



University of
Sheffield

I Have No Mouth, and I Must Glean:
Assessing Phonetically and Phonologically
Motivated Language Understanding

Tyler Cameron Loakman

Supervisors:

Prof. Chenghua Lin

Prof. Robert Gaizauskas

A thesis submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy in Computer Science

in the

Department of Computer Science

January 25, 2026

Declaration

I, Tyler Cameron Loakman, hereby declare that this thesis has been composed solely by myself and that it has not been submitted, in whole or in part, in any previous application for a degree. Except where stated otherwise by reference or acknowledgement, the work presented is entirely my own.

Acknowledgements

Firstly, I would like to thank my supervisor, Prof. Chenghua Lin, whose willingness to allow me to explore niche topics was integral to creating this body of work.

Thanks also go to my parents, Noel and Adele, for their support and for ensuring that I never forgot the meaning of the abbreviations NLP, NLG, CL, ML, or PhD, by regularly needing me to explain them.

Additionally, I would like to thank Dr. Phil Harrison, whose guidance was instrumental in convincing me to take a chance and apply for a PhD position in a computer science department. Without it, this thesis would never have existed.

The support of friends, co-authors, and colleagues has also been greatly appreciated. Particular thanks go out to Joey James, Liam Thorne, Will Ravenscroft, Jasivan Sivakumar, Meg Thomas, Tomas Goldsack, Carolina Scarton, Tom Pickard, Chen Tang, Aaron Maladry, Cliodhna Hughes, Yucheng Li, Sabrina McCallum, and Annie Baker.

Finally, I would like to thank Rafferty (Figure 1), also known as “Raffiki”, “Raffikins”, and “The Cat™”, for providing stress relief and ensuring my ego stayed in check by never once appearing to care when he sees me.

This work was supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications, funded by UK Research and Innovation [grant number EP/S023062/1].



Figure 1: *Cat.*

Abstract

Whilst language models have long been capable of generating grammatical and coherent text, their understanding of phonetically and phonologically motivated language remains weak. Primarily, this is due to such models being trained exclusively on orthographic text, losing much of the information contained in speech. This thesis proposes approaches to assessing and alleviating this phonetic and phonological deficit, regarding both explicit linguistic knowledge and the impact on downstream tasks such as creative language generation, humour explanation, and in silico sound symbolism experimentation. First, the thesis establishes a standard for the human evaluation of subjective language. By exploring the impact of demographic variables on the interpretation of such language, the thesis establishes best practices for assessing creative generation, such as the phonological word-play that is explored in subsequent chapters. It then introduces the task of tongue twister generation to assess the ability of pre-trained language models to implicitly learn phonological patterns from orthographic text, introducing a phonologically informed generation pipeline, two novel tongue twister datasets, a constrained decoding module, and phoneme-level evaluation metrics. Following this, the thesis transitions towards assessing the understanding and reasoning of large language models, exploring their ability to explain the humour underlying different types of jokes, from phonologically inspired heterographic puns to long-form topical humour. Multimodal phonetic and phonological understanding is then assessed by replicating a range of classic psycholinguistic experiments regarding sound symbolism using vision language models, demonstrating their capacity for in silico experimentation. Multimodality is further explored through the novel

task of spectrogram and waveform interpretation, assessing the integration of visual information with parametric linguistic knowledge. The thesis concludes with a comparative investigation into the perspectives of artificial intelligence practitioners and the general public, retrospectively assessing the desire for creative language generation and *in silico* experimentation, whilst exploring the wider societal impact of such technologies overall.

Contents

I	Introduction	6
1	Thesis Overview	7
1.1	Thesis Aim	8
1.2	Thesis Structure	10
1.3	Research Questions	11
1.4	Thematic Progression of Research	18
 Survey I		
<i>From Sounds to Symbols: A Survey of Phonetically and Phonologically</i>		
Informed Language Generation		20
1.5	Introduction	23
1.5.1	Related Work	24
1.5.2	Survey Outline	24
1.6	Upstream Foundational Tasks	25
1.6.1	Grapheme-to-Phoneme Conversion	25
1.6.2	Language Modelling	28
1.7	Downstream Tasks	32
1.7.1	Poetry	32
1.7.2	Humour	35

CONTENTS

1.7.3 Translation	37
1.8 Conclusion	39
1.9 Limitations	39
1.10 Appendix	40
1.10.1 Survey Inclusion Criteria	40

Survey II

***Who's Laughing Now?* An Overview of Computational Humour**

Generation and Explanation	41
1.11 Introduction	44
1.11.1 Related Work	44
1.11.2 Survey Outline	45
1.12 The Role of Computational Humour	45
1.12.1 Theories of Humour	47
1.13 Humour Generation	48
1.13.1 Pun Generation	49
1.13.1.1 Homographic Pun Generation	49
1.13.1.2 Heterographic Pun Generation	50
1.13.1.3 Combined Pun Generation	51
1.13.2 Non-Pun Humour Generation	52
1.14 Humour Explanation	54
1.14.1 Explanation through Classification	54
1.14.2 Natural Language Explanation	56
1.15 Conclusion	60
1.16 Limitations	60
1.17 Ethics Statement	61

CONTENTS

II Human Evaluation 62

2	<i>The Iron(ic) Melting Pot: Reviewing Human Evaluation in Humour, Irony, and Sarcasm Generation</i>	63
2.1	Introduction	66
2.2	Related Works	67
2.3	Introducing Humour, Irony, and Sarcasm	69
2.3.1	Humour	69
2.3.2	Irony	71
2.3.3	Sarcasm	73
2.4	Critical Survey	75
2.4.1	Scope and Selection	75
2.4.2	Subject Prevalence	76
2.4.3	Reporting Transparency	77
2.5	Proposed Reporting Criteria & Discussion	81
2.6	Conclusion	84
2.7	Limitations	84
2.8	Appendix	86
2.8.1	Surveyed Generation Papers	86

III Tongue Twister Generation 87

3	<i>TwistList: Resources and Baselines for Tongue Twister Generation</i>	88
3.1	Introduction	91
3.2	Related Works	93
3.3	Methodology	95
3.3.1	Task	95

CONTENTS

3.3.2	Dataset (TwistList)	95
3.3.3	Models (TwisterMisters)	98
3.4	Results	98
3.5	Case Study	102
3.6	Conclusion	103
3.7	Limitations	104
3.8	Ethics Statement	105
3.9	Appendix	105
3.9.1	Human Participants	105
3.9.2	Implementation Details	106
4	<i>Train and Constrain: Phonologically Informed Tongue Twister Generation from Topics and Paraphrases</i>	107
4.1	Introduction	110
4.1.1	Contributions	111
4.2	Phonetics & Phonology	113
4.2.1	Place & Manner of Articulation	113
4.2.2	Phonological Features	114
4.2.3	Phonetic Transcription Standards	114
4.3	Related Work	115
4.3.1	Constrained Generation	115
4.3.2	Knowledge Distillation	117
4.4	Dataset (TwistList 2.0)	117
4.4.1	TwisterLister Pipeline	118
4.4.2	Style-Transfer Paraphrase Generation	121
4.4.3	Output Refinement	122

CONTENTS

4.4.4	Quality Control & Enhancement	124
4.4.5	TwistList 2.0 Summary	126
4.5	Fine-Tuning Methodology	129
4.5.1	Tasks	129
4.5.2	Models	129
4.5.3	Automatic Metric Suite	131
4.5.4	Human Evaluation Protocol	134
4.6	Fine-Tuning Results	136
4.6.1	Automatic Evaluation	136
4.6.2	Human Evaluation	141
4.6.3	Human vs. Automatic Metrics	143
4.7	Constrained Decoding Methodology	146
4.7.1	Task	146
4.7.2	Phoneme-Aware Constrained Decoding (PACD)	147
4.7.3	Models	151
4.8	Constrained Decoding Results	151
4.8.1	Automatic Evaluation	152
4.8.2	Human Evaluation	155
4.9	Qualitative Analysis	158
4.9.1	Topic-to-Twister (Fine-Tuned)	158
4.9.2	Topic-to-Twister (Constrained Decoding)	162
4.9.3	Style-Transfer (Fine-Tuned)	166
4.10	Conclusion	171
4.11	Limitations	172
4.12	Appendix	173
4.12.1	Evaluation Rubric	173

CONTENTS

4.12.2	Implementation Details	176
IV	Humour Explanation	177
5	<i>Comparing Apples to Oranges: Analysing LLM Understanding from Traditional Puns to Topical Humour</i>	178
5.1	Introduction	181
5.2	Related Works	183
5.3	Dataset	185
5.3.1	Existing Datasets	185
5.3.2	Our Dataset	185
5.3.3	Ground Truth Joke Explanations	187
5.3.4	Dataset Enhancement	188
5.4	Methodology	188
5.4.1	Hypotheses	188
5.4.2	Models	189
5.4.3	Evaluation Criteria	190
5.5	Human Evaluation	192
5.5.1	Explanation Success Rate	194
5.6	Automatic Evaluation	196
5.7	Case Study	197
5.8	Discussion	198
5.9	Conclusion	199
5.10	Limitations	200
5.11	Ethics Statement	201
5.12	Appendix	202

CONTENTS

5.12.1	Implementation Details	202
5.12.2	Dataset Enhancement Details	203
5.12.3	Additional Qualitative Analysis	204
5.12.4	LLM-as-a-Judge Protocol	206
5.12.5	Additional Annotator & LLM-as-a-Judge Results	207
5.12.6	Additional Automatic Evaluation	210
V	Multimodal Understanding	212
6	<i>With Ears to See and Eyes to Hear: Sound Symbolism Experiments</i>	
	with Multimodal Large Language Models	213
6.1	Introduction	216
6.2	Related Works	218
6.3	Shape Symbolism	221
6.3.1	Methodology	222
6.3.2	Results	224
6.4	Magnitude Symbolism	226
6.4.1	Methodology	227
6.4.2	Results	228
6.5	Iconicity Ratings	229
6.5.1	Methodology	229
6.5.2	Results	230
6.6	Discussion	231
6.6.1	Sound Symbolism Source	231
6.6.2	Future Directions	233
6.7	Conclusion	233

CONTENTS

6.8	Limitations	234
6.9	Ethics Statement	234
6.10	Appendix	236
6.10.1	Implementation Details	236
6.10.2	Iconicity Prompt	237
6.10.3	Full Results	238
7	<i>Seeing isn't Hearing: Benchmarking Vision Language Models at Interpreting Spectrograms</i>	240
7.1	Introduction	243
7.2	Related Work	244
7.3	Dataset	245
7.3.1	Word List	245
7.3.2	Speech Synthesis	246
7.3.3	Spectrograms & Waveforms	246
7.4	Methodology	247
7.4.1	Task	247
7.4.2	Models	248
7.5	Results	249
7.5.1	Automatic Speech Recognition & Phonetician Baselines	251
7.6	Conclusion	252
7.7	Limitations	252
7.8	Appendix	254
7.8.1	Dataset Characteristics	254
7.8.2	Implementation Details	254

CONTENTS

VI Perspectives on AI Development 257

8	<i>Do as I Say, Not as You Do: Assessing Public Versus Practitioner Perspectives on the Development of Artificial Intelligence and Language Technologies</i>	258
8.1	Introduction	261
8.2	Related Works	263
8.3	Methodology	264
8.4	Artificial Intelligence Usage	266
8.5	Training Data Collection	268
8.6	Art & Creativity	271
8.7	Thought & Consciousness	273
8.8	Societal Impact	275
8.8.1	Accessibility	277
8.8.2	Research	278
8.8.3	Healthcare	278
8.8.4	Wellbeing	279
8.8.5	Information Integrity	279
8.9	Conclusion	280
8.10	Limitations	281
8.11	Appendix	281
8.11.1	Comparison to Existing Work	281
8.11.2	Original Question Phrasing & Statistical Testing	282
8.11.3	Additional Demographic Information	282

CONTENTS

VII	Conclusion	285
9	Thesis Conclusions	286
9.1	Research Questions Discussion	286
9.2	Impact	293
9.3	Future Work	295
9.3.1	Phonetically Motivated Language Understanding	296
9.3.2	Computational Humour Processing	298
9.3.3	Demographic-Aware Humour Generation	298
9.3.4	Human-in-the-Loop Humour Generation	299
9.3.5	Complex Humour Explanation	300
9.3.6	Intent-Aware Humour Explanation	300

List of Publications

LIST OF PUBLICATIONS

The following publications concern works authored during the PhD programme associated with this thesis. Publications that form part of this thesis are in **bold**.

First-Author

- T Loakman, C Lin. ***Do as I Say, Not as You Do: Assessing Public Versus Practitioner Perspectives on the Development of Artificial Intelligence and Language Technologies.***
(Under Review)
- T Loakman, J James, C Lin. ***From Sounds to Symbols: A Survey of Phonetically and Phonologically Informed Language Generation.***
(Under Review)
- T Loakman, J James, C Lin. ***Seeing isn't Hearing: Benchmarking Vision Language Models at Interpreting Spectrograms.***
(Proceedings of IJCNLP-AACL, [Loakman et al., 2025a](#))
- T Loakman, W Thorne, C Lin. ***Who's Laughing Now? An Overview of Computational Humour Generation and Explanation.***
(Proceedings of INLG, [Loakman et al., 2025d](#))
- T Loakman, W Thorne, C Lin. ***Comparing Apples to Oranges: Analysing LLM Understanding from Traditional Puns to Topical Humour.***
(Findings of EMNLP, [Loakman et al., 2025c](#))
- T Loakman, C Tang, C Lin. ***Train & Constrain: Phonologically Informed Tongue-Twister Generation from Topics and Paraphrases.***
(Computational Linguistics, [Loakman et al., 2025b](#))

LIST OF PUBLICATIONS

- T Loakman, Y Li, C Lin. *With Ears to See and Eyes to Hear: Sound Symbolism Experiments with Multimodal Large Language Models*.
(Proceedings of EMNLP, [Loakman et al., 2024](#))
- T Loakman, C Lin. *ReproHum #0087-01: Human Evaluation Reproduction Report for “Generating Fact-Checking Explanations”*.
(HumEval Workshop, LREC-COLING, [Loakman and Lin, 2024](#))
- T Loakman, A Maladry, C Lin. *The Iron(ic) Melting Pot: Reviewing Human Evaluation in Humour, Irony and Sarcasm Generation*.
(Findings of EMNLP, [Loakman et al., 2023a](#))
- T Loakman, C Tang, C Lin. *TwistList: Resources and Baselines for Tongue Twister Generation*.
(Proceedings of ACL, [Loakman et al., 2023b](#))

Co-Author

- S Wu et al. *COIG-P: A High-Quality and Large-Scale Chinese Preference Dataset for Alignment with Human Values*.
(Findings of EACL, 2026)
- Y Wang, C Xiao, C-Y Hsiao, ZY Chang, C-L Chen, **T Loakman**, C Lin. *Drivelology: Challenging LLMs with Interpreting Nonsense with Depth*.
(Proceedings of EMNLP, [Wang et al., 2025b](#))
- C Tang, H Zhang, **T Loakman**, B Yang, S Goetze, C Lin. *CADGE: Context-Aware Dialogue Generation Enhanced with Graph-structured Knowledge Aggregation*.
(Proceedings of INLG, [Tang et al., 2024b](#))

LIST OF PUBLICATIONS

- S Wang, G Zhang, H Wu, **T Loakman**, W Huang, C Lin. *MMTE: Corpus and Metrics for Evaluating Machine Translation Quality for Metaphorical Language*. (Proceedings of EMNLP, [Wang et al., 2024b](#))
- C Tang, **T Loakman**, C Lin. *A Cross-Attention Augmented Model for Event-Triggered Context-Aware Story Generation*. (Computer Speech & Language, [Tang et al., 2024a](#))
- C Tang, H Zhang, **T Loakman**, C Lin and F Guerin. *Enhancing Dialogue Generation via Dynamic Graph Knowledge Aggregation*. (Proceedings of ACL, [Tang et al., 2023a](#))
- C Tang, H Zhang, **T Loakman**, C Lin, F Guerin. *Terminology-Aware Medical Dialogue Generation*. (Proceedings of ICASSP, [Tang et al., 2023b](#))
- T Pickard, **T Loakman**, and M Pandya. *shefnlp at SemEval-2023 Task 10: Compute-Efficient Category Adapters*. (SemEval, [Pickard et al., 2023](#))
- H Huang, C Tang, **T Loakman**, F Guerin, C Lin. *Improving Chinese Story Generation via Awareness of Syntactic Dependencies and Semantics*. (Proceedings of IJCNLP-AAACL, [Huang et al., 2022](#))
- C Tang, Z Zhang, **T Loakman**, C Lin, F Guerin. *NGEP: A Graph-Based Event Planning Framework for Story Generation*. (Proceedings of IJCNLP-AAACL, [Tang et al., 2022b](#))

LIST OF PUBLICATIONS

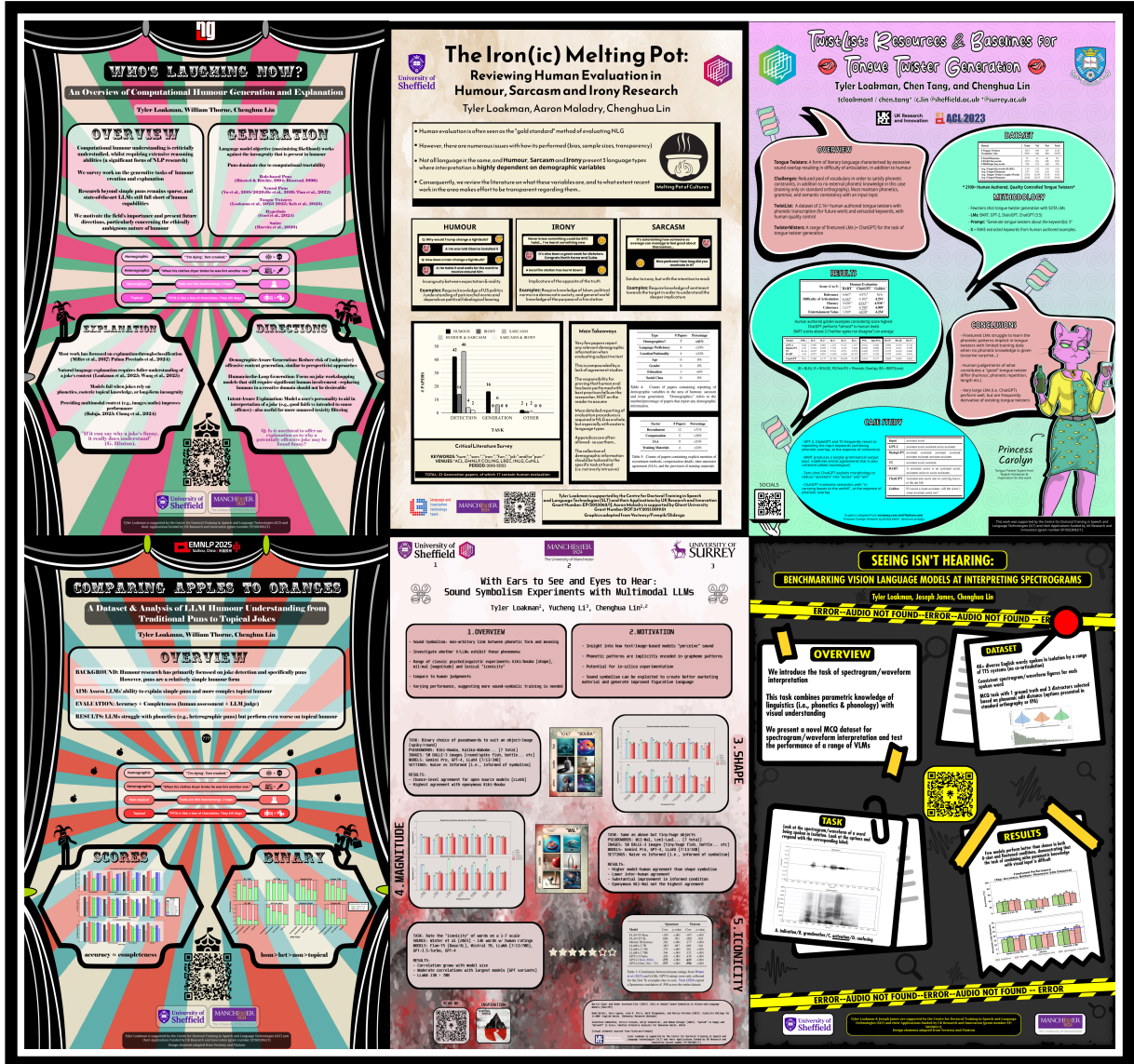


Figure 2: A subset of the conference posters designed for publications within this thesis (in order of appearance).

Part I

Introduction

Chapter 1

Thesis Overview

Language model pre-training primarily focuses on learning grammatical correctness and semantic coherence from text ([Common Crawl Foundation, 2025](#); [Gao et al., 2020](#)). However, most human communication occurs through speech, of which text is a lossy representation ([Yadavalli et al., 2025](#); [Wolf et al., 2023](#); [Bloomfield, 1933](#); [de Saussure, 1916](#)). This information loss stems from factors such as opaque orthography and the standardisation and codification of spoken language into writing, compressing variations into a singular representation ([Marjou, 2021](#)). Despite this, myriad language forms exploit the characteristics of speech, such as articulatory difficulty, phonetic similarity, and symbolic associations between sounds and semantics ([Guerini et al., 2015](#); [Yorkston and Menon, 2004](#); [Redfern, 1987](#)). Whilst such language can be written, readers must relate a text’s orthographic form to its phonetic and phonemic forms to gain a full appreciation ([Wei et al., 2022](#); [McCutchen and Perfetti, 1982](#)). Due to this mismatch between spoken communication and textual representations, language models trained exclusively on orthographic text struggle to both generate and understand such language ([Suvarna et al., 2024](#); [Cuskley et al., 2024](#)), whilst research in the domain remains comparatively scarce, owing to the required expertise in phonetics and phonology.

1.1 Thesis Aim

To systematically address the disconnect between text and speech, the thesis aims to investigate the abilities of orthographic large language models (LLMs) and vision language models (VLMs) in the processing of textual and visual representations of language that exploit phonetic and phonological properties that are not explicitly rendered in written text. Specifically, this thesis addresses the following overarching research question:

RQ0. To what extent can large language models and vision-language models bridge the inherent information loss between orthographic text and spoken language to generate, reason about, and visually interpret phonetic, phonological, and acoustic phenomena?

To answer this question, this thesis presents a range of novel tasks for assessing phonetic and phonological understanding, including i) generating tongue twisters to investigate the understanding of phonetic similarity and alliteration in generative models; ii) reasoning about phonetic similarity through the task of pun explanation; iii) exhibiting human-like sound symbolism associations between an entity’s visual characteristics and speech sounds; and iv) visually interpreting representations of speech in the form of spectrograms and waveforms.

As a secondary theme of this thesis, phonetic and phonological understanding is repeatedly explored through the lens of subjective creative language, specifically through humour. For example, tongue twisters present a language form that explicitly relies on articulatory characteristics to reach desirable outcomes. Whilst other phonetically

CHAPTER 1. THESIS OVERVIEW

motivated language exists such as poetry, articulatory characteristics are integral in tongue twisters whilst being an optional characteristic in poetry. Furthermore, the aim of the articulatory difficulty in tongue twisters is to encourage mispronunciation for humorous effect. Consequently, pun explanation further explores a practical use of phonetic and phonological knowledge to explain jokes whilst simultaneously allowing an investigation into how models reason about both phonetics and humour, further increasing the practical implications of findings. As humour is a creative and subjective domain, the thesis additionally explores best practices in the human evaluation of such language as well as the ethical implications of AI research on creative domains.

The thesis therefore transitions sequentially from evaluating human subjectivity in creative language, to testing explicit generative capabilities under phonological constraints, to evaluating cross-modal and abstract semantic-acoustic associations, and finally to assessing the socio-ethical implications of automating linguistic domains that rely on human creativity. A visual overview of the central theme of this thesis is presented in Figure 1.1, demonstrating the information loss between speech and text and the range of tasks investigated.

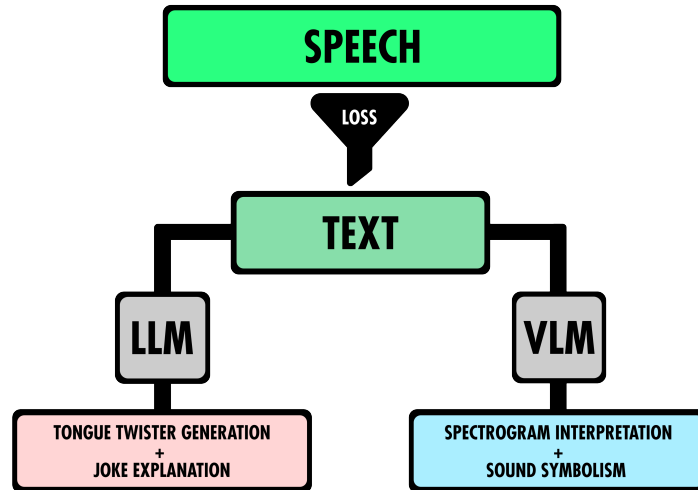


Figure 1.1: *Illustrated thesis outline of the tasks performed with large language models (LLMs) and vision language models (VLMs).*

1.2 Thesis Structure

The structure of this thesis concerns Parts, Chapters, and Research Questions. Parts refer to overarching research themes that group together Chapters (e.g., tongue twister generation or multimodal understanding). On the other hand, each Chapter relates to one publication and one primary Research Question. The exception is the current Introduction, which contains two publications in the form of literature surveys regarding phonetically/phonologically motivated language generation (Survey I) and humour generation/explanation (Survey II), providing wider context to the domains explored in this thesis.¹ A full table of concordance between Parts, Chapters, and Research Questions is presented in Table 1.1. In the next section, the primary research questions posed and addressed by this thesis are introduced.

Part	Chapter	Research Question
Introduction	1	
Human Evaluation	2	RQ1
Tongue Twister Generation	3	RQ2
	4	RQ3
Humour Explanation	5	RQ4
Multimodal Understanding	6	RQ5
	7	RQ6
Perspectives on AI Development	8	RQ7
Conclusion	9	

Table 1.1: *Table of concordance.*

¹Both literature surveys were performed towards the end of the thesis project and therefore cite published versions of work contained within this thesis.

1.3 Research Questions

RQ1. To what extent do demographic factors affect the interpretation of the intensity of humour, irony, and sarcasm; how transparently are human evaluation procedures reported in existing works; and how can the reporting of human evaluation be standardised to account for this subjectivity?

Chapter 2: *The Iron(ic) Melting Pot: Reviewing Human Evaluation in Humour, Irony, and Sarcasm Generation* (Findings of EMNLP, [Loakman et al., 2023b](#))

Summary: This chapter addresses RQ1 by surveying known demographic factors that affect the interpretation of the complex, subjective language forms of humour, irony, and sarcasm. Assessing common practices in human evaluation of these language types, it reveals the need for standardised criteria for performing and reporting human evaluation of subjective language, in addition to the lack of comprehensive and transparent reporting in existing works. The observations made within this chapter inform the depth and transparency of the design, implementation, and reporting of all human evaluation presented in subsequent chapters.

Individual Contributions: *Tyler Loakman* - Conceptualisation, humour/irony/sarcasm survey, transparency survey, standardised evaluation recommendations. *Aaron Maladry* - irony/sarcasm survey, standardised evaluation recommendations. *Chenghua Lin* - Research supervision.

RQ2. To what extent can grapheme-based, pre-trained language models learn phonological constraints from human-authored samples to generate tongue twisters?

Chapter 3: *TwistList: Resources and Baselines for Tongue Twister Generation* (Proceedings of ACL, [Loakman et al., 2023b](#))

Summary: This chapter addresses RQ2 by assessing the ability of small pre-trained language models to generate tongue twisters, a form of phonologically constrained language designed to be difficult to articulate. This chapter introduces the TwistList dataset, consisting of 2.1k+ human-authored tongue twisters from various online platforms. Additionally, naive phoneme-overlap metrics are introduced to assess the level of sound repetition present in tongue twister outputs. The chapter identifies the challenges that arise when generating phonologically conditioned language, including the difficulty language models face in achieving sound repetition whilst maintaining semantic coherence and the tendency for larger models to produce derivative, non-creative outputs.

Individual Contributions: *Tyler Loakman* - Conceptualisation, dataset creation, phoneme evaluation metrics, human evaluation protocol. *Chen Tang* - Experimental design, automatic evaluation, human evaluation platform, paper writing. *Chenghua Lin* - Research supervision.

RQ3. To what extent does incorporating synthetic fine-tuning data and constrained decoding improve the ability of grapheme-based language models to generate tongue twisters, and what insights are provided by phonologically informed metrics?

Chapter 4: *Train & Constrain: Phonologically Informed Tongue Twister Generation from Topics and Paraphrases* (Computational Linguistics, [Loakman et al., 2025b](#))

Summary: Building on the findings of Chapter 3, this chapter addresses RQ3 by developing a tongue twister generation pipeline (TwisterLister) using LLMs alongside a phonetically transcribed dictionary and semantically embedded topic phrases. This pipeline is used to extend the original TwistList dataset into TwistList 2.0, consisting of 17k tongue twisters. Using this dataset, a range of smaller models are fine-tuned to investigate the effects of fine-tuning data volume on the ability of language models to learn phonological properties from orthographic text alone. Also presented in this chapter is a phoneme-aware constrained decoding module (PACD), a constraint-based decoding algorithm that generates tongue twisters without the need for fine-tuning. Finally, the phoneme evaluation suite is extended to include metrics that incorporate phonemic edit distance, extending the metrics introduced in Chapter 3 to take into account the feature similarity of different phonemes, rather than treating all sounds as equally distinct.

Individual Contributions: *Tyler Loakman* - Conceptualisation, TwisterLister pipeline, constrained decoding module, phonemic edit distance metrics, human evaluation protocol, paper writing. *Chen Tang* - Experimental design, automatic evaluation, paper writing. *Chenghua Lin* - Research supervision.

RQ4. To what extent does joke typology affect the ability of large language models to explain humour, and how does the transition from sound-based wordplay to topical humour expose reasoning limits?

Chapter 5: *Comparing Apples to Oranges: Analysing LLM Understanding from Traditional Puns to Topical Humour* (Findings of EMNLP, [Loakman et al., 2025c](#))

Summary: Continuing with the theme of humour previously investigated via tongue twisters, this chapter addresses RQ4 by critically analysing the ability of LLMs to explain a range of jokes. By comparing human ratings of explanation accuracy and completeness across homographic and heterographic puns, in addition to topical and non-topical jokes from Reddit, this chapter highlights that existing research has been narrowly focused on a small subset of humour types. Regarding phonetic and phonological understanding, this chapter identifies the tendency of LLMs to hallucinate phonetic similarities between words when tasked with explaining heterographic puns. Topical humour is additionally explored, where understanding is reliant on detecting subtle references to real-world entities and events, presenting further challenges.

Individual Contributions: *Tyler Loakman* - Conceptualisation, dataset creation, gold standard explanations, human evaluation protocol, paper writing. *William Thorne* - Automatic evaluation, paper writing. *Chenghua Lin* - Research supervision.

RQ5. To what extent do large language models and vision language models reflect human psycholinguistic associations regarding sound symbolism?

Chapter 6: *With Ears to See and Eyes to Hear: Sound Symbolism Experiments with Multimodal Large Language Models* (Proceedings of EMNLP, [Loakman et al., 2024](#))

Summary: Transitioning from humour but remaining focused on phonetics and phonology, this chapter addresses RQ5 by exploring the ability of LLMs to engage with psycholinguistic phenomena by replicating a range of classic psycholinguistics experiments regarding sound symbolism with text-based LLMs and vision-enabled VLMs. Specifically, the alignment between models and humans on shape and magnitude symbolism tasks is explored using a novel dataset, in addition to evaluating LLMs on the task of providing iconicity scores for English words. Consequently, this chapter explores the extent to which the cognitive semantic effects of phonetic and phonological phenomena seen in humans also apply to models that do not have access to audio and must instead learn implicitly from orthographic text.

Individual Contributions: *Tyler Loakman* - Conceptualisation, shape/magnitude symbolism dataset creation, experimental design, paper writing. *Yucheng Li* - Experimental design. *Chenghua Lin* - Research supervision.

RQ6. To what extent can vision language models synthesise parametric knowledge of phonetics and phonology to correctly interpret visual representations of speech?

Chapter 7: *Seeing isn't Hearing: Benchmarking Vision Language Models at Interpreting Spectrograms* (IJCNLP-AACL, [Loakman et al., 2025a](#))

Summary: Moving away from sound symbolism yet continuing with multimodality, this chapter addresses RQ6 by investigating whether VLMs can correctly predict spoken words from visual representations in the form of spectrograms and waveforms, combining parametric knowledge of acoustic phonetics with visual inputs. This chapter presents a novel dataset of spectrogram and waveform figures for 4k+ English words spoken in isolation by text-to-speech models. A novel multiple-choice-question dataset is then generated, consisting of the ground truth answer alongside three distractors that are selected based on their phonemic edit distance to the ground truth. Overall, this chapter demonstrates the difficulty of incorporating linguistic knowledge contained in pre-training corpora with esoteric data visualisations.

Individual Contributions: *Tyler Loakman* - Conceptualisation, dataset creation, experimental design, expert phonetician evaluation, paper writing. *Joseph James* - Experimental design, paper writing. *Chenghua Lin* - Research supervision.

RQ7. To what extent do the perspectives of practitioners and the general public differ towards the development and use of artificial intelligence technology, encompassing factors such as creativity, consent, and wellbeing?

Chapter 8: *Do as I Say, Not as You Do: Assessing Public Versus Practitioner Perspectives on the Development of Artificial Intelligence and Language Technologies*
(Under Review)

Summary: Following the transition from creative language generation to explanation, and finally towards multimodal understanding, this chapter addresses RQ7 by assessing the ethical and subjective nature of artificial intelligence (AI) use and development. By surveying practitioners (academics and industry professionals) and the general public, the differences of opinion on fundamental areas of AI are understood. This work highlights the lack of alignment between those who create such technologies and those most impacted by them, whilst critically evaluating the ethics and impact of contributions from earlier chapters, including the desire for creative language generation and whether in silico experimentation is societally desirable.

Individual Contributions: *Tyler Loakman* - Conceptualisation, questionnaire design, participant recruitment, research survey, paper writing. *Chenghua Lin* - Research supervision.

1.4 Thematic Progression of Research

Whilst each chapter of this thesis is focused on an independent research publication, the overall progression of this thesis will now be outlined. As later chapters concern the generation and explanation of creative language, including tongue twisters and jokes, in addition to abstract psycholinguistic phenomena, Chapter 2 presents an overview of human evaluation concerning the subjective language forms of humour, sarcasm, and irony. A critical survey of existing generative works is performed, demonstrating that little existing research transparently reports relevant factors that may affect human evaluation of subjective language forms. Consequently, a framework of evaluation best practices is developed, which informs the performance and reporting of human evaluation in later chapters. Chapter 3 presents the task of topic-based tongue twister generation, introducing the largest dataset of human-authored tongue twisters at the time of publication, alongside naive evaluation metrics for assessing phoneme repetition. Chapter 4 builds upon this work by introducing a phoneme-aware LLM-based pipeline for generating synthetic tongue twisters, resulting in a larger tongue twister dataset for analysing the effects of fine-tuning data volume on the ability of smaller language models to learn the implicit relationship between text and speech. This chapter further introduces a style-transfer variant of tongue twister generation, a phoneme-aware constrained decoding approach for tongue twister generation without the need for model fine-tuning, and phonologically informed evaluation metrics based on phonemic edit distance, therefore taking into account the similarity of related speech sounds. Following work on generating humorous, phonologically motivated language in the form of tongue twisters, Chapter 5 turns attention toward generating explanations of jokes, quantifying the difficulty of semantics-based homographic puns, sound-based heterographic puns, and more complex topical humour. Multimodality is then introduced in Chapter 6 and

CHAPTER 1. THESIS OVERVIEW

Chapter 7. Firstly, Chapter 6 investigates whether VLMs understand the phenomenon of sound symbolism, comparing model and human preferences on tasks pertaining to shape symbolism, magnitude symbolism, and linguistic iconicity. Chapter 7 then tests VLMs on a further meta-linguistic task of interpreting spectrograms and waveforms of speech, returning to the topic of esoteric information retrieval explored through topical humour explanation in Chapter 5. The chapter presents a novel dataset of speech visualisations and a phonologically informed multiple-choice task where distractors are selected based on phonemic edit distance from the ground truth. Finally, Chapter 8 takes a broader look at the development of artificial intelligence (AI) technologies as a whole, whilst simultaneously taking a retrospective look at the ethics and impact of the creative language generation and *in silico* experimentation performed in earlier chapters. Chapter 8 assesses the differences of opinion between practitioners and the general public and the implications for future model alignment and societal impact. Whilst Chapter 3 and Chapter 4 concern creative language generation through tongue twisters, rising contention regarding the role of AI in creative industries prompted the transition from explicit generation, towards the *explanation* of creative language demonstrated in Chapter 5, and the abstract vision-based tasks presented in Chapter 6 and Chapter 7. Consequently, Chapter 8 acts as an overview of the ethical and practical concerns of future AI developments. For each chapter, additional contextualising statements are provided to reiterate the addressed research questions and thematic progression.

The publications included in this thesis have been edited from their published versions for additional clarity and consistency. The underlying data, methodologies, and findings remain unchanged.

Survey I:

From Sounds to Symbols: A Survey of Phonetically and Phonologically Informed Language Generation

Under Review

Tyler Loakman^a, Joseph James^a, and Chenghua Lin^b

^aDepartment of Computer Science, University of Sheffield, UK

^bDepartment of Computer Science, University of Manchester, UK

Contextual Statement

As the central research theme of this thesis concerns phonetic and phonological language processing, this survey presents a wide overview of existing work in the primary research area of the thesis.

Abstract

Text, although written using symbols and characters, is not void of connection to the predominantly spoken form of language. Whilst syntax and semantics are the cornerstones of meaningful text, a substantial portion of human communication relies heavily on the specific sounds and sound patterns observed in language. In this survey paper, we present an outline of the areas of computational linguistics, specifically natural language generation (NLG), that are either dependent upon, or benefit from, an additional understanding of phonetics and phonology. We cover a range of subdisciplines, from practical areas such as and grapheme-to-phoneme conversion (G2P) to creative endeavours such as poetry and humour generation. We aim for this overview to situate the role of phonetics and phonology in NLG, and invite the community at-large to consider how the study of such features may be able to positively impact different domains through fusing characteristics of both speech and text.

1.5 Introduction

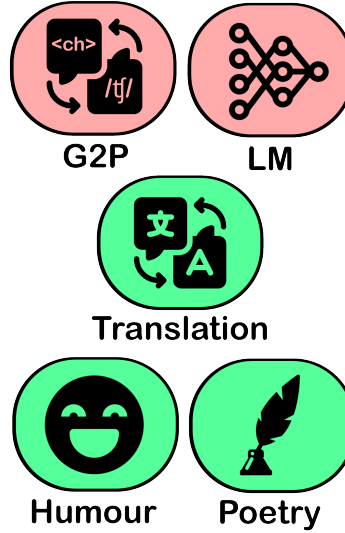


Figure 1.2: *Illustration of the areas surveyed in this work. G2P refers to grapheme-to-phoneme conversion, whilst LM refers to the general domain of language modelling. We distinguish between upstream foundational tasks (red) and downstream applications (green).*

To successfully generate coherent language, computational models must understand several key areas of linguistics. Fundamentally, the area of syntax dictates the ordering of words and the grammatical relationships between constituents (Moravcsik, 2006; Van Valin, 2001), whilst semantics addresses the meaning of language and the message that is to be conveyed (Kroeger, 2018; Hofmann, 1993). The former is essential to not generate random assortments of words, avoiding “cat the sat on mat his” in favour of “his cat sat on the mat”, whilst the latter is what explains the nonsensical nature of the sentence “They quickly ran *upwards down* the stairs”. Both of these elements are essential to the majority of written language. However, as computational approaches such as large language models (LLMs) become adept at handling such concerns, the challenges presented by other areas of linguistics become clearer. Of these, the related areas of phonetics and phonology concern the production, perception, and distribution

of speech sounds across human languages (Rogers, 2013; De Lacy, 2007; Wells, 1982b). Such elements are essential to many aspects of human communication, even in its written form, complementing syntax and semantics and being the fundamental elements that separate prose from poetry and other creative uses of language.

1.5.1 Related Work

To the best of our knowledge, this work presents the first survey paper dedicated to the topic of phonetically and phonologically motivated developments in NLG. However, there exists a range of survey papers that cover individual areas in depth, particularly those where conventional approaches do not rely explicitly on such characteristics. Cheng et al. (2024) provides an overview of G2P in multilingual settings, covering methods, datasets and evaluation metrics, Gonçalves Oliveira (2017) cover the area of intelligent poetry generation, Watanabe and Goto (2020) address the automatic generation of lyrics, and Loakman et al. (2025d) and Amin and Burghardt (2020) survey approaches to humour generation. We focus on novel approaches to incorporating phonetic and phonological knowledge into these tasks, rather than overall paradigms and state-of-the-art approaches, where the novelty is not linguistically motivated. The primary aim of this survey is one of breadth concerning the inclusion of phonetic and phonological knowledge in natural language generation, demonstrating the richness of research in this subdiscipline for a domain-adjacent readership, rather than per-topic depth for area specialists.

1.5.2 Survey Outline

In this survey, we focus primarily on works where phonetic and phonological knowledge has been incorporated into the process of natural language *generation*. We do,

however, take a broad view of generation to also include elements such as the natural language explanation of phonetically and phonologically motivated text. Regarding paper inclusion, we utilised a systematic backward snowball approach beginning with articles found on the ACL Anthology and Google Scholar using a list of domain-relevant keywords, considering only the first page of results (see Appendix 1.10.1 for the full list of terms). From there, cited works were obtained and considered for inclusion. In §1.6.1 we outline grapheme-to-phoneme conversion, a fundamental step for most other phonetically and phonologically informed works. In §1.6.2 we cover the broad topic of language modelling and approaches to incorporating relevant knowledge without a specific downstream task (including prompting methods, knowledge probing, and general evaluation). Following this, we address the large scale generative tasks of poetry generation in §1.7.1 and humour generation in §1.7.2, followed by translation in §1.7.3.

1.6 Upstream Foundational Tasks

1.6.1 Grapheme-to-Phoneme Conversion

Grapheme-to-phoneme conversion (G2P) is a common prerequisite for the processing of text using phonetic and phonological knowledge for downstream tasks, particularly for languages with opaque deep orthography. G2P centres on the process of converting *graphemes*, the character-level writing units of a language, to *phonemes*, the smallest meaningful units of sound. As such, it allows the investigation of the sounds that would be produced if written text was to be read aloud, assuming a specific accent. For instance, G2P may convert the word *bath* into /bɑːθ/ for a British speaker of Southern Standard British English (SSBE, Wells, 1982b), /baθ/ for a speaker of General Northern English (Wells, 1982b), or /bæθ/ for a speaker of General American (Wells, 1982a).² It

²These constitute three distinct collections of accents, each with variation in practice.

is therefore paramount to the task of G2P that phonetic and phonological knowledge is available in the form of transcribed training data such as pronunciation dictionaries or aligned speech and text (Ribeiro et al., 2023; Route et al., 2019; Aquino et al., 2019).

G2P is commonplace in tasks such as automatic speech recognition (ASR) and speech synthesis (i.e., text-to-speech), particularly in multilingual and low-resource settings (Cheng et al., 2024). The inclusion of phonetic transcriptions provides the spoken form of words, allowing the transfer of knowledge across languages with different orthographic conventions but shared phoneme inventories. This has motivated approaches that exploit phonological similarity and cross-lingual transfer to extend models to under-resourced languages (Matogawa et al., 2024; Hammond, 2023; Li et al., 2022a; Deri and Knight, 2016). In situations where textual data is scarce, the incorporation of multimodal speech data to learn pronunciations has also been well researched (Ribeiro et al., 2023; Route et al., 2019; Aquino et al., 2019), facilitating cross-modal transfer to alleviate resource constraints.

Owing to the ever-growing scale of model requirements, including quantities of training data, a growing focus in G2P is computational efficiency. Recent works such as Kim et al. (2022) present methods to make G2P more appropriate for complex, real-time applications. They propose a non-autoregressive transformer for bilingual G2P, achieving accuracy comparable to autoregressive baselines whilst offering substantially faster inference, making them suitable for low-latency applications. Similarly, LiteG2P (Wang et al., 2023a) presents a highly parameter-efficient and parallelised hybrid connectionist temporal classification (CTC) model that incorporates G2P mapping dictionaries alongside maximum mapping length restrictions and masking rules to suppress invalid phoneme predictions for each grapheme.

A persistent challenge that faces G2P, and has measurable downstream impacts (particularly regarding puns and rhyme) is that of homographs. Homographs concern

words that have identical orthography yet differ in their pronunciation based on context (e.g., “take a bow” versus “a bow and arrow”). G2P systems capable of homophone disambiguation are becoming increasingly common, such as SoundChoice (Ploujnikov and Ravanelli, 2022), which leverages BERT-based semantic embeddings to incorporate contextual information, allowing for sentence-level disambiguation of homographs. Similarly, Han et al. (2024a) present an in-context knowledge retrieval approach to G2P, where an LLM is prompted with the sentence and its phonetic transcription, using semantic context to find the correct pronunciation of homographs.

Similar to the linguistic ambiguity posed by homographs, the prevalence of code-mixing in many language communities presents an increased need for multilingual G2P models capable of handling switches between languages and writing systems. To this end, Cho et al. (2020) present a multi-stage code-mixed G2P model for Korean which applies transliteration at the sentence level to handle agglutinative morphology. They decompose the task into stages, which includes detecting and separating non-Korean segments in *eojeols*³ and transliterating them into Korean pronunciations to improve the handling of code-mixed orthographies. Additionally, Zhu et al. (2022) demonstrated that byte-level models (e.g., ByT5) outperformed standard token-level models (e.g., mT5) trained on a large scale multilingual G2P dataset. Rather than relying on predefined subword tokenisation, ByT5 works directly on UTF-8 bytes which avoids segmentation errors and vocabulary mismatches, thereby reducing the phoneme error rate, especially in low-resource languages.

A further challenge to G2P is presented with underspecified orthographies, seen in languages such as Hebrew, Arabic, and Persian, where the given writing systems do not directly reflect all phonological contrasts present in a language’s spoken form. In particular, Persian (i.e., Farsi) has become a popular focus for G2P research due to its

³An *eojeol* is a fundamental linguistic unit of Korean, consisting of a sequence of characters separated from the rest of a text with whitespace.

underspecified orthography, in which short vowels (e.g., ezāfe particles) are not typically represented in writing. [Rahmati and Sameti \(2024\)](#) present *GE2PE*, a Persian G2P model to handle unwritten short vowels, ezāfe detection, and homograph disambiguation. GE2PE handles inputs at the sentence level using ByT5 to incorporate morphological, syntactic, and semantic contexts which would otherwise not be possible at the word-level. Similarly, [Qharabagh et al. \(2025b\)](#) show *Finglish* prompting (i.e., romanised Farsi) to improve G2P accuracy without additional training. They further introduce a sentence-level benchmark for polyphone disambiguation and ezāfe prediction, also releasing Kaamel-Dict, an open-licensed Persian G2P dictionary. Further research has also explored alternative disambiguation strategies for handling Persian homographs using intermediaries such as *Finglish* ([Qharabagh et al., 2025a](#); [Bertina et al., 2025](#)).

1.6.2 Language Modelling

Language modelling concerns the development of computational models of linguistic structure and meaning, most essentially via being able to predict the most probable next word given a previous sequence ([Chang and Bergen, 2024](#)). We first discuss more general literature that addresses the inclusion of phonetic and phonological knowledge in the language modelling process, abstracted away from any particular downstream application, covering aspects such as prompting approaches, phonetic/phonological knowledge probing, and general phonetics/phonology benchmarks.

Rather than relying entirely on retroactive knowledge injection, phonetic and phonological knowledge can be incorporated into language models from the offset via being included in the pretraining stage. To this end, [Goriely et al. \(2024\)](#) present a pretraining paradigm where LLMs are trained on continuous streams of phonemes by removing word boundaries and applying character-level tokenisation. This approach allows model to learn enhanced phoneme-based knowledge at the cost of a slight decrease in perfor-

mance on traditional language understanding tasks. Similarly, [Bunzeck et al. \(2024\)](#) compare grapheme- and phoneme-based Llama models and observe that grapheme models perform better overall on grammatical tasks, whereas phoneme models trade some performance for improved understanding of phonological tasks. Along the same lines, [Gale et al. \(2023\)](#) present BORT, a BART-based model pre-trained on orthographic and phonemic data through a self-supervised phoneme-to-grapheme (P2G) translation task that aims to restore orthography. Experimenting with pronunciation, spelling, and noise transforms, BORT is fine-tuned on clinically relevant tasks in speech language pathology and shows substantial gains over baselines. Performing pretraining on phonemic representations has also been explored for ASR ([Masumura et al., 2020](#)). [Tsvetkov et al. \(2016\)](#) introduce *polyglot* language models, which train a single recurrent neural network (RNN) over phonetic symbol sequences from multiple languages using shared representations. They condition the models on typological features of the target language to guide cross-lingual transfer. Their results show that polyglot models generalise better to held-out phonetic sequences than comparable monolingual models, and phonetic feature representations learned in multilingual settings are of higher quality than those trained in isolation. An analysis of the embedding space shows that the models capture features such as vowel length, nasality, and consonant-vowel distinctions, with similar sounds clustering across languages.

Abstracted from downstream tasks, the phonetic and phonological understanding of models can also be assessed through observing the characteristics learnt during training, and the development of explicit phonetic word embeddings. For example, [Gutscher and Pucher \(2024\)](#) investigate whether articulatory characteristics, dialectal differences, and speaker variances can be encoded in embeddings from audio segments, using quinphones⁴ in the embeddings of Austrian German speakers. They show that

⁴A type of phone representation that takes into account coarticulation from preceding and proceeding sounds.

CHAPTER 1. LITERATURE SURVEYS: FROM SOUNDS TO SYMBOLS

characteristic phones are captured effectively in the embedding space, allowing for better classification of dialects. Finally, [Malisz et al. \(2024\)](#) release PRODIS, a large multi-speaker Polish speech database with a phoneme-level language model, enabling the study of how prosody and discourse factors shape acoustic patterns. There also exists a range of BERT-based encoding models for speech tasks using phonetic features ([Nguyen et al., 2023](#); [Zhang et al., 2022](#); [Sundararaman et al., 2021](#)). [McLaughlin et al. \(2025\)](#) investigate how Llama 3.2 internally represents phonetic information despite not explicitly training on such data. By probing and using causal intervention techniques, they show that embeddings can encode recoverable phonetic features, identifying what they coin a “phoneme mover head” that can promote phonetic similarity in rhythm tasks. They further find that the model’s internal vowel space partially aligns with the IPA vowel quadrilateral, indicating that the model can converge on phonetic representations without explicit training. Furthermore, [Sharma et al. \(2021\)](#) present work on learning phonetic word embeddings. Using X-SAMPA mappings for English ([Wells, 1995](#)) and mappings proposed by [Malviya et al. \(2016\)](#) for Hindi, they develop a distance metric based on perceptual similarity that takes into account the salience of different sound changes (such as vowel changes being more salient than consonants). Phonetic representations of words within the resulting embedding space are shown to correctly group perceptually similar words.

Work has also been conducted on how to integrate phonetic knowledge into prompting strategies, alleviating the necessity of model fine-tuning. [Lei et al. \(2025\)](#) present a phonetic retrieval-augmented framework for ASR with LLMs, where entities are selected using normalised phonetic distance and provided as context. By including phonetically similar entities in the prompt, the model is able to re-rank and refine candidate ASR outputs, improving the performance of named entity recognition. [Nguyen et al. \(2025\)](#) examine the performance gap between Latin and non-Latin script languages,

demonstrating that even state-of-the-art multilingual models underperform on languages that are not written in the Latin script. By proposing a prompting method that includes phonemic transcriptions, it allows for language-independent representations that would be otherwise obscured by graphemic differences. This method, therefore, improved multilingual generation and translation, thereby reducing the performance disparity between languages with different writing systems.

With the increasing use of LLMs as the basis of phonetic and phonological tasks, it becomes crucial for such systems to be evaluated. [Suvarna et al. \(2024\)](#) introduce PhonologyBench, a benchmark that evaluates LLMs across three different tasks, including G2P, syllable counting, and rhyme generation. Overall, even the strongest models underperformed human baselines, with the largest discrepancy observed in syllable counting. This weakness extends even to simpler orthographic counting tasks such as counting the frequency of <r> in the word “strawberry” ([Xu and Ma, 2025](#)). [Zouhar et al. \(2024\)](#) present PWESuite, a benchmark for phonetic word embeddings that evaluates how models capture phonological relationships. Two levels of understanding are evaluated: *intrinsic* aspects, which include word retrieval and correlation with sound similarity and *extrinsic* aspects, which include cognate detection and sound analogies.⁵ Experiments show that whilst subword-based embeddings encode some phonetic structure, dedicated phonetic embeddings consistently outperform them across all tasks. Finally, [Gao et al. \(2024\)](#) present SLIQ-LL, a benchmark targeting spoken language learning by evaluating LLMs on conceptual and applied phonological questions. In this dichotomy, conceptual questions refer to things such as “*How many possible places of articulation are there?*”, whilst applied questions asked for this knowledge to be used in practice, such as asking “*What syllable carries primary stress in [word]?*”.

⁵For example, a sound analogy may be that /tm/ is to /dm/, what /sm/ is to /zm/, as a result of the change in voicing within the first segment.

Results demonstrated that LLMs can learn conceptual knowledge reasonably well but fail to apply it directly to tasks involving pronunciation, stress, or intonation.

1.7 Downstream Tasks

1.7.1 Poetry

In this section, we define poetry as an umbrella term encompassing traditional poetry such as sonnets (Spiller, 1992), haikus (Yasuda, 1973), and limericks (Vaughn, 1904), alongside the generation of song lyrics that rely on linguistic techniques and areas of phonetics and phonology such as rhyme, metre, alliteration, consonance, and assonance. Poetry and lyrics of these kinds exploit patterns in phonetics and phonology for literary effects such as sound symbolism (Loakman et al., 2024; Ramachandran and Hubbard, 2001; Sapir, 1929), and to meet constraints imposed by the given medium. For example, a haiku traditionally consists of three lines, in which the first line has five syllables, the second has seven, and the third again has five, whilst song lyrics are often designed to complement a musical backing track and are therefore length-restricted.

Owing to the integral nature of characteristics such as rhyme to poetry, most approaches focus on training phoneme-aware models for poetry generation. McGregor et al. (2016) present an early attempt at using a phoneme-based model for the task of poetry generation. They modelled the patterns of phoneme co-occurrence in poetry with an information-theoretic metric that uses phoneme transcriptions and syllable information from the CMUDict. Whilst this model was trained to have phonetic and phonological awareness to meet form constraints, the approach was at the expense of not maintaining coherent semantics. Hopkins and Kiela (2017) present a similar approach, also training a phoneme-level model. Using a combination of CMUDict transcriptions and finite state automata based on the G2P rules presented in Black

et al. (1998), they consider the output semantics by incorporating a stage to convert the learnt phoneme sequences back into words using an HMM and n-gram language model, overcoming the semantic incoherence that could be caused by incorrect homophone selection during transliteration. They increase the robustness of their approach to unseen poetry forms by decoupling the content and form aspects of poetry, using a cascade of weighted finite state transducers to identify syllable stress patterns, and therefore rhythm. Ghazvininejad et al. (2016) likewise explore meeting rhythm and rhyme constraints alongside topic-conditioned poetry generation. Despite the efficacy of incorporating linguistic knowledge, some approaches have also forgone the use of phonetically transcribed data, opting instead to train on lines of unlabelled orthographic text annotated with control codes (Keskar et al., 2019) associated with the metre and rhyme scheme, leading to moderate implicit learning (Ormazabal et al., 2022). Regarding rap lyric generation, Nikolov et al. (2020) incorporate an explicit rhyme enhancement step into generation, using BERT to mask the final word in a line and then predicting the 200 most likely replacements, selecting the candidate that maximises the sequence of overlapping vowels with the last word of the previous line. Other approaches to song lyric generation have focused on backwards decoding - predicting the rhyming words for each line before decoding the rest of a given line, therefore prioritising the presence of rhyme (Sun et al., 2023; Liu et al., 2022a; Xue et al., 2021; Van de Cruys, 2020).

A common limitation of NLP approaches to processing language such as poetry is the focus on broad subword tokenisation strategies that break up phonetically relevant sequences. One approach to remedying this is with character-level tokenisation strategies. For example, Tikhonov and Yamshchikov (2018) trained both orthographic character-based and phoneme-based LSTMs to generate poetry that imitates the styles of specific well-known English- (e.g., Shakespeare and Poe) and Russian-language (e.g., Pushkin and Esenin) poets - confirming that the phonetic characteristics of the output are

integral to correctly mimicking poetic voice. [Potash et al. \(2015\)](#) also train LSTMs for rap lyric generation, incorporating rhyme knowledge via adding special tokens to the start and end of a line or verse to aid the model in implicitly learning that final tokens in different lines are likely to rhyme. Similar to [Tikhonov and Yamshchikov \(2018\)](#), they incorporate additional features to aid in mimicking the styles of different artists. [Potash et al. \(2018\)](#) extends this work into a benchmark for rap lyric ghostwriting. Building upon the abilities of character-based models for phonetic tasks, [Belouadi and Eger \(2023\)](#) train an end-to-end token-free poetry generation model, ByGPT5, using rhyme and metre classifiers alongside a novel dataset of English and German poetry quatrains that was annotated with an alliteration metric.

Finally, as the fields of music generation ([Pathariya et al., 2024](#); [Zhu et al., 2023](#)) and singing speech synthesis ([Bai et al., 2025a](#); [Cho et al., 2021](#)) expand, so too has the generation of combined lyric and melody generation. For example, [Watanabe et al. \(2018\)](#) present a melody-conditioned lyrics model that exploits syllable counts to ensure alignment between melody and lyrics whilst taking into account the discourse-level semantic structure of lyrics. Further work to this end has been performed ([Chen and Lerch, 2020](#); [Sheng et al., 2020](#)), including for Portuguese ([Gonalo Oliveira, 2015](#); [Gonalo Oliveira et al., 2007](#)), Tamil ([Ramakrishnan A et al., 2009](#)), and Chinese ([Lee et al., 2019](#)). More recently, [Tian et al. \(2023\)](#) presented an unsupervised approach leveraging automatic alignment between syllables and musical melody. They use a plan-to-lyric approach, using GPT-2 to generate full words before retrieving the word’s syllable stress from CMUDict, identifying whether or not it meets the required stress-duration alignment constraint to be included in the final lyrics. Additional research has also focused on the reverse task, generating appropriate melodies for a pre-existing set of lyrics ([Yu et al., 2025](#); [Duan et al., 2023](#)).

In recent years, there has been increased work on benchmarking the performance of pre-trained language models in poetry generation. Abdibayev et al. (2021) present BPoMP, a poetry-focused variant of Warstadt et al. (2020)’s minimal-pair specific benchmark, BLiMP. BPoMP tests models’ abilities to distinguish between correct and corrupted limericks from the OEDILF project,⁶ the latter of which are created by shuffling rhyming words, shuffling entire lines, or replacing an element of a rhyming word pair with a non-rhyming substitute. Of the three tasks, the tested models struggled the most at identifying corrupted limericks when a rhyming word was replaced with a non-rhyming substitute, demonstrating that these pre-trained models did not understand the phonetic characteristics of the texts, focusing instead on the semantics, which remained constant in this condition. Additionally, Walsh et al. (2024) present a benchmark to test LLMs at identifying different poetic forms.

1.7.2 Humour

Whilst there are myriad forms of humour, features of phonetics and phonology are paramount to the forms of heterographic puns and tongue twisters. Heterographic puns comprise wordplay that exploits homophony and near-homophony (i.e., words that sound similar, yet are spelt differently). For instance, “*What do you call a fake noodle? An impasta*” relies on the phonetic similarity between “pasta” [pastə] and “-poster” [pɒstə]. To generate such humour, computational models must be able to identify words that share similar phonetic characteristics, despite having differing orthography. On the other hand, tongue twisters are difficult to articulate due to the rapid succession of phonemes that share similar phonological features (Clements and Ridouane, 2011; De Lacy, 2007; Klausenburger, 1970) which cause confusion when being realised as phones - for example, alternating between /s/ and /ʃ/ in “*She sells sea shells by the sea*

⁶<https://www.oedilf.com/db/Lim.php>

CHAPTER 1. LITERATURE SURVEYS: FROM SOUNDS TO SYMBOLS

shore”. Incorporating phonetic and phonological knowledge enables models to construct sequences that exploit these characteristics.

Midway between poetry/lyric generation and humour, tongue twister generation aims to create a grammatically valid sequence of words that features high levels of alliteration, generating humour through articulatory difficulty. [Keh et al. \(2023\)](#) present PANCETTA, a series of approaches involving training naive and phoneme-aware models in topic-to-twister and style-transfer settings. They include phoneme-level awareness by pre-training models on phonemic representations of WikiText data. Other approaches have also been taken to observe whether features of tongue twisters can be learnt implicitly from orthography ([Loakman et al., 2025b, 2023b](#)). Two phoneme-aware evaluation metrics were introduced in [Loakman et al. \(2025b\)](#) based on phoneme edit distance ([McCoy and Frank, 2018](#); [Mortensen et al., 2016](#)), which assessed the similarity of output phonemes in phonological feature space.

Elsewhere in computational humour, early work in pun generation used template-based joke schema that facilitated the easy injection of specialised linguistic knowledge, such as phonetic transcriptions ([Hong and Ong, 2009](#); [Ritchie et al., 2007](#); [Binsted, 1996](#); [Binsted and Ritchie, 1994](#)). For example, in order to create heterographic puns, [Binsted \(1996\)](#) and [Binsted and Ritchie \(1994\)](#) made use of a pre-existing corpus of English homophones, substituting parts of a word with semantically different yet phonetically similar elements, such as turning “tranquillizer” into “trunkquillizer” within a joke concerning elephants. Furthermore, [Stock and Strapparava \(2005\)](#) present HAHAcronym, a method for producing acronym-based jokes that leverages CMUDict to identify candidate words that can replace the original whilst maintaining phonetic similarity, such as using “intimidation” in place of “investigation” for the Federal Bureau of Investigation (FBI). Rule-based approaches remain in present works, such as [Dehouck and Delaborde \(2025\)](#), who present work on generating *Monsieur et Madame*

puns based on French given names. Their approach involves taking a phonetically transcribed name and mapping it to a French term from Wiktionary using a prefix tree. Further concessions are made for the discrepancy in written and spoken French, with phonological distinctions being relaxed in order to generate a larger pool of pun word candidates whilst using less accent-specific matching. More modern approaches to pun generation have relied on neural architectures, combining phonetics (e.g., homophone pairs) with humour mechanisms such as incongruity (Chen et al., 2024; Sun et al., 2022b; Tian et al., 2022b; He et al., 2019).

Baluja (2025) observe that prompting LLMs with synthetic speech of puns being read aloud alongside the original text results in an improvement in generated pun explanations. This suggests that exploiting the audio of the heterograph’s pronunciation, rather than identifying similar-sounding words through orthography alone, is beneficial. In the case of homographs, this is likely the result of the homonym subset, whereby two words are spelt the same, yet both their meaning and pronunciation differ (e.g., “*I like to read*”, /i:ɪd/ versus “*Yesterday, I read the book*”, /ɪɛd/). Furthermore, recent work has shown that LLMs still have considerable difficulty with phonetically motivated humour, struggling more to both generate (Xu et al., 2024) and explain (Zangari et al., 2025; Loakman et al., 2025c) heterographic puns than homographic puns.

1.7.3 Translation

Finally, we turn to the topic of translation. In particular, phonetic and phonological knowledge is incorporated into translation systems that are specifically designed to process texts of the formats previously discussed. In such cases, performing translation such that it can be understood by a target audience requires not only the preservation of meaning, but also phonetic and phonological constraints, such as maintaining the same metrical structure or location of rhyming couplets, which may be achieved through

techniques such as constrained decoding (Bogoychev and Chen, 2023; Nakazawa et al., 2021; Susanto et al., 2020; Post and Vilar, 2018).

Early work into the translation of loan words found that including both grapheme- and phoneme-based methods helped mitigate issues that each faced when applied separately. Oh et al. (2006) showed that whilst grapheme-based models could capture regularities in spelling, they often failed when pronunciation diverged from what would be expected through orthography. In such cases, phoneme-based models showed better performance in preserving sound information that was not encoded in graphemes. For example, when the English loan word “data” (/dɛɪtə/ or /dætə/) is transliterated, in Korean it becomes “de-i-teo” and in Japanese it becomes “de-e-ta”, reflecting each language’s phonology. Tian et al. (2022a) present a transliteration model for English to Arabic proper nouns that utilises a phonemic memory module to guide the mapping between English phonemes and Arabic graphemes during decoding. This results in outputs being biased towards phonetically similar transliterations, making them more accurate in transliterating names without changing the pronunciation.

Liu et al. (2019) integrate textual and phonetic embeddings to improve robustness to homophones, whilst Mhaskar et al. (2024) present a method that aligns phoneme counts between source and target sentences. With the introduction of phoneme-count compliance, Mhaskar et al. (2024) demonstrate that their approach significantly outperforms baselines for English-Hindi translation. In addition, Rosales Núñez et al. (2019) present a phonetic normalisation approach to improve the translation of noisy, user-generated content in French. They utilise character-based G2P, with the output being used to generate phonemic candidates for translation. This allows for non-standard spellings, accent variations, and typographical errors to be corrected.

Humour translation remains an actively investigated area, having dedicated shared tasks such as CLEF’s JOKER (Ermakova et al., 2026). Taylor et al. (2025) present work

on translating puns from English to French, presenting the highest scoring submission. To achieve this, they use contrastive learning with phonetic-semantic embeddings learnt from the databases Lexique (New et al., 2004) and PanPhon (Mortensen et al., 2016). Chakrabarty et al. (2021) show that a multilingual fine-tuned poetry model significantly outperforms models 30x larger that were not exposed to poetry during training. Whilst the results focus on semantics, the findings also suggest links to phonetics and phonology due to poetry frequently relying on rhyme. Similar links have been identified in the processing of Sanskrit and Persian (Sandhan et al., 2025; Khanmohammadi et al., 2023).

1.8 Conclusion

In this survey, we have outlined a range of key foundational and application-based areas where incorporating knowledge of phonetics and phonology is either integral or of substantial benefit. We have focused primarily on unique use cases of phonetic and phonological knowledge in these domains, aiming to provide the reader with a concise, accessible introduction to a subfield of NLG where linguistic knowledge is paramount. Suggested future work for research in this area can be found in Section 9.3.

1.9 Limitations

We aim for this work to act as an overview of the ways in which the NLG community has incorporated knowledge from the areas of phonetics and phonology into different tasks. As a result, this survey does not aim for high levels of depth in any specific subdiscipline, instead we refer the reader to relevant recent works in the respective areas of interest for deeper architectural insight (especially where these details do not pertain to the incorporation of phonetic/phonological knowledge itself). The purpose of

this survey is to act as an overview for works incorporating knowledge from a specific branch of linguistics that is less frequently considered in common generation tasks.

Although we have striven to include a wide range of works in the target areas we explore, it remains possible that we missed some works in the area, and this survey is not to be taken as a definitive list of all of the research in NLG that has incorporated phonetic/phonological knowledge.

Additional Acknowledgments

Joseph James is supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [grant number EP/S023062/1].

1.10 Appendix

1.10.1 Survey Inclusion Criteria

The following seed terms were used to determine paper inclusion in the followed backward snowball approach: *G2P*, *grapheme to phoneme conversion*, *poetry generation*, *rap generation*, *lyric generation*, *melody generation*, *pun generation*, *joke generation*, *tongue twisters*, *phonetics*, *phonology*, *phonological*, *poetry translation*, *joke translation*, *phonetic translation*.

Survey II:

Who's Laughing Now? An Overview of Computational Humour Generation and Explanation

INLG ([Loakman et al., 2025d](#))

Tyler Loakman^a, William Thorne^a, and Chenghua Lin^b

^aDepartment of Computer Science, University of Sheffield, UK

^bDepartment of Computer Science, The University of Manchester, UK

Contextual Statement

This survey addresses the secondary research theme of the thesis in the form of humour generation and explanation. In Chapter 3 and Chapter 4, the humorous language of tongue twisters is explored as a way through which to study model understanding of phonetic similarity and sound repetition, whilst Chapter 5 explores humour explanation including heterographic puns that rely on phonetic similarity. The present survey offers a wider exploration of the domain of generative computational humour processing as a whole in order to situate the work included in later thesis chapters.

Abstract

The creation and perception of humour is a fundamental human trait, positioning its computational understanding as one of the most challenging tasks in natural language processing (NLP). As an abstract, creative, and frequently context-dependent construct, humour requires extensive reasoning to understand and therefore create, making it a pertinent task for assessing the common sense knowledge and reasoning abilities of modern large language models (LLMs). In this work, we survey the landscape of computational humour as it pertains to the generative tasks of creation and explanation. We observe that, despite the task of understanding humour bearing all the hallmarks of a foundational NLP task, work on generating and explaining humour beyond puns remains sparse, whilst state-of-the-art models continue to fall short of human capabilities.

1.11 Introduction

Humour serves as a foundational element of human communication, acting as a way through which to express emotion, build interpersonal relationships, and experience levity and entertainment (Ritschel and André, 2018). However, humour may arise from a myriad sources (Dyner, 2009), including simple, innocuous wordplay such as puns, all the way to deeply contextualised topical references that require layered reasoning to both create and interpret (Highfield, 2015; Laineste, 2002). The position of humour as a distinctly human experience, in addition to the challenges its processing presents, even to humans (Bell and Attardo, 2010; Mak and Carpenter, 2007; Wierzbicki and Young, 1978), makes generative computational humour tasks such as joke creation and explanation a formidable domain for analysing the common sense reasoning capabilities of modern large language models (LLMs).

1.11.1 Related Work

Several surveys have examined different aspects of computational humour. Early foundational work by Ritchie (2001) provides a comprehensive overview of the emerging field of computational humour. More recent reviews have focused on specific subdomains of computational humour. Ramakristanaiah et al. (2021) and Kalloniatis and Adamidis (2024) survey humour recognition and detection, whilst Kenneth et al. (2024) review humour style classification. Ganganwar et al. (2024) explore sarcasm and humour detection in code-mixed Hindi, and Nijholt et al. (2017) cover the unique domain of humour in human-computer interaction. The most comparable works to ours are Amin and Burghardt (2020), who present a survey of text-based computational humour generation, and Nguyen and Ng (2024), who survey works on meme understanding. Our survey provides an up-to-date account of the field in the age of LLMs, an additional

focus on humour explanation, and suggestions for future work based on real-world ethical considerations, rather than solely technical innovations.

1.11.2 Survey Outline

This survey is structured around the two primary generative tasks in computational humour. In §1.13 we explore humour generation, broken down by humour type, whilst in §1.14 we explore humour explanation, broken down into explanation through classification (§1.14.1) and natural language explanation (§1.14.2).

1.12 The Role of Computational Humour

Significant effort and resources are currently being invested into enhancing the reasoning capabilities of language models (Wu et al., 2024b; Yin et al., 2024; Servantez et al., 2024). Reasoning or thinking models embed established prompting techniques such as *chain-of-thought* (Wei et al., 2023) or *tree of thought* (Yao et al., 2023) into the generation process, typically producing a verbose stream of consciousness in order to elicit the necessary steps to solve complex tasks. Scaling test-time compute in this way has seen a significant improvement on many benchmarks (Geiping et al., 2025; Snell et al., 2025); however, the current status quo relies heavily on more technical and formalised reasoning tasks such as code generation and arithmetic. Whilst results in such domains can be deterministically validated (Wang et al., 2024a,c), and large quantities of samples can be generated (Xu et al., 2025), verbal and common-sense reasoning remains a fundamental task with wider applicability to the end-users of such technologies (Trichelair et al., 2019). We argue, therefore, that computational humour processing is an essential task for addressing the limitations in how models process natural language pragmatics and cultural knowledge.

CHAPTER 1. LITERATURE SURVEYS: WHO’S LAUGHING NOW?

To emphasise the existing focus on mathematical reasoning, of the benchmark suites used in recent releases (i.e., Kimi K2⁷ and Grok4⁸), nine focus on STEM problem solving: GPQA (Rein et al., 2024), USAMO (Petrov et al., 2025), HMMT,⁹ AIME-25,¹⁰ LiveCodeBench (Jain et al., 2024), SWE (Jimenez et al., 2024), OJBench (Wang et al., 2025c); or structured reasoning: ARC-AGI-2 (Chollet et al., 2025), and ACEBench (Chen et al., 2025). On the other hand, only Tau2 (Barres et al., 2025) touched upon communicative competence, albeit through customer service scenarios.

Computational humour presents a uniquely demanding arena for model evaluation that complements the existing array of benchmarks. Lacking from existing options are foundational elements of humour, such as phonetic understanding and pragmatic inference. Consider the joke in which a husband and wife are solving a crossword. The joke centres on the husband giving clues such as “*Emphatic no, five letters*”, and “*Pistol, 3 letters*”, resulting in guesses of “*never*” and “*gun*”. This continues until the string of words guessed by the wife reads “never gun ugh give ewe Up”. Whilst meaningless in isolation, the spoken realisation of the sequences equates to “*Never gonna give you up*”, an instance of the popular *Rickrolling* internet phenomenon.¹¹ The creation and comprehension of this joke depends on the recognition of Rick Astley’s song of the same title, the cultural phenomenon of *Rickrolling*, and an understanding of the phonetic similarity present in homophones.¹²

Moreover, humour frequently depends on understanding what was *not* said or inferring the opposite of explicit statements (Yus, 2013), which directly conflicts with the semantic-similarity-based retrieval systems that would typically be employed when

⁷<https://moonshotai.github.io/Kimi-K2/>

⁸<https://x.ai/news/grok-4>

⁹<https://www.hmmt.org/>

¹⁰<https://artofproblemsolving.com>

¹¹<https://en.wikipedia.org/wiki/Rickrolling>

¹²This is made especially hard as there is not a 1-to-1 mapping, with “gun” and “ugh” taking the place of “gonna”.

addressing jokes that reference post-training events (Barnett et al., 2024; Akila and Jayakumar, 2014). It is for these reasons that we assert the importance of humour understanding as a subdomain of NLP in the LLM age. The field of computational humour remains a fruitful area for continued work (Ignat et al., 2024; Lima Inácio and Gonçalo Oliveira, 2023); one that is critically overlooked and under-researched.

1.12.1 Theories of Humour

Humour, as a uniquely human experience, has been the focus of thinkers for centuries, becoming the subject of heavy debate. As such, several prominent theories have arisen that attempt to explain the perception of humour, a subset of which are presented below:

- **Relief Theory** suggests that humour, and particularly laughter, is the result of releasing psychological energy (Freud, 1963; Spencer, 1875; Kant, 1790).
- **Superiority Theory** posits that the experience of humour and comedy is born out of the perception that one individual is superior to another, thus making the inferior individual the subject of humour (Aristotle, 1902; Plato, 1892; Hobbes, 1660).
- **Incongruity Theory** states that humour is the perception of something that conflicts with established mental patterns and expectations, therefore being incongruous (Morreall, 2024; Tu et al., 2014).
- **Benign Violation Theory** postulates that humour arises from situations that are simultaneously harmless (i.e., benign) and are incongruous with expectations (i.e., violating a norm) (McGraw et al., 2012; McGraw and Warren, 2010).

Whilst the superiority and relief theories offer accounts of the experience of humour, they do not present simple interpretations that can be easily formulated linguistically and computationally. It is for this reason that, of the approaches to humour processing that are directly grounded in theory, incongruity and the general sense of norm violation remain essential elements (Tian et al., 2022b; He et al., 2019; Valitutti et al., 2013).

We present examples of textual linguistic humour in Figure 1.3 and an example of a multimodal meme in Figure 1.4.

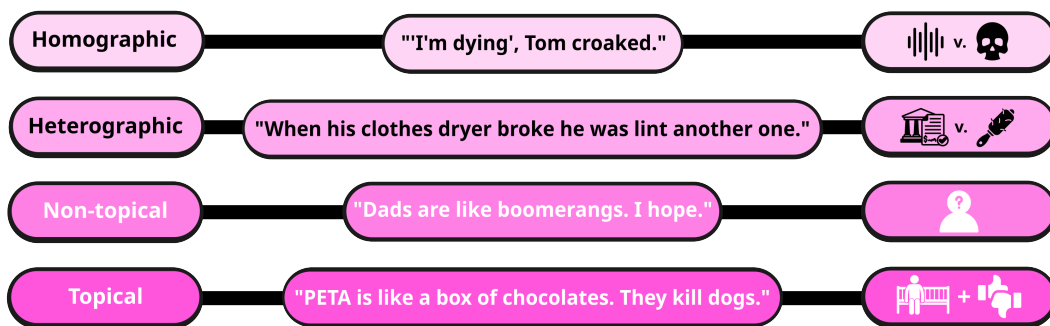


Figure 1.3: Examples of the four broad textual joke types from *Lookman et al. (2025c)*.

1.13 Humour Generation

The ability to compose novel jokes requires an implicit understanding of the cognitive mechanisms that underlie humour. As outlined in the *incongruity* and *benign violation* theories of humour, a necessary feature of humorous language is that it violates an expectation within the reader, either in the form of cultural or situational norms (e.g., topical and contextual humour), or linguistic norms (e.g., puns). This core property poses a challenge to computational approaches and to creative language generation more broadly: the language modelling objective aims to maximise the log likelihood of an output sequence, thereby working against the very incongruity that humour

demands (Loakman et al., 2025b). The following subsections explore historical attempts to address or incorporate this contention for pun generation.

1.13.1 Pun Generation

Pun generation has dominated research in computational humour, spanning from early rule-based approaches (Lessard, 1992) to contemporary neural methods. The historical prevalence of puns is largely a result of their computational tractability and theoretical grounding in Incongruity Theory, providing both an entry point and capacity to produce large quantities of training/evaluation data. Early works, such as JAPE (Binsted, 1996; Binsted and Ritchie, 1994) and STANDUP (Ritchie et al., 2007), leverage WordNet (Fellbaum, 1998) for semantics and Unisyn¹³ for phonetic similarity. However, such classical approaches rely heavily on fixed schema, limiting overall creativity. For a deeper coverage of early literature in pun generation, we refer the reader to the prior surveys of Ritchie (2001) and Amin and Burghardt (2020).

The remainder of this section is split according to the distinction of heterographic and homographic puns (Redfern, 1987), including their combination.

1.13.1.1 Homographic Pun Generation

Homographic puns exploit polysemy, the phenomenon of multiple meanings existing for a single word, to create humour through semantic ambiguity. Early work in automatic pun generation focused on this category due to the accessibility of semantic resources like WordNet (Miller, 1994).

Yu et al. (2018) present the first neural approach to homographic pun generation, training a conditional encoder-decoder LSTM on unlabelled Wikipedia text to create sentences that could support the semantics of two words simultaneously. This model

¹³<https://www.cstr.ed.ac.uk/projects/unisyn/>.

then uses a constrained beam search algorithm to jointly decode the two distinct senses of the same word, generating puns without requiring any pun-specific training data.

In a follow-up work, [Yu et al. \(2020\)](#) explore pun generation through a lexically constrained rewriting approach that first identifies constraint words supporting semantic incongruity for a sentence, then rewrites it with explicit positive and negative constraints. Their method achieved state-of-the-art results in both automatic and human evaluations.

Building on this, [Luo et al. \(2019\)](#) introduced Pun-GAN, a GAN-based model ([Goodfellow et al., 2014](#)) that employed a discriminator module with word-sense disambiguation capabilities to assess how well a generated sentence supported the polysemy of the target homographic pun word, aiming to maximise semantic ambiguity. The generator was trained via reinforcement learning using the discriminator’s output as a reward signal to encourage the production of sentences that could support two word senses simultaneously. However, the key shortcoming of this approach was its tendency to produce generic outputs that prioritised ambiguity over general pun quality.

AmbiPun ([Mittal et al., 2022](#)) takes as input a homograph with two distinct word senses and proposes that ambiguity comes from context rather than the pun word itself. The approach first produces a list of related concepts through a reverse dictionary, then utilises one-shot GPT-3 to generate context words from both concepts before generating puns that incorporate these contextual elements. They achieve a 52% success rate in human evaluation, significantly outperforming baselines but remaining tied to a complex, multi-staged pipeline.

1.13.1.2 Heterographic Pun Generation

Heterographic puns present additional challenges due to the requirement of modelling phonetic similarity between different surface forms whilst maintaining semantic coherence. These puns exploit words that sound alike but are spelt differently, requiring

systems to understand both phonetic relationships and semantic contexts (Kao et al., 2016).

He et al. (2019) generate heterographic puns through an unsupervised retrieve-and-edit framework based on the *local-global surprisal principle*. Given a pair of homophones (e.g., “died” and “dyed”), they first retrieve candidate sentences containing the alternative word from a corpus, replace the alternative word with the pun word, and insert a semantically related topic word to the start of the pun. The approach is a direct instantiation of Incongruity Theory: the foreshadowing topic word creates a strong association with the pun word in the global context whilst maintaining the local context of the alternative word in the sentence.

1.13.1.3 Combined Pun Generation

More recently, Tian et al. (2022b) proposed a unified framework that addresses both homographic and heterographic pun generation by incorporating three key linguistic attributes: ambiguity, distinctiveness, and surprise. Their approach demonstrates that principled integration of humour theories can improve generation quality across different pun types.

Chen et al. (2024) propose a multi-stage curriculum learning approach using direct preference optimisation (DPO) (Rafailov et al., 2023) to align models with the ability to create valid linguistic structures in support of both homographic and homophonic pun creation. The authors ultimately released the ChinesePun dataset of Chinese humour, presenting one of the few works that does not focus predominantly on English.

Recent systematic evaluation by Xu et al. (2024) provides evidence that large language models struggle considerably more with the generation of heterographic puns than homographic puns. This difficulty likely stems from the need to infer phonetic characteristics of words, which is challenging for models operating primarily on text-

based representations (Baluja, 2025). The authors additionally identify what they term “lazy pun generation”, whereby models incorporate both senses of the pun word in a single generation (e.g., “The sailor’s pay was docked after he struggled to dock on time”). This phenomenon nullifies the intended effects of surprisal, eliminating any ambiguity that would require cognitive effort to resolve.

Sun et al. (2022b) demonstrates a departure from stand-alone pun generation, releasing CUP, a novel task for context-situated puns. CUP integrates puns more naturally into real-world conversations by demanding contextual awareness. The authors extract keywords from context sentences using RAKE (Rose et al., 2010) and use a T5 model to create sentences that support both meanings of the pun word. Logically, this approach can be viewed as a hybrid of Mittal et al. (2022), He et al. (2019), and Luo et al. (2019).

1.13.2 Non-Pun Humour Generation

Whilst puns have received the most attention in computational humour research, broader forms of humour generation present a greater challenge due to their association with cultural knowledge, timeliness, and more complex incongruities.

Goel et al. (2024) proposed an approach combining template extraction and infilling using BERT (Devlin et al., 2019) with LLMs (specifically GPT-4, OpenAI et al., 2023, and Zephyr-7B, Tunstall et al., 2024) to generate set-up and punchline style jokes (e.g., “*Why did the chicken cross the road? To get to the other side*”). To achieve this, tokens from jokes are masked, with BERT attention weights used to determine which elements of a given joke can be masked, maintaining the essential structure, whilst removing overly topic-specific terms. As a result, the system learns to create joke structures that can then be filled to create novel joke instances.

On the other hand, [Chung et al. \(2024b\)](#) present UNPIE, a benchmark to assess the understanding of vision language models (VLMs) when reasoning about lexical incongruities. To achieve this, they take as input written puns and generate an image to serve as a visual representation of the pun, demonstrating that such visual clues may be beneficial in pun understanding tasks, helping to identify the location of a pun in a given text.

[Horvitz et al. \(2020\)](#) focus on the generation of satirical headlines, a type of language that is both humorous and dry, being used to criticise individuals and entities on complex topics. In doing so, they tackle the linking of relevant real-world knowledge from sources such as Wikipedia and CNN, to relevant satirical headlines from TheOnion,¹⁴ which are then used to fine-tune GPT-2 ([Radford et al., 2019](#)).

[Tian et al. \(2021\)](#) explore humour through the lens of hyperbole with HypoGen. To do so, they curate a dataset of hyperbolic phrases following the “*so [X] that [Y]*” pattern (e.g., “*My personality is so dry that a cactus flourishes inside*”), using variants of CoMET models ([Bosselut et al., 2019](#)) to learn the commonsense and counterfactual relationships present in hyperbolic language to assist generation.

Finally, in recent years, there has been a growing body of work in the humour-adjacent domain of tongue twister generation - texts where entertainment and humour arise from mispronunciations stemming from complex phonetics. Such approaches involve training keyword-to-twister and style-transfer models for tongue twister generation, either via training on graphemes ([Loakman et al., 2025b, 2023b](#); [Keh et al., 2023](#)) or phonemes ([Keh et al., 2023](#)). Research in this domain has also highlighted the benefit of incorporating explicit phonetic and phonemic information into the generation of language formats that rely on such characteristics.

¹⁴<https://theonion.com/>

1.14 Humour Explanation

“...if it can say why a joke’s funny, it really does understand”

- Geoffrey Hinton.¹⁵

The above quote from Hinton (“the Godfather of AI”) refers directly to the use of joke explanation as a milestone achievement in the marketing of Google’s PaLM LLM (Chowdhery et al., 2023). Detection approaches, whilst offering objective and easily automated evaluation, remain susceptible to statistical flukes. Explanation generation eliminates this vulnerability through a vanishingly low probability of accidentally producing coherent comedic analysis. Moreover, the act of explanation provides valuable insight into the inner workings of what still exist as nearly opaque black boxes. This value comes at the expense of significantly more challenging, expensive and labour-intensive evaluation methodologies. In its present state, the field of humour explanation remains critically understudied. To provide necessary context, we draw upon select literature covering humour classification and detection.

1.14.1 Explanation through Classification

Humour explanation, although understudied, is not a completely isolated evaluation of comprehension. Whilst not encompassing the full scope of a natural language explanation, humour detection tasks models with identifying the linguistic traits common to humour. As such, we treat humour detection as a precursor to explanation.

The majority of existing works on humour explanation do not focus on providing textual natural language explanations, instead focusing on other indications of a joke’s source, such as word senses or identifying the type of humour being expressed.

¹⁵From the 16th June 2025 episode of The Diary of a CEO podcast. See <https://youtu.be/giT0ytynSsg?si=00iN3DM2Fv58fp8o&t=4460>.

CHAPTER 1. LITERATURE SURVEYS: WHO’S LAUGHING NOW?

Miller et al. (2017) presents an overview of Task 7 from SemEval 2017, which concerned detection and interpretation (i.e., classification) of puns. Specifically, the pun interpretation task consisted of assigning word sense keys from WordNet (Fellbaum, 1998) to the punning word contained in a text. One approach to the task, taken by Oele and Evang (2017), splits a given text into two parts at all possible locations, with the split where both parts have low semantic similarity being where the two meanings of the pun word are best separated. A word sense disambiguation (WSD) model is then used to retrieve the relevant senses. Similar techniques based on mapping and comparing the semantics of different partitions of the text with the pun word are shown in multiple submissions (e.g., Hurtado et al., 2017; Indurthi and Oota, 2017). Whilst assigning sense keys to a pun word offers a minor explanation to a user, the requirement for the pun word to be pre-identified nullifies the utility of such systems in practice.

Palma Preciado et al. (2024) present an overview of the JOKER shared task from CLEF 2024, where Task 2 aimed to classify humorous texts based on their genre and the linguistic techniques used, including irony, sarcasm, exaggeration, incongruity/absurdity, self-deprecation and wit/surprise. In total, they received 54 submissions to the shared task, with techniques ranging from training BERT classifiers (e.g., Narayanan et al., 2024) and ensembles of classic machine learning classifiers (e.g., Bartulović and Váradi, 2024), to zero-shot prompting of LLMs like GPT-4 (e.g., Wu et al., 2024a).¹⁶

However, explanation through classification presents a series of issues. Firstly, and most obviously, whilst such approaches can demonstrate a model’s understanding of humour, assigning the correct label to a given joke is unlikely to present a human user with valuable information that would aid their interpretation. Secondly, classification requires the creation of a taxonomy under which to categorise different jokes. Owing to the complexity of humour, high-quality, robust taxonomies are challenging to create. For

¹⁶See Palma Preciado et al. (2024) for a full overview of submissions.

instance, Task 2 from JOKER (Palma Preciado et al., 2024) has considerable overlap between joke categories. A prerequisite for both irony and sarcasm is the violation of a norm, whilst *incongruity* is presented as a separate category. This is additionally true of *wit* (a subjective judgement of cleverness and novelty) and *surprise*.

1.14.2 Natural Language Explanation

Whilst explanation through the lens of classification is a valid approach for simple joke formats such as puns, more complex humour formats such as more esoteric, context-dependent jokes, benefit from natural language explanations (in addition to being more friendly to the proposed end-user). Natural language explanations rectify the coarseness of explanation through classification. Such approaches do not require a pre-defined taxonomy of humour types to identify, but instead test the overall abilities of models to provide tailored, specific explanations for every humorous text.

Xu et al. (2024) generates natural language explanations for puns using chain-of-thought prompting with a range of LLMs. They find that most LLMs can identify the punning word in both heterographic and homographic examples, but most models struggle with correctly identifying the alternative intended meaning of heterographic puns, which rely on phonetic similarity. The authors identify a series of mistakes commonly made by models, including failure to recognise the joke as a pun, incorrectly identifying the alternative senses of the pun word, and failing to provide the meanings of the pun word, but without contextualisation into a natural language explanation.

Loakman et al. (2025c) extend these findings and use a range of LLMs on a range of joke formats, including homographic puns, heterographic puns, long-form humour, and topical jokes from Reddit. Their results further confirm that LLMs struggle with heterographic puns (owing to their lack of phonetic knowledge), but additionally demonstrate that longer incongruity-based jokes and jokes that rely on

CHAPTER 1. LITERATURE SURVEYS: WHO’S LAUGHING NOW?

esoteric topical knowledge present even more difficulty to LLMs. Whilst a zero-shot non-chain-of-thought approach is used, the authors find that the Llama models distilled from Deepseek R1, are among the worst-performing. They hypothesise that this is a result of the ambiguity in real-world references within topical humour, leading to early misunderstandings being propagated through the reasoning process. A further, smaller-scale evaluation of ChatGPT’s humour explanation ability is presented by [Jentzsch and Kersting \(2023\)](#). Similarly, [Wang et al. \(2025b\)](#) investigate the ability of LLMs to explain “drivelology”, a linguistic form characterised as “nonsense with depth”, incorporating humour alongside other elements such as sarcasm, irony, and tautologies. Whilst they investigate the ability for LLMs to correctly categorise the type of “drivel” being used, they additionally perform zero-shot explanation generation and likewise find that models struggle significantly with creating high-quality explanations.

Identifying the common failure of models to understand jokes where phonetic characteristics play a vital role (e.g., heterographic puns), [Baluja \(2025\)](#) investigate whether access to speech audio of the joke being read aloud leads to improved performance. They assessed the ability of Gemini 1.5 ([Team, 2024](#)) to explain jokes with and without access to text-to-speech readings, demonstrating approximately a 2.5% to 4% improvement in explanation performance. Whilst moderate, such findings indicate the affordances provided by multimodal prompting in language forms that rely heavily on modalities that are absent from text (i.e., pronunciations), suggesting room for further performance gains alongside improvements in the fusion of modalities.

In the realm of multimodal humour, [Hessel et al. \(2023\)](#) present work on the explanation of visual jokes in the form of humorous captions from the New Yorker Cartoon Caption Contest. In such a task, models must understand the visual cartoon in order to disambiguate pun words or correctly establish the key incongruity, owing to the text alone being ambiguous. For instance, one cartoon presents a barbershop



Figure 1.4: *An example of a meme, a form of multimodal humour that incorporates visual elements, often alongside text. In this instance, the humour arises from the absurdity of the belief that asking about cheese is political, whilst being enhanced by the confused T-pose Kermit the Frog render. The original meme has been edited to remove expletives.*

with a hole in the roof and a spring coming out of a barbershop chair, with the caption reading “He’ll be back”. This phrase typically means that someone will return to an establishment even if they were dissatisfied (e.g., they provide an essential service or the individual’s criticism was unreasonable). However, in this instance, the visual cues reveal that “He’ll be back” is a literal statement, referring to the effects of gravity on the man who was ejected from his chair by a spring. To evaluate this, [Hessel et al. \(2023\)](#) fine-tune GPT-3 on human explanations, as well as perform 5-shot prompting with GPT-4. The results showed that access to visual information from the cartoons resulted in a better explanation in 84.7% of cases (via human preference judgements). However, whilst in-context learning with 5-shot GPT-4 outperforms fine-tuned GPT-3, it was shown that human-authored explanations are still preferred in 68% of instances.

On the other hand, the lion’s share of work in multimodal humour explanation specifically investigates memes, a type of media (typically an image) that is copied

and propagated rapidly across the internet, often comprising text captions and related imagery (see Figure 1.4 for an example). Hwang and Shwartz (2023) present MemeCap, a dataset tailored to meme understanding, consisting of 6.3k memes from the r/Memes subreddit alongside crowdsourced captions that explain the semantics the meme is trying to convey. Additionally, Khan et al. (2024) present a dataset of 13k+ memes, including those with audio. In this instance, an LLM pipeline was utilised to generate explanations of the memes, which were then refined by human annotators. Interestingly, the authors aim to generate explanations that achieve the aim of entertaining and amusing readers (therefore being somewhat humorous themselves), rather than being strictly objective accounts of the humour.

Park et al. (2024) acknowledge that author’s intent contributes significantly to meaning, ergo perception. They define the task of *intent description generation*, accompanied by a dataset of 950 samples that are annotated with perceived intention and the necessary context. We believe that intent-aware systems are a compelling direction for future study, which we explore further in §9.3.6.

Furthermore, Agarwal et al. (2024) present MemeMQA, a multimodal Q&A dataset for asking questions regarding the content of memes in order to better understand them. From this, they develop ARSENAL, a multimodal meme understanding pipeline that uses LLMs to reason about a given meme in relation to a specific question.

The site Know Your Meme¹⁷ tracks the provenance of memes as they develop, making it a wealth of knowledge regarding the origin and growth of memes over time. From this, Tommasini et al. (2023) built the Internet Meme Knowledge Graph, comprising two million edges to represent the semantics of multimodal memes.

In a more directly application-based setting, Jha et al. (2024) explore the explanation of memes being used explicitly for purposes of cyberbullying. Using the MultiBully

¹⁷<https://knowyourmeme.com/>

dataset (Maity et al., 2022), they add annotations to highlight pertinent visual and linguistic aspects of a meme that clarify its intent as cyberbullying, allowing the training of enhanced bullying detection models.

1.15 Conclusion

In this work, we have explored the current landscape of computational humour generation and explanation, outlining the approaches taken in existing work and identifying promising future directions. We posit that computational humour processing remains an unsolved task with a wide range of real-world applications, as well as being one of the most promising yet overlooked domains through which to evaluate the reasoning abilities of modern LLMs. We make the case that future work should focus on increasing the breadth of humour studied, in addition to explicitly considering the ethics of computational humour. Suggested future work for research in this area can be found in Section 9.3.

1.16 Limitations

We position this work as an overview of computational humour generation and explanation. As a result, we focus on the breadth of research available, rather than presenting in-depth accounts of the exact technical novelty of existing works. Additionally, whilst we aim to be comprehensive with our overview, some instances of relevant research may have been missed. Furthermore, we have not extensively explored other areas of computational humour processing, such as automatic metrics and human evaluation paradigms, available datasets, or approaches to humour detection.

1.17 Ethics Statement

Relating to the ethical considerations explored in this work, whilst we believe that humour generation and explanation are worthwhile pursuits, we acknowledge and appreciate the stances taken by other individuals. This relates specifically to ethical considerations surrounding the generation of creative language in any form, the generation of humour that is potentially offensive, and whether or not explaining potentially offensive humour equates to an endorsement of the content of such humour. Such decisions should continue to be a source of discussion in the NLP community as a whole, and we encourage individual researchers working in these domains to explicitly state their stance in published works.

Part II

Human Evaluation

Chapter 2

The Iron(ic) Melting Pot:

Reviewing Human Evaluation in Humour, Irony, and Sarcasm Generation

Findings of EMNLP ([Loakman et al., 2023a](#))

Tyler Loakman^a, Aaron Maladry^b, and Chenghua Lin^c

^aDepartment of Computer Science, University of Sheffield, UK

^bLT³, Ghent University, Belgium

^cDepartment of Computer Science, The University of Manchester, UK

Contextual Statement

This chapter addresses RQ1, investigating the impact of evaluator demographics on the interpretation of the intensity of subjective language, whilst assessing existing research regarding the transparency of human evaluation reporting. By presenting a set of standardised reporting criteria for the evaluation of subjective language, the insights from this chapter informed the design of human evaluation protocols and reporting transparency in subsequent chapters concerning subjective and creative phonologically-motivated language, including tongue twisters (Chapter 3 and Chapter 4) and jokes (Chapter 5).

Abstract

Human evaluation is often considered to be the gold standard method of evaluating a natural language generation (NLG) system. However, whilst its importance is accepted by the community at large, the quality of its execution is often brought into question. In this work, we argue that the generation of more esoteric forms of language - humour, irony, and sarcasm - constitutes a subdomain where the characteristics of selected evaluator panels are of utmost importance, and every effort should be made to report demographic characteristics wherever possible, in the interest of open science. We support these claims with an overview of each language form and an analysis of examples in terms of how their interpretation is affected by different participant variables. We additionally perform a critical survey of recent works in NLG to assess how well evaluation procedures are reported in this subdomain, noting a severe lack of open reporting concerning the demographic characteristics of participants, as well as a significant reliance on crowdsourcing platforms for recruitment.

2.1 Introduction

The aim of a natural language generation (NLG) system is to generate coherent, well-formed text, usually given an input prompt (Wang et al., 2022; Raffel et al., 2020), outline (Tang et al., 2022b; Goldfarb-Tarrant et al., 2020), topic (Van de Cruys, 2020), or dataset (Kale and Rastogi, 2020; Chen et al., 2020). As a result, NLG contains many subdomains such as summarisation (Goldsack et al., 2023, 2022), data-to-text generation (Kale and Rastogi, 2020; Chen et al., 2020), dialogue generation (Tang et al., 2023b), story generation (Tang et al., 2022b; Goldfarb-Tarrant et al., 2020), and humour generation (Loakman et al., 2023b; Sun et al., 2022b). Whilst it may be possible for any competent speaker of the target language to sufficiently understand generated outputs, many language forms are more subjective, such as humour, irony, and sarcasm, where interpretations are intertwined with broader demographic characteristics, as well as individual and idiosyncratic factors.

Human evaluation is viewed as the foremost form of evaluation for any NLG system (Howcroft et al., 2020), owing largely to its direct capture of the opinions of the (usually) human end-user.¹ This is made even more prominent when considering the low concurrent validity that many easily deployed automatic evaluation metrics have with human judgements (Zhao et al., 2023a; Alva-Manchego et al., 2021; Reiter, 2018). Whilst such evaluation procedures are considered preferable over automatic metrics (particularly in cases where no reliable automatic metrics are available), this does not mean that they are infallible. Numerous recent works have outlined the inconsistency with which human evaluation procedures are executed, and increasing attention is being brought to the downstream effects that such oversights may cause (Zajko, 2021).

In this paper, we present an overview of three particular types of language: humour, irony, and sarcasm - demonstrating how different evaluator demographics may affect

¹Sometimes NLG outputs are used as inputs to further NLG models (Zhang et al., 2022b).

the interpretation of such language and therefore its evaluation in relevant systems. Furthermore, we present a critical survey and assessment of 17 NLG research papers from 2018 to 2023 in the ACL Anthology concerning the generation of these types of language, analysing reporting transparency of human evaluation procedures.

2.2 Related Works

In recent years, the execution of human evaluation in NLG has become an even more significant topic, with myriad top-tier venues hosting dedicated tracks and workshops specifically focused on the evaluation of evaluation *itself*, such as the HumEval workshops at EACL 2021, ACL 2022 and RANLP 2023.² As a product of such venues, multiple recent works have been proposed that aim to highlight common inadequacies in multiple aspects of human evaluation, including the design of assessment criteria, the recruitment and training of participants, the use of biased response elicitation techniques (Parmar et al., 2023), and the transparency of reporting these processes.

Howcroft et al. (2020) assess a 20-year time-span of work within NLG (165 papers between 2000 and 2019) in regard to how evaluation procedures are performed, finding that there is substantial variation in how evaluation criteria are named and portrayed. Resulting from this, they produce normalised mappings between different criteria descriptions and their underlying evaluation focus and call for human evaluation criteria to be standardised, in addition to the creation of evaluation sheets outlining best practices for response elicitation from evaluators. Amidei et al. (2019) also assess NLG research in terms of response elicitation methods with a particular focus on the use of Likert scales and common deviations from best-practice within 135 papers from a 10-year period (e.g., whether or not to view them as ordinal or interval in nature, and

²<https://humeval.github.io>

CHAPTER 2. THE IRON(IC) MELTING POT

consequently how to perform statistical significance testing). Further issues have also been identified by [Schoch et al. \(2020\)](#) with the framing of evaluation questions and participant recruitment methods, including the presence of demand characteristics in evaluators being more common on crowdsourcing platforms. This is because participants vie for limited positions and wish to give researchers the results they believe they desire, under the assumption that it may lead to future work. Furthermore, these issues are exacerbated by cases wherein researchers reject annotations outright due to not falling in line with their expectations, or pay inadequately for services on crowdsourcing platforms ([Huynh et al., 2021](#); [Shmueli et al., 2021](#)). However, whilst automatic metrics may have many drawbacks for evaluating NLG, human evaluation has been suggested to be imperfect by its very nature, with human judgements being inconsistent. For example, [Clark et al. \(2021\)](#) find that human evaluators range significantly in their preconceptions of the quality of text NLG systems can produce, resulting in evaluators having biased perspectives when assessing text quality (a bias that is only somewhat remedied by the provision of training materials to standardise criteria definitions among evaluators). In order to partially mitigate these evaluation pitfalls, [van der Lee et al. \(2019\)](#) suggest a range of best practices, including the reporting of confidence intervals for results, using multiple ratings of each criterion, and using randomisation and counterbalancing when assigning evaluators to conditions (e.g., human- vs model- generated text). Finally, due to many of the problems that we outline in this work, the domain is currently starting to demonstrate a paradigm shift toward understanding the role of human disagreement as a valuable feature, rather than noise, resulting in the rise of perspectivist approaches ([Uma et al., 2022](#); [Plank, 2022](#); [Meaney, 2020](#); [Pavlick and Kwiatkowski, 2019](#)). Additionally, whilst the use of such language can lead to different responses from different people, this is compounded in human-computer interaction where the communicative divide is even more pronounced ([Shani et al., 2022](#); [Shapira et al., 2023](#)).

2.3 Introducing Humour, Irony, and Sarcasm

In the following section, we discuss three primary language types where interpretations and opinions are highly affected by a range of nuanced participant variables. For each type, we outline the nature of this language alongside a range of examples and reflect upon how the interpretation of these examples may differ across demographics.

2.3.1 Humour

Type	No.	Example
Humour	1.	Q: “Why would Trump change a lightbulb?” A: “He was told Obama installed it.”
	2.	Q: “How does a man change a lightbulb?” A: “He holds it and waits for the world to revolve around him.”
Non-humour	3.	“Colourless green ideas sleep furiously.”

Table 2.1: *Examples of humour.*

As the first language form to discuss, humour is also one of the most complex. The general consensus within cognitive and linguistic theories of humour is that an essential element for something to be found humorous is the existence of some degree of incongruity between expectations and reality (Raskin, 1985). Consequently, the experience of humour arises from the resolution of these incongruities within one’s mind. This, in turn, is how puns may elicit humour, when the initial interpretation of a text primes the interpretation of one semantic sense of a word, before that expectation is violated when the text is viewed as a whole. For example, in “the greyhound stopped to get a hare cut”, the interpretation of “hare” as in the rabbit-like animal is primed by the mention of the “greyhound” which chases hares, whilst “hare cut” forces the reader to reinterpret “hare” as the homophone “hair”, thus resolving the incongruity from the non-collocational phrase “hare cut” (He et al., 2019).

CHAPTER 2. THE IRON(IC) MELTING POT

Whilst puns are the most frequently studied form of humour in NLG (Tian et al., 2022b; He et al., 2019; Mittal et al., 2022), primarily due to their simple exploitation of phonetics and word senses, humour may also arise from more complex incongruities which are more likely to be resolved by individuals with specific world knowledge. Numerous works have ascertained a link between different demographic variables and humour perception. For example, Jiang et al. (2019) explore the literature on how individuals in the Western and Eastern worlds view humour as suitable under different conditions, and differ also in the exact forms of humour they opt to produce. Table 2.1 presents examples of such humour.

In Example 1, the perception of humour is dependent on factors including an individual’s knowledge of politics and their ideological leaning, as well as region.³ The reader is therefore assumed to have external background knowledge of US politics in order to infer that the joke concerns Donald Trump’s record of reversing actions made by his predecessor, Barack Obama. Additionally, due to the joke using Trump as the target, in effect calling him petty, whether or not this joke is seen as humorous is also dependent on one’s political leanings. This is due to Republicans, the political party that Trump represented, being less likely to find comments mocking him to be humorous. In Example 2, the joke is based on the shared knowledge of individuals in a patriarchal society that benefits men. As a result, the joke uses men as the target and comments on a presumed level of arrogance. In addition, this example requires the reader to interpret both literal and metaphorical meanings of something “revolving around” something else. Consequently, factors that could affect the perception of humour here are one’s familiarity with a patriarchal society and culture (as it can be imagined that such assumptions about male arrogance are not shared by members of a matriarchal culture). Additionally, due to the joke targeting men, there is the

³In this case, US politics is quite well-known globally, but a similar joke for a different government could be imagined, where universal knowledge would be far less likely.

possibility for some men to be offended at the assumptions that it is making, and it is known that gender generally affects the perception of certain types of jokes, such as these (Lawless et al., 2020). Furthermore, a large body of work has investigated the different tendencies of particular genders towards different patterns of language use, which in some cases may also affect the perception of different language types (Rabinovich et al., 2020; Coates, 2016). Finally, Example 3 (Chomsky, 1957) is an instance of an incongruous sentence that is semantically nonsensical (yet grammatically valid), where this incongruity does not give rise to humour (in the general case).

2.3.2 Irony

Type	No.	Example
Irony	1.	“Just found out about the train strike today, guess I’ll have to use my wings and fly there myself.”
	2.	“It’s been a great week for dictators. Congrats North Korea and Cuba.”
	3.	“Never knew something could be 80% halal... I’ve learned something new I guess.”
	4.	“A local fire station has burnt down.”
Non-irony	5.	“I went to the hospital yesterday. It was a big building.”

Table 2.2: *Examples of irony.*

Although the exact definition of irony varies, within natural language processing (NLP) research, the general consensus is that verbal irony is a form of figurative language where the producer of the language intends to convey the opposite meaning of the literal interpretation (Camp, 2012; Burgers, 2010; Grice, 1978). Consequently, similar to within the humour domain, an ironic statement violates the true beliefs of the producer and the expectations of the audience. It is for this reason that many ironic (and sarcastic) statements may also be considered humorous - a well-known feature of British comedy programming (Gervais, 2011). This means that a correct understanding of irony

CHAPTER 2. THE IRON(IC) MELTING POT

presupposes true or accepted knowledge about whatever entity, person or situation is being referred to.

A range of ironic texts is presented in Table 2.2. Irony can, for example, be factual, where it is generated by making a factually incorrect statement, such as in Example 1, where a human asserts that they have wings. Alternatively, it can also rely on social conventions, habits, or shared knowledge among members of a community. For instance, in order to understand the irony in Example 2, the receiver is expected to know that most people would not genuinely applaud the actions of someone they call a dictator. Whilst some knowledge and social conventions are ubiquitous, many ironic expressions also assume specific knowledge that is often only shared within a particular social group. For example, within the statement seen in Example 3, the receiver of the message is required to understand that “halal” is most often used as a binary concept, and therefore something being a “percentage” halal is infeasible, thus requiring additional background knowledge about Islam and its conventions. A single individual can, of course, be a member of several different social groups that share common ideals and knowledge, such as political conviction, free-time activities, or music taste, without even having to consider major cultural or geographical divergences. This simple fact, that irony strongly relies on common-sense knowledge, makes it an inherently subjective phenomenon, both for detection and the evaluation of generative models. In addition, ironic statements can also rely on situational or contextual knowledge, in which case they are then classified as *situational*, as opposed to *verbal*. An example of situational irony is demonstrated in Example 4, where a fire station burning down is unexpected due to containing multiple resources explicitly for the extinguishing of such fires. Contextual information is frequently essential for most people to recognise most forms of irony (Wallace et al., 2014). Finally, Example 5 presents a non-ironic statement, where expectations about concepts are not violated, as a hospital is usually

Type	Example
Sarcasm	1. “It’s astonishing how someone so average and unremarkable can manage to feel good about themselves.” 2. “Nice perfume. How long did you marinate in it?”
Non-sarcasm	3. “You’re an idiot.” 4. “Thank you so much for buying me a Ferrari!”

Table 2.3: *Examples of sarcasm.*

assumed to be a large building. Recent works have taken into account the plethora of variables affecting the interpretation of irony, with [Freunda et al. \(2023\)](#) developing a corpus of irony from a perspectivist approach, finding that irony perceptions vary with generation and nationality (even within speakers of the same language).

2.3.3 Sarcasm

Traditional definitions usually consider sarcasm to be a subcategory of irony. In addition to having a strong negative connotation and aggressive tone, this form of verbal irony is intended to ridicule someone or something ([Filik et al., 2019](#)).

Table 2.3 presents a selection of sarcastic texts. The clearly aggressive intent inherent in sarcasm is presented in Examples 1 and 2. In Example 1, the intention is to express that someone with seemingly no special qualities is able to have a sense of pride, whilst in Example 2, the choice of the word “marinate” implies that the target of the comment is wearing too much perfume, similar to how meat may marinate in a glaze for many hours. Examples 3 and 4 present non-sarcastic utterances, with Example 3 being a blatant insult, and Example 4 demonstrating the expected gratitude someone would have if they had been bought a supercar (assuming no deeper contextualised meaning, such as it being a toy car).

As sarcasm is considered to be a subcategory of irony, research into irony detection or generation may often include sarcasm and treat them as one and the same ([Maladry](#)

CHAPTER 2. THE IRON(IC) MELTING POT

et al., 2024; Van Hee, 2017; Filatova, 2012; Jijkoun and Hofmann, 2009). Moreover, studies by Gibbs (1986) and Bryant and Fox Tree (2002) indicate that there is an ongoing semantic shift, where verbal irony is perceived more often as sarcasm in popular use. Recent papers on sarcasm generation (Oprea et al., 2021; Mishra et al., 2019) do not systematically discuss their definition of sarcasm in relation to the concept of verbal irony. However, one may assume that sarcasm is synonymous with verbal irony, since it would be suboptimal for generative systems to create content that intentionally ridicules people in an explicitly hurtful way. In short, both for pragmatic reasons as well as conceptual changes, it makes sense to consider irony and sarcasm to be the same. Thus, the same motivations that make irony subjective apply to sarcasm as well. On top of that, detection and generation tasks concerning sarcasm specifically also demand that evaluators are able to discern the aggressive tone and retrieve the intention of the author or speaker, which can vary significantly. For example, Phillips et al. (2015) find age to be a significant contributing factor in an individual’s ability to correctly perceive intended sarcasm, with older individuals struggling more than younger ones. Furthermore, Rockwell and Theriot (2001) find gender differences in the patterns of sarcasm use to be affected by both the gender of the speaker, as well as the gender of the recipient interlocutor. Additionally, the use of sarcasm has been shown to vary alongside sociocultural contexts, with Blasko et al. (2021) noting that sarcasm use is more prevalent in individualist countries with a lower power distance such as the U.S than collectivist countries with higher power distance such as China.⁴ It logically follows, therefore, that these differences may also impact one’s ability to perceive intended sarcasm, and consequently, how they would react to sarcastic utterances in different contexts.

⁴Power distance refers to the distribution of perceived power (political, financial, etc.) across a society. A lower power distance means that members of a society are more “equal”, for example, when comparing a blue-collar worker to a politician.

2.4 Critical Survey

Following existing works that critically review human evaluation procedures in NLG, we present a critical survey of works in the area of computational humour, irony, and sarcasm. We begin by reviewing the scope and range of our literature search for this review and the prerequisite criteria for incorporation. Following this, we motivate the need for increased care in this specific domain by briefly exploring the prevalence of research in the area, both in terms of generative works and those on the detection of such language. As previously discussed, irony and sarcasm are highly related language forms, resulting in the common use of the terms interchangeably. However, to ensure that the intended definition is respected (the correctness of which cannot always be inferred with complete certainty), we report irony and sarcasm separately based on the titles of the papers.

2.4.1 Scope and Selection

The papers included in this survey pool were all taken from the ACL Anthology for the venues of *ACL, EMNLP, COLING, LREC, INLG, and CoNLL, between 2018 and 2023.⁵⁶ Initially, all papers concerning humour, irony and sarcasm were selected. This was performed by keyword matching titles on the following word stems: “hum-”, “sarc-”, “iron-”, “fun-”, “jok-” and “pun-”.⁷ The selected papers include main conference, Findings, and workshop submissions, in addition to one system demonstration, resulting in 259 papers. Submissions to shared tasks were then excluded, reducing the pool to 135 papers.⁸ The resulting papers were then verified to ensure relevance and categorised

⁵⁶Where the proceedings were available at the time of writing.

⁶INLG and CoNLL contributed 0 papers that met our criteria for inclusion.

⁷In order to capture “humour”, “sarcasm”, “irony”, “funny”, “joke”, and “puns”, including all derivations and regional spelling.

⁸Due to shared task submissions often attracting approaches with limited novelty, whilst having a standardised evaluation procedure.

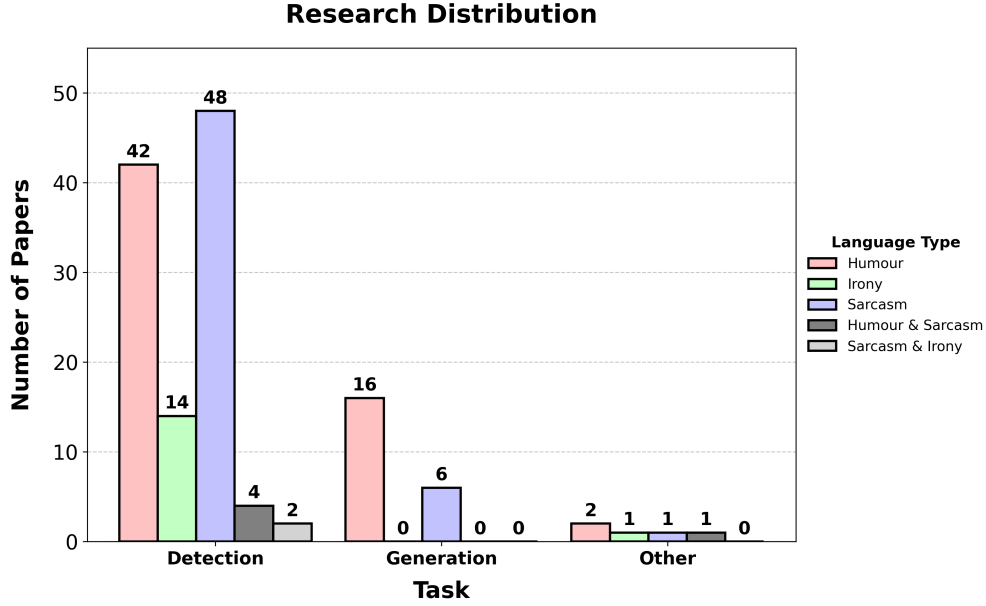


Figure 2.1: Breakdown of papers by task and language type. Shared task submissions are excluded, but survey and dataset papers are included. The “other” task category refers to papers with contributions such as annotation schema (Pamulapati et al., 2020) and cognitive analyses (Zeng and Li, 2022).

broadly into generation and detection (the latter of which may be binary, multi-class, or estimation). These criteria resulted in 135 remaining papers, of which 22 were focused on the generation of humour, sarcasm and/or irony, and 108 on detection (with the remaining five being classified ed as “other”). We then excluded four of the remaining 22 generation papers that were surveys or dataset papers where human evaluation was performed on non-computationally generated examples. This resulted in a final selection of 18 generation papers, all of which are listed in Appendix 2.8.1.

2.4.2 Subject Prevalence

As seen in Figure 2.1, there has been a significant number of accepted papers on these topics within this 6-year scope of collection. In addition, this search also revealed five shared tasks within this period, including a sarcasm detection shared task at the

Figurative Language Processing workshop at ACL 2020 (Ghosh et al., 2020), Task 7 of SemEval 2020 and 2021 on humour detection (Faraj and Abdullah, 2021; Hossain et al., 2020), a WANLP 2021 shared task on sarcasm detection (Abu Farha et al., 2021), and SemEval-2022 Task 6, also on sarcasm detection (Abu Farha et al., 2022).⁹

In addition to the works captured by our literature search and the aforementioned shared tasks, it is reasonable to think that the prevalence of recently released large language models (LLMs) such as ChatGPT and GPT-4 would result in further increases in research into this field, now that models are widely available that can better handle the generation and explanation of complex topics such as humour. Not explicitly included in this survey, but relevant, are also more general works on anthropomorphism in language-based AI agents (Gros et al., 2022; Abercrombie et al., 2021), in which the ability to detect, generate, and respond to humour, irony, and sarcasm are paramount if anthropomorphism is deemed desirable (Oprea et al., 2022).

2.4.3 Reporting Transparency

Of the 135 remaining papers, we now move our attention to the 18 focused on generation. All but one of the 18 papers contained human evaluation, and therefore, the remaining 17 papers form the basis of the following analysis.

Participant Demographics. As explored in Section 2.3, the interpretation of subjective language is shown to vary based on a range of characteristics. Specifically, characteristics such as age and gender are widely believed to measurably impact subjective language perception (Rockwell and Theriot, 2001; Phillips et al., 2015; Rabinovich et al., 2020; Lawless et al., 2020). Such characteristics are also frequently the focus of perspectivist approaches that aim to create models that better represent individuals

⁹All shared tasks were on detection rather than generation, due to the simpler nature of evaluation.

CHAPTER 2. THE IRON(IC) MELTING POT

with particular demographic characteristics (Uma et al., 2022; Plank, 2022; Meaney, 2020; Pavlick and Kwiatkowski, 2019). Furthermore, much subjective language relies on one’s understanding of a particular topic. It is for this reason that we survey location/nationality as a proxy for geographically-bounded knowledge (e.g., knowing about a particular nation’s politics or traditions). Education level and social class¹⁰ further aims to capture world knowledge and preferences. Outside of culturally relevant factors, language proficiency is arguably the most essential criteria to report as understanding subtle cultural cues to humour, sarcasm, and irony, first requires sufficient vocabulary and grammatical knowledge to recognise surface level semantics.

As shown in Table 2.4, less than half of the assessed works (7 of 17) reported *any* demographic information. Of those that were more transparent in regard to evaluator demographics, language proficiency, and location (usually as a proxy for nationality) were the most common details given. A single paper (Mishra et al., 2019) reported evaluator education.

Importantly, even in papers that outlined some demographic information, the reporting was frequently crucially inadequate. For example, Oprea et al. (2022) provide the full information sheet provided to participants but reduce the demographic information to “We target everyone registered as living in ⟨country⟩ on the Prolific Academic platform”, where the valuable information of *which* country is redacted. Additionally, the one paper that reported education (Mishra et al., 2019) was only to the extent that the two evaluators “had a better understanding of the language and socio-cultural diversities” - a statement that may range from meaning native speakers of the target language in the

¹⁰The inclusion of social class in this survey is an Anglocentric variable. In the UK, social class is frequently reflected in political alignment, employment industries, and leisure activity choices (Le Roux et al., 2008). Such relationships may not manifest as clearly in other regions. We therefore do not view the lack of reporting of such a criteria as a critique, but include it to give an overall view of reported variables.

Factor	Count	Percentage
Language Proficiency	4	<24%
Location/Nationality	4	<24%
Age	0	0%
Gender	0	0%
Education	1	<6%
Social Class	0	0%
Any	7	<42%

Table 2.4: *Counts of papers reporting demographic factors.*

desired culture, to people with doctorates in sociology. None of the surveyed papers outlined evaluator age, gender, or social class.

Recruitment. As can be seen in Table 2.5, 12 of the 17 surveyed papers contained explicit mention of participant recruitment methods. Upon further investigation, we find that all 12 instances concern the use of crowdsourcing platforms, of which nine are Amazon Mechanical Turk (AMT), two are Prolific Academic (Oprea et al., 2021, 2022), and one does not explicitly mention the platform (Tian et al., 2022b). Of the 12 papers that report recruitment, compensation amounts are provided in less than half (five papers), ranging from approximately 9.23 USD¹¹ (Oprea et al., 2021, 2022) to 20 USD per hour (Tian et al., 2022b; Mittal et al., 2022).

Transparent reporting of recruitment methods is important due to allowing inference of potential participant characteristics that may not be reported elsewhere. For instance, research in the social sciences has often been criticised for focusing primarily on WEIRD participants, western and educated individuals from industrialised, rich and democratic societies (Henrich et al., 2010). However, as access to language technology becomes more democratised, cheaper to access, and easier to use (Simons et al., 2022; de Dios-Flores et al., 2022), the majority of end users of these technologies are unlikely to be well

¹¹This study paid 0.38 GBP for approximately three minutes of work. The USD conversion is the exchange rate at the time of publication.

CHAPTER 2. THE IRON(IC) MELTING POT

represented by this set of characteristics. This tendency is further reflected by the overwhelming bias towards work on English, which leads to the demotion of other non-digitally-accessible languages and reinforces existing inequalities (Søgaard, 2022). In addition to these characteristics, there has also historically been a bias towards male participants in scientific research (García-González et al., 2019; Holdcroft, 2007), something that is also common in AI due to the field of computer science still being male-dominated (Breidenbach et al., 2021; European Commission and Directorate-General for Research and Innovation, 2019). Not only does this result in AI research not involving people from key demographics who will be affected by these technologies, but the research performed on the generated language will lack reproducible evaluation, as heterogeneous evaluation panels are not conducive to easy comparison across studies.¹² Whilst the substantial reliance on crowdsourcing platforms, particularly Mechanical Turk, may result in a different set of participants to the typical WEIRD demographic in social science and health research, the lack of reliable worker filters and allowances on what demographic data can be collected by requesters nevertheless results in a biased sample. Earlier works on the use of crowdsourcing have even touted the low cost as a benefit, whilst paying as low as 1/10th of a US cent per individual task (Callison-Burch, 2009). Perspectives such as these, therefore, result in biased demographics being used for evaluation and annotation at best, and the outright exploitation of people from particular socioeconomic backgrounds and regions at worst.

In terms of inter-annotator agreement (IAA), we find that just over half of the surveyed papers include explicit mention of statistical tests for agreement levels. IAA is one method through which the variation amongst evaluator opinions can best be analysed, in order to understand how subjective a task is, and consequently how best to assess the quality of an NLG system’s output where there is no “ground truth”

¹²We exclude cases where research was intended for a specific demographic group, where a homogeneous evaluator panel may be desirable.

Factor	Count	Percentage
Recruitment	12	<71%
Compensation	5	<30%
Inter-Annotator Agreement	9	<53%
Training Materials	4	<24%

Table 2.5: *Counts of papers reporting logistical factors.*

for a given criterion (e.g., the humour, irony, and sarcasm discussed here). Finally, we note that only four (of 17) papers convey any information regarding the level of training given to participants, whether that constitutes the presence of qualifier tasks on crowdsourcing platforms (Sun et al., 2022a,b) or extensive documentation on the nature of the task (Oprea et al., 2022, 2021). In tasks that are inherently subjective, the provision of adequate training materials can help to account for variation in the understanding of different individuals by orienting opinions, ensuring that all evaluators are working from the same definition of a particular phenomenon. Especially in cases where IAA is not reported, the provision of training given to evaluators is useful to the research community for estimating the extent to which evaluators were aligned.

2.5 Proposed Reporting Criteria & Discussion

Our analysis has shown inadequate reporting of relevant demographic information and details of evaluation procedures. To address this, we propose that future work on the evaluation of generative tasks include an “evaluation statement”, akin to the data statements proposed by Bender and Friedman (2018). We suggest the following details:

Evaluation Logistics

- number of evaluators used

CHAPTER 2. THE IRON(IC) MELTING POT

- the split of samples across evaluators (i.e., does everyone rate every sample)
- recruitment method used (e.g., internal emails, posters, crowdsourcing)
- compensation given to evaluators (e.g., gift cards, cash). Additionally, the choice to not monetarily compensate evaluators should also be divulged.

Demographic Information

- age (e.g., range, group, generation)
- language proficiency & “native” language (L1)
- gender identity
- cultural background (e.g., religion, political leaning, country of origin, education level)
- task proficiency (i.e., expert or layperson)

Evaluation Process

- clear and operationalised definitions of the evaluated phenomena (with explicit mention if a term with a largely accepted definition is being used in a non-standard way)
- amendments made in light of potential pilot studies
- training materials and examples of criteria given to evaluators
- agreement study (i.e., reporting inter-annotator agreement)

Whilst the language proficiency of the participants is relevant to all tasks, we recognise that other characteristics such as gender and cultural background (e.g., religion and political leaning) may not always apply. Still, it is beneficial to report which factors were taken into consideration, as they may become relevant in hindsight, and in the case of dataset annotations, facilitate more extensive use of any collected data in future works. Similarly, task proficiency may sometimes seem irrelevant for subjective and intuitive tasks. However, [Iskender et al. \(2021\)](#) and [Lloret et al. \(2018\)](#) demonstrate the preference towards expert evaluators for complex linguistic tasks. Additionally, authors may be reluctant to present only moderate agreement for subjective tasks due to concerns regarding evaluation quality. However, this criterion is highly relevant as it describes the expected degree of subjectivity for a task, a concept the community should be able to embrace. In addition, we also encourage increased interest in perspectivist approaches to the study of these language forms. Finally, the intentional omission of many of these criteria (in particular, IAA and language proficiency) may be considered intentionally deceptive, and therefore at odds with the foundations of open science.

2.6 Conclusion

In this work, we reiterate the importance of human evaluation in assessing the outputs of NLG systems but demonstrate that particular types of language, such as humour, irony, and sarcasm, demand additional attention during the construction of human evaluation processes. We do not discourage the use of the “non-native” or L2+ speakers of a target language for evaluation (or any other minority demographic), nor do we likewise encourage the use of “native” or L1 target-language speakers (or any other majority demographic). Rather, we encourage additional consideration from the community when selecting human evaluator panels, including analysing results in light of these demographic characteristics and openly reporting such decisions. To this end, we provide a list of suggested criteria to form an “evaluation statement”. Additionally, we advocate for the use of stratified sampling techniques to gain more representative evaluator panels.

2.7 Limitations

The limitations of this work are as follows. We only survey papers from 2018 to 2023 to capture recent evaluation behaviours at the time of publication. However, this results in a small number of available papers. This is naturally exacerbated by the choice of focusing solely on humour, sarcasm, and irony, rather than language generation more generally. However, we believe that broad surveys of NLG have been performed in recent years to high standards, and instead wish to use these specific language forms as explicit examples of the type of language that is most affected by these factors. We additionally convey that transparent reporting and consideration of evaluation practices is applicable to NLG as a whole, and not just to the subdomains we focus on. Furthermore, we exclusively cover the use of human evaluation in generation tasks.

CHAPTER 2. THE IRON(IC) MELTING POT

However, many of the same ideas apply equally to the selection of annotators during dataset creation for detection tasks. All statements and opinions expressed in this work are by nature in good faith, and we accept, appreciate and welcome perspectives from other individuals from different backgrounds to support, challenge, or revise any of the claims made herein. Regarding the selection of demographic variables analysed in Table 2.4, we acknowledge that these criteria present only a subset of those that are potentially relevant to the interpretation of subjective language. Furthermore, some criteria reflect the context of the survey’s authors at a UK institution. Social class, for instance, is defined differently in different areas, some of which may not have as clear a cultural difference as traditionally seen in the UK. It is for this reason that we do not directly criticise works for not reporting such demographic details, but aim to present a general overview of the type of information that is present or absent. Whilst we encourage the collection of metadata about participants in NLG research, we emphasise that this data should only be collected where the correct ethics application procedures have been followed by the respective researchers, and that collected information should be only as fine-grained as is needed to accurately characterise the participant panel (for example, one’s home-town is not necessary when nationality or cultural identity would suffice, and neither is one’s exact age when 10-year age brackets may be sufficient). Additionally, the critical survey was performed at the start of this research project, before Automatic Rolling Review incorporated mandatory reporting strategies such as the Responsible NLP Checklist. Consequently, modern reporting practices may have improved.

Additional Acknowledgements

Aaron Maladry is supported by Ghent University [grant BOF.24Y.2021.0019.01.]

2.8 Appendix

2.8.1 Surveyed Generation Papers

The full list of generation-based research papers that contain human evaluation from the critical survey are presented in Table 2.6.

Title	Author (Year)
A Neural Approach to Pun Generation	Yu et al. (2018)
Punny Captions: Witty Wordplay in Image Descriptions	Chandrasekaran et al. (2018)
A Modular Architecture for Unsupervised Sarcasm Generation	Mishra et al. (2019)
Pun-GAN: Generative Adversarial Network for Pun Generation	Luo et al. (2019)
Pun Generation with Surprise	He et al. (2019)
Can Humor Prediction Datasets be used for Humor Generation? Humorous Headline Generation via Style Transfer	Weller et al. (2020)
R ³ : Reverse, Retrieve, and Rank for Sarcasm Generation with Commonsense Knowledge	Chakrabarty et al. (2020)
Homophonic Pun Generation with Lexically Constrained Rewriting	Yu et al. (2020)
“Judge me by my size (noun), do you?” YodaLib: A Demographic-Aware Humor Generation Framework	Garimella et al. (2020)
Chandler: An Explainable Sarcastic Response Generator	Oprea et al. (2021)
Should a Chatbot be Sarcastic? Understanding User Preferences Towards Sarcasm Generation	Oprea et al. (2022)
Context-Situated Pun Generation	Sun et al. (2022b)
ExPUNations: Augmenting Puns with Keywords and Explanations	Sun et al. (2022a)
A Unified Framework for Pun Generation with Humor Principles	Tian et al. (2022b)
AMBIPUN: Generating Puns with Ambiguous Context	Mittal et al. (2022)
“When Words Fail, Emojis Prevail”: A Novel Architecture for Generating Sarcastic Sentences With Emoji Using Valence Reversal and Semantic Incongruity	Kader et al. (2023)
Multi-modal Sarcasm Generation: Dataset and Solution	Zhao et al. (2023b)

Table 2.6: *List of generation papers in the critical survey.*

Part III

Tongue Twister Generation

Chapter 3

TwistList: Resources and Baselines for Tongue Twister Generation

ACL 2023 ([Loakman et al., 2023b](#))

Tyler Loakman^a, Chen Tang^b, and Chenghua Lin^c

^aDepartment of Computer Science, University of Sheffield, UK

^bDepartment of Computer Science, University of Surrey, UK

^cDepartment of Computer Science, The University of Manchester, UK

Contextual Statement

Progressing from the factors that affect human evaluation of subjective language in Chapter 2, this chapter addresses RQ2, focusing on the generation of a form of phonologically motivated humour known as tongue twisters, using a topic-to-twister generative task. This chapter, therefore, investigates the extent to which grapheme-based, pre-trained language models can learn the phonological characteristics of tongue twisters by introducing a novel dataset (TwistList) to fine-tune a range of small pre-trained language models (TwisterMisters), additionally presenting naive phoneme-overlap based metrics (PO/IPO) to identify the sound repetition inherent in tongue twisters.

Abstract

Previous work on phonologically-grounded language generation has focused on domains such as lyrics and poetry. In this work, we present investigations on the generation of tongue twisters - a form of language that is required to be phonologically conditioned to maximise sound overlap, whilst maintaining semantic consistency with an input topic, and still being grammatically correct. We present TwistList, a large annotated dataset of tongue twisters consisting of 2.1k+ human-authored examples. We additionally present several benchmark systems (TwisterMisters) for the proposed task of tongue twister generation, including models fine-tuned on in-domain data and zero-shot solutions. We present the results of automatic and human evaluation to demonstrate the performance of pre-trained models in this task with limited or no task-specific fine-tuning and no explicit phonological knowledge. We find that the task of tongue twister generation is challenging for models under these conditions, yet some models are capable of generating acceptable examples.

3.1 Introduction

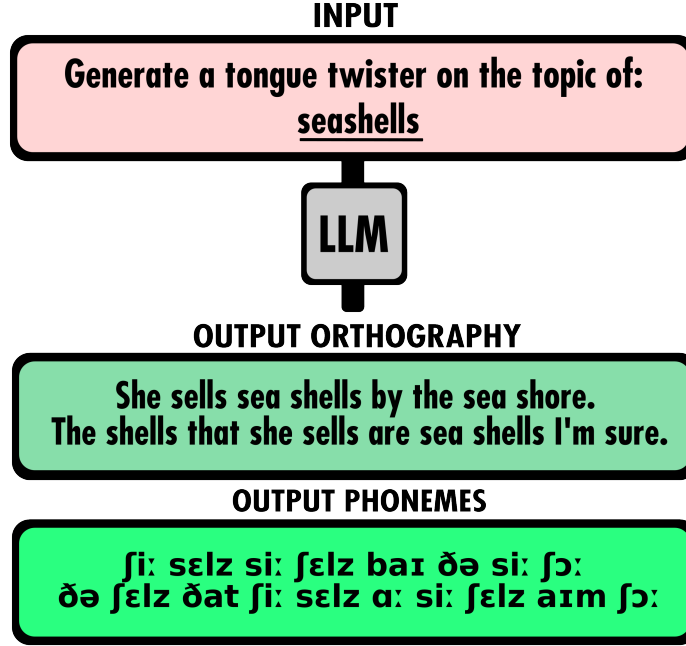


Figure 3.1: *Outline of the topic-to-twister generation procedure.*

Phonologically constrained language generation is a subdiscipline of computational creativity in natural language generation (NLG), predominantly encompassing lyric and poetry generation (Tian and Peng, 2022; Wöckener et al., 2021; Xue et al., 2021; Zhang et al., 2020a; Agarwal and Kann, 2020), as well as humour generation (Sun et al., 2022b; He et al., 2019; Yu et al., 2018), and continues to prove challenging for myriad reasons. Primarily, such works require the inclusion of factors such as metre and rhyme, which involve careful consideration of candidate vocabulary on the syllable level, leading to a reduced pool of allowable vocabulary once these constraints are in place.

In this paper, we present work on the generation of tongue twisters, a type of phonologically constrained language that is rarely explored in the NLG community. As a form of creative generation, tongue twisters can facilitate numerous useful applications, including: (i) being used as a pedagogical tool (Sugiharto et al., 2022; Somoff, 2014;

Wilshire, 1999); (ii) as a source of humorous entertainment stemming from unintentional mispronunciations; (iii) as a stylistic device for engaging children in reading (e.g., Dr. Seuss stories, Geisel, 1965); (iv) as a method of designing memorable slogans and tag lines (Guerini et al., 2015; Yorkston and Menon, 2004); and (v) as stimuli in neuroscience research (Wong et al., 2019; O’Halloran, 2020; Kember et al., 2017).

Tongue twister generation posits unique challenges compared to other generation tasks. One of the most pertinent features of tongue twisters is the presence of high levels of phonological overlap across words (Wilshire, 1999). Consequently, whilst other types of creative generation may require only *some* output tokens to consider phonology and phonetic realisation, tongue twisters present an extreme version of this problem where the characteristics of almost all generated tokens must be considered. This leads to a very small vocabulary from which to choose semantically relevant words and presents further challenges with maintaining grammatical validity.

To summarise our contributions, we present:

- **TwistList**, a large annotated dataset of human-authored tongue twisters, containing 2.1k+ examples with human quality evaluation.
- **TwisterMisters**, a series of baseline models for tongue twister generation using popular state-of-the-art language models.
- Extensive automatic and human evaluation to assess the ability of language models to implicitly model the complex phonological phenomena in tongue twisters.

3.2 Related Works

¹Numerous efforts have been made toward the generation of creative language forms, with a range of findings regarding whether or not popular LLMs truly exhibit human-level creativity in these tasks. [Chakrabarty et al. \(2024\)](#) present work applying the Torrance Tests of Creative Thinking (TTCT) to objectively analyse the outputs of LLMs and human authors on a narrative writing task, finding that LLM generations perform measurably worse than humans, passing 3-10x fewer criteria outlined by the TTCT. However, it is important to note that this was in comparison to professional authors, who represent a very niche subset of the best human writers available. On the other hand, [Gómez-Rodríguez and Williams \(2023\)](#) compare human and LLM-authored narratives and find that LLMs are able to match or surpass human performance on several of the evaluation criteria they present. However, in this case, the “creativity” of the task was somewhat diminished by having prescribed rules about the topic, characters, and writing style, where the task may be more construed as emulating the writing of an existing work. However, similarly, both [Franceschelli and Musolesi \(2024\)](#) and [Clark et al. \(2021\)](#) observe that humans are infrequently able to distinguish creative works written by other humans from those authored by LLMs, with the latter often achieving very high-quality outputs. Overall, there is clear potential and room for improvement in the field of automatically generating creative language forms.

A popular trend is to investigate the extent to which models can be trained to generate language forms where training data is scarce. [Wöckener et al. \(2021\)](#) investigate this for the generation of poetry using $\approx 16k$ and $\approx 67k$ quatrains of English and German poetry, respectively. They notice difficulties in GPT-2 learning sub-lexical phenomena, including rhyme, from this number of training examples alone. However, poetry presents

¹This related works section has been expanded with content from the longer [Loakman et al. \(2025b\)](#) paper presented in Chapter 4.

CHAPTER 3. TWISTLIST

a highly restrictive form of literary language where many types contain formal constraints regarding length, syllable count, and metrical patterns. Additionally, [Van de Cruys \(2020\)](#) presents work on the generation of Shakespearean sonnets, another literary niche that contains an even more limited number of available training samples. They therefore approach the task via adding constraints at decoding time in a pipeline approach that includes the stages of content planning, rhyme generation, and output polishing to imbue literary sensibilities into the outputs of models trained exclusively on non-literary text. Finally, [Popescu-Belis et al. \(2023\)](#) use data synthesis techniques to increase the amount of rhyming data they can train GPT-2 ([Radford et al., 2019](#)) on, to realise GPoeT, which shows an increased ability to generate consecutive rhymes.

The only known work on tongue twister generation at the time of conducting this research is by [Keh et al. \(2023\)](#), who present models that train on graphemes and phonemes, and take either a starting prompt to be continued or keywords around which to theme an output. They release TT-Corp, a dataset of 644 tongue twisters with parallel non-twister equivalents. We differentiate our work through the release of a dataset that is over 3x larger and which has undergone substantial human quality control. Furthermore, we assess the results of a wider range of popular pre-trained models on this task, including ChatGPT, without explicit injection of phonological knowledge. Our experimental results show that the majority of popular, pre-trained language models rely on pure word repetition to generate tongue twisters, whilst some (i.e., BART) are able to generate more sophisticated examples. Additionally, very large zero-shot models (i.e., ChatGPT) are able to generate convincing tongue twisters almost on par with human equivalents.

Other forms of language where phonemic and phonetic information are essential have also been generated, including rap ([Xue et al., 2021](#); [Manjavacas et al., 2019](#); [Potash et al., 2018](#)) and song lyrics more generally ([Tian et al., 2023](#); [Chang et al., 2023](#);

Zhuo et al., 2023; Zhang et al., 2022a). Computational research on creative works is not restricted to such domains, however, with extensive work also existing in the area of narrative generation (Hong et al., 2023; Tang et al., 2022a; Chen et al., 2021), humour generation (Loakman et al., 2023a; Sun et al., 2022b; Tian et al., 2022b), metaphor processing (Wang et al., 2024b, 2023b; Li et al., 2023a, 2022b), and music generation (Li et al., 2024; Yuan et al., 2024b; Yu et al., 2023).

3.3 Methodology

3.3.1 Task

We formulate the task of tongue twister generation as follows: for a given set of keywords, we aim to generate a tongue twister T , whereby T comprises a sequence of words $\{w_1, w_2, \dots, w_n\}$. The generated output must satisfy the following constraints: (1) the output should be semantically related to the input keywords; (2) the output should show maximal levels of phonological overlap across words; and (3) the output should be grammatically valid (Wilshire, 1999). Of these requirements, phonological overlap is the most central to defining text as a “tongue twister”.

3.3.2 Dataset (TwistList)

Dataset Construction. We present TwistList, an annotated dataset of 2.1k+ human-authored tongue twisters for use by the community. The examples contained therein come from a variety of sources available on the web.² For each tongue twister, phonemic transcription is provided using the g2pE package, in addition to keywords extracted with RAKE and BERTopic to represent the topic of the tongue twister.³ Following

²The source of each tongue twister is stated for each entry.

³g2pE is available at: <https://github.com/Kyubyong/g2p>

CHAPTER 3. TWISTLIST

	Train	Val	Test	Total
<i>Dataset Totals</i>				
# Samples	1,912	106	107	2,128
# Vocabulary Entries	9,556	946	880	10,358
# RAKE Keywords	3,333	316	288	3,567
# BERTopic Keywords	250	132	160	250
<i>Sample Averages</i>				
Avg. Twister Length (words)	15.01	16.59	13.54	15.01
Avg. RAKE Keywords	3.16	3.32	3.01	3.16

Table 3.1: *Descriptive statistics of the TwistList dataset.*

experimentation with both RAKE and BERTopic, only RAKE keywords are used in training due to human preference and issues regarding the use of BERTopic on short texts (where frequently no keywords are extracted). The main statistics of the dataset are presented in Table 3.1.

Feature	Content
RAKE Keywords	sells thick socks
BERTopic Keywords	short shorts socks sock
Graphemes	Seth at Sainsbury’s sells thick socks.
Phonemes	[S EH1 TH] [AE1 T] [S EY1 N S B ER0 IY0 Z] [S EH1 L Z] [TH IH1 K] [S AA1 K S]

Table 3.2: *An example entry from TwistList.*

Quality Control. Quality control on TwistList was performed in multiple ways. Firstly, it was ensured that only sufficiently unique tongue twisters were kept in the dataset, as determined by removing examples with over 90% word overlap (rather than keeping variants of the same tongue twister, such as “Peter Piper picked a pickled pepper” versus “Peter *the* Piper picked...”). Additionally, non-standard spellings were manually converted to standard US English⁴ to avoid grapheme-to-phoneme

⁴For example, where phonetic spellings or letter substitutions such as “k” for “c” were used for literary and visual effect, such as “kwik” for “quick”.

Data Source	Sample Quality (%)			
	High	Medium	Low	κ
RAKE Keywords	82	18	0	0.321
BERTopic Keywords	15	85	0	0.445
Tongue Twisters	88	6	4	0.321

Table 3.3: *TwistList* quality control results. κ refers to Fleiss’ Kappa (Fleiss, 1971). Good tongue twisters that are deemed a bit longer (3%) or shorter (3%) than preferred are considered medium quality.

conversion errors.⁵ Similarly, tongue-twisters containing obscure vocabulary (such as the names of medications and dinosaurs) were excluded to further minimise errors. An annotation platform was developed, on which three human evaluators, who are native speakers of English, were provided with 100 randomly sampled instances from the dataset to rate the quality of the resulting tongue twisters and the associated extracted keywords. The full dataset contains 2.5k+ tongue twisters, of which 2,128 are kept for training/development/testing after filtering examples with insufficient extracted keywords and excessive similarity to existing entries.

To summarise, three annotators evaluated the quality of TwistList on a three-point scale (i.e., high, medium, and low quality, based on overall impressions), with 88% of assessed tongue twisters being considered high quality, and 6% considered medium quality. An example from TwistList is provided in Table 3.2. As Table 3.3 shows, the final dataset can be considered high quality, owing to fair-moderate approval and agreement across evaluators. Demographic information of the evaluators can be found in Appendix 3.9.1.

⁵g2pE uses the CMU Pronouncing Dictionary to retrieve transcriptions, which is an American English resource.

3.3.3 Models (TwisterMisters)

Fine-Tuned. For the fine-tuning, we select a range of popular models for language generation: including GPT-2 (117M) (Radford et al., 2019), DialoGPT (117M) (Zhang et al., 2020c), BART (139M) (Lewis et al., 2020), and T5 (220M) (Raffel et al., 2020). These were fine-tuned with RAKE keywords extracted from human-authored tongue twisters as the input, and the tongue twister text from TwistList as the target, representing baselines fine-tuned on in-domain data. At inference time, the prompt “Generate tongue twisters about the keyword(s): X” is used, where X refers to the input consisting of one or more RAKE keywords. Full training details are presented in Appendix 3.9.2.

Zero-Shot. We additionally use a TwisterMister that does not require fine-tuning. Specifically, we utilise OpenAI’s GPT-3.5-based ChatGPT with the same prompt used for fine-tuning, but for zero-shot generation.⁶ Each request to ChatGPT was submitted as part of a separate session to avoid the effects of extended dialogue influencing outputs. At the time of writing, ChatGPT had been utilised in order to set a practical upper-bound of what may be expected from models without explicit phonological knowledge.

3.4 Results

Automatic Evaluation. We present the results of automatic evaluation on generated outputs and gold examples in Table 3.4 for the following metrics: Perplexity (PPL), BLEU (B-1/B-2) (Papineni et al., 2002), ROUGE (R-1/R-2/R-L) (Lin, 2004), and BERTScore precision, recall, and F-score (Zhang et al., 2020b) (BS-P/BS-R/BS-F).

⁶No direct comparison is made to PANCETTA (Keh et al., 2023) as no code had been publicly released at the time of publication, and essential implementation details are absent from the paper.

Model	Fluency	BLEU		ROUGE			Phonemic		BERTScore		
	PPL↓	B-1↑	B-2↑	R-1↑	R-2↑	R-L↑	PO↓	IPO↓	BS-P↑	BS-R↑	BS-F↑
GPT-2	8.40	0.007	0.003	1.30	0.12	1.32	0.022	0.020	0.690	0.810	0.744
DialoGPT	<u>3.83</u>	0.038	0.025	7.72	3.61	7.64	<u>0.069</u>	<u>0.089</u>	0.754	<u>0.831</u>	0.790
BART	1.65	0.073	0.051	11.88	6.11	10.35	0.075	0.120	0.795	0.845	0.819
T5	10.16	<u>0.057</u>	<u>0.038</u>	<u>9.70</u>	<u>4.57</u>	<u>9.57</u>	0.689	0.727	0.795	0.818	<u>0.806</u>
ChatGPT	N/A	0.200	0.137	36.77	20.66	33.44	0.093	0.157	0.888	0.894	0.883

Table 3.4: Automatic evaluation results. The phoneme-based metrics for the gold standard references are: $PO = 0.385$ and $IPO = 0.417$. Due to the one-to-many issue in creative language generation, the presented reference-based metrics are imperfect. The best performing fine-tuned model for each metric is presented in **bold**, whilst the second best is underlined. Arrows indicate the preferred direction of change.

PPL, BLEU and ROUGE are standard metrics in language generation to assess quality, whilst BERTScore assesses semantic similarity to a gold reference. Additionally, we propose two new metrics, phoneme overlap (PO) and initial phoneme overlap (IPO). PO refers to the average overlap of all phonemes across tokens, whereas IPO is the ratio of unique word-initial phonemes to the total number of words.

The phoneme-based metrics reward longer outputs. We argue that, all things equal, a longer tongue twister is better than a shorter one as it provides more entertainment and more opportunities for mispronunciation. Moreover, perfect scores on PO and IPO can be achieved through the repetition of a single word. Whilst this does not lead to high-quality outputs, these metrics are intended exclusively to be indicators of the phonemic characteristics, rather than an overall guide to quality. In both cases, higher levels of overlap result in lower (i.e., better) scores, and the highest (i.e., worst) achievable score is 1.

The results in Table 3.4 show a rather clear scaling, with the performance ranking on most metrics (except perplexity and phoneme overlap) being identical. On the models explicitly fine-tuned for this task, GPT-2 is shown to be the worst, whilst BART performs the best. We hypothesise that GPT-2’s poor performance is due to its simple

causal language modelling objective alongside its decoder-only architecture (which is also in DialoGPT). Furthermore, although T5 performed well on the automatic metrics, manual inspection revealed that it often misinterpreted the task instructions, choosing to select its own keywords from the entire prompt text, rather than using only the provided keyword list. On the other hand, the training-free zero-shot model, ChatGPT, was shown to perform best on all metrics. This is to be expected as ChatGPT is substantially larger than any other tested model, with various pre-training objectives and reinforcement learning, leading to performant zero-shot capabilities. This further demonstrates that language models struggle to learn phonological patterns implicitly from text, especially in English, which has high levels of irregular orthography. Furthermore, with limited data, language models struggle to learn the unusual probability distributions underlying tongue twisters, where word choices are intentionally obscure and anti-euphonious. Additionally, due to the wealth of training data seen by ChatGPT, many examples were likely observed during training.

Human Evaluation. Due to tongue twisters being a creative domain where articulation abilities are tested, we also perform human evaluation. Three evaluators were asked to rate 100 outputs from the best performing fine-tuned baseline (BART), in addition to ChatGPT and gold standard references from TwistList on the following criteria: *Relevance* (i.e., how relevant the tongue twister is given the keyword inputs), *fluency* (i.e., how grammatically valid the output is), *difficulty of articulation* (i.e., how difficult a tongue twister is to say), *coherence* (i.e., how much sense the output makes), and *entertainment value* (i.e., how entertaining the output is). All ratings were given on a 5-point Likert scale. Evaluator demographics are presented in Appendix 3.9.1.

The mean scores of human evaluation in Table 3.5 align with expectations, with gold examples performing best on all metrics, and ChatGPT placing second on all

CHAPTER 3. TWISTLIST

but difficulty of articulation.⁷ BART is able to produce outputs that are deemed to be the second most difficult to articulate, which we infer may be the result of slight morphological variants of input keywords being used repeatedly, making distinguishing between them during articulation quite challenging (whilst not being able to exploit deeper phonological relationships). The moderate score on fluency (3.028) suggests instances of poor grammar may also hinder articulation abilities when expected grammatical structures are not found, leading to an interaction between grammatical validity and articulatory difficulty. Additionally, ChatGPT scoring the lowest for articulatory difficulty may be due to occasionally misunderstanding the requirements of a tongue twister, sometimes producing rhymes or standard prose. However, ChatGPT scores well for relevance and fluency, highlighting its ability to produce high-quality, coherent language. Perhaps most interestingly, none of the average BART scores fall below the midpoint score of 3 on any human evaluation criteria, despite being a small model fine-tuned on only 2128 examples, with no additional phonological knowledge.

Whilst the pattern of extreme repetition is seen in many of the fine-tuned models (with the exception of BART, which is demonstrated to be capable of producing slightly more sophisticated outputs), overall assessment of the generated tongue twisters reveals interesting patterns. Firstly, the limits of GPT-3.5-Turbo are made apparent, such as the input “*silver shiny ship sank*” resulting in the output “*How much wood would a woodchuck chuck if a woodchuck could chuck silver shiny ships?*”, a clear derivation of a famous woodchuck-related tongue twister that likely appears multiple times in pre-training corpora. Additionally, comments from evaluators reveal that ChatGPT’s output is often considered more of a rhyme than a tongue twister. However, examples such as these are also found in the human-authored gold examples, demonstrating that there is no steadfast, consistent opinion as to what constitutes a (good) tongue twister.

⁷We exclude relevance for the gold examples as the input keywords were extracted from these examples originally.

Metric (Score 1–5)	Human Evaluation		
	BART	ChatGPT	Gold
Relevance	4.667*	<u>4.971</u> ***	-
Difficulty of Articulation	<u>4.143</u> *	4.102*	4.291 *
Fluency	3.028**	<u>4.915</u> **	4.938 **
Coherence	3.217*	<u>4.798</u> *	4.909 *
Entertainment Value	3.269*	<u>4.070</u> *	4.254 *

Table 3.5: *Human evaluation results. The best scores are in **bold**, and the second best are underlined. We calculate Fleiss’ κ for each metric, and mark the agreement fair* (0.21 - 0.40), moderate** (0.41-0.60), and substantial*** (0.61 - 0.80). The relevance of gold samples is omitted as input keywords were extracted from these samples.*

3.5 Case Study

Within the example in Table 3.6, GPT-2 resorts to simply repeating the input, successfully achieving phonological overlap, but failing to be grammatically valid or particularly sophisticated. This pattern is also demonstrated by DialoGPT and T5. Conversely, BART is able to introduce tokens unseen in the input to create an almost grammatically valid output (the primary mistake being indefinite article agreement, where “a” should be “an”). BART’s output is also semantically and logically coherent, with “A assistant assist is an assistant assist” being valid, yet redundant, whilst “assistants assist to assist assistants” is also comprehensible. This example demonstrates why evaluators with high English proficiency and linguistics education were selected, as the same word may have different parts of speech, creating outputs that may seem grammatically invalid at first glance, yet do follow the rules of English.⁸ Further investigation is needed to ascertain whether the models are intentionally exploiting this lexical ambiguity, or if human evaluators are demonstrating apophenia, where patterns are found in what is effectively noise (Brugger, 2001). Finally, ChatGPT utilises morphology to exploit the

⁸https://en.wikipedia.org/wiki/Buffalo_buffalo_Buffalo_buffalo_buffalo_buffalo_Buffalo_buffalo

similarity of the plural noun “assistants” and the phrase “assist ants”, and provides a continuation that is in line with the expected behaviour of ants. In comparison to the gold example, ChatGPT’s output may be considered more interesting topic-wise, at the expense of not being as phonologically complex (e.g., “carrying leaves to the ant hill” contributes heavily to semantics, whilst not being recognisable as part of a tongue twister in isolation).

Model	Generated Tongue Twisters (Input: “assistant assist”)
GPT-2	assistant assist assistant assist assistant
DialoGPT	assistant assistant assistant assistant assistant assistant assistant assistant
BART	A assistant assist is an assistant assist, assistants assist to assist assistants.
T5	assistant assist assistant
ChatGPT	Assistant ants assist ants in carrying leaves to the ant hill.
Gold	<i>If I assist a sister-assistant, will the sister’s sister-assistant assist me?</i>

Table 3.6: *Example model outputs. Gold refers to the human-authored reference tongue twisters from TwistList.*

3.6 Conclusion

In this work, we presented a range of experiments on the topic of tongue twister generation, a form of phonologically-constrained language generation that aims to maximise phonological overlap, whilst conveying meaningful semantics. We motivate the potential application domains for such generated language and provide a large annotated dataset of tongue twisters, TwistList, to encourage further work. Finally, we present a series of benchmark models alongside automatic and human evaluation to assess generation quality.

3.7 Limitations

Whilst the system presented within this paper is capable of allowing human-in-the-loop contributions (via selecting the input keywords on which to condition the output), it is not able to produce tongue twisters that take explicit advantage of particular features of speech sounds such as place and manner of articulation, in order to create more advanced outputs that exploit phoneme relatedness (rather than exact matches). The same can be said of our proposed metrics, PO and IPO, which do not account for similarity across sounds that share manner or place of articulation (e.g., “she sells sea shells”). Additionally, whilst commonly known tongue twisters may follow a particular format, such templates have not been enforced here. We also do not demonstrate the capabilities of these systems if they were trained on phonemic transcriptions, as we aim only to assess their performance when training on graphemes, observing implicit learning. Furthermore, the presented finetuning set up remains similar across all trained models. Whilst this allowed for consistency in training, it also means that individual models were not fine-tuned explicitly to best match each model’s intended paradigm (e.g., text completion, dialogue). However, inputs were formatted to match general expectations (e.g., task prefixes for T5). Statistical significance testing was omitted from cross-model evaluations due to the main contribution of this work being the identification of a general capability gap between language models and humans in the task of tongue twister generation. Whilst inter-model performance differences are discussed, the purpose of this work is primarily exploratory and to provide a novel dataset rather than to suggest a specific model for the task of tongue twister generation or present a state-of-the-art architectural solution.

3.8 Ethics Statement

Regarding the generation of tongue twisters, language generation is a necessarily creative domain that has the ability to reproduce content that some individuals may find offensive. Care was taken to check outputs in the human evaluation set for any such materials, and if they had been produced, they would have been removed. Additionally, no egregiously offensive material has been provided in the TwistList dataset. However, the distinction between offensive and humorous content is complex, and not all examples may be suitable for all individuals.

3.9 Appendix

3.9.1 Human Participants

To follow best practices for human evaluation, the relevant demographic information of participants is outlined below.

Dataset Evaluation. All evaluators involved in the quality control process of TwistList are native speakers of English, and either have or are working towards university-level qualifications. Additionally, two of the three evaluators have extensive education in linguistics or modern languages. No monetary incentive was provided.

Generation Evaluation. All evaluators involved in the evaluation of the quality of generated tongue twisters are native speakers of English, and either hold or are working towards university-level qualifications in Linguistics, Modern Languages, or NLP. Additionally, all evaluators cited the United Kingdom as their country of socialisation, and no participants reported language processing difficulties that could affect results. No monetary incentive was provided.

Materials Provided to Evaluators. Additionally, evaluators for both the dataset and generation outputs were presented with calibration examples to demonstrate the type of outputs that would be presented and the logic behind particular scores, minimising individual interpretations of the scoring criteria. All evaluation was performed on a custom online annotation platform.

3.9.2 Implementation Details

Table 3.7 presents the hyperparameters used for model fine-tuning. The specific models used were: GPT-2 (openai-community/gpt2), DialoGPT (microsoft/DialoGPT-small), BART (facebook/bart-base), and T5 (google-t5/t5-base), ChatGPT (gpt-3.5-turbo) was accessed through the Chat Completions API.

Hyperparameter	Setting
Epochs	5
Batch Size	32
Optimiser	Adam
Learning Rate	8e-5
Maximum Generation Length	100
Maximum Source Length	150
Hardware	Nvidia Tesla A5000

Table 3.7: *Fine-tuning hyperparameters.*

Chapter 4

Train and Constrain:

Phonologically Informed Tongue

Twister Generation from Topics and

Paraphrases

Computational Linguistics ([Loakman et al., 2025b](#))

Tyler Loakman^a, Chen Tang^b, and Chenghua Lin^c

^aDepartment of Computer Science, University of Sheffield, UK

^bDepartment of Computer Science, University of Surrey, UK

^cDepartment of Computer Science, The University of Manchester, UK

Contextual Statement

Continuing with tongue twister generation, this chapter addresses RQ3 by investigating the scaling of tongue twister fine-tuning data through a phoneme-aware large language model (LLM) pipeline for synthetic data creation (TwisterLister), resulting in an expanded dataset (TwistList 2.0) to investigate the effects of fine-tuning data volume on the ability of small language models to learn phonological characteristics from orthography alone. Consequently, this chapter progresses the integration of LLMs within this thesis from gold-standard reference benchmarks to elements of an active pipeline, reflecting the growing presence of such technologies around the time of publication and the shifting focus of the field towards LLM-powered solutions. This chapter additionally improves upon previously presented naive metrics (PO/IPO) by introducing phoneme-aware variants based on weighted phonemic edit distance (oPED/iPED) that consider feature similarity. This chapter further introduces the style-transfer tongue twister generation task, expanding upon the topic-to-twister approach presented in the previous chapter. Finally, a phoneme-aware constrained decoding module is introduced for generating topic-conditioned tongue twisters without the need for model fine-tuning (PACD).

Abstract

Previous work in phonetically and phonologically grounded language generation has focused on domains such as puns and poetry. In this work, we present new findings on the generation of English tongue twisters - a form of language that is required to be conditioned on a phonological level to maximise sound overlap, whilst maintaining semantic consistency with an input topic or phrase and still being grammatically correct. We present *TwisterLister*, a pipeline for generating phonologically informed tongue twisters from large language models (LLMs) that we use to generate *TwistList 2.0*, the largest annotated dataset of tongue twisters to date, consisting of 17k+ examples from a combination of human and LLM authors. Our generation pipeline involves the use of a phonologically constrained vocabulary alongside LLM prompting to generate novel, non-derivative tongue twister examples. We additionally present the results of automatic and human evaluation of smaller models fine-tuned on *TwistList 2.0* to demonstrate the extent to which phonologically motivated language types can be generated without explicit injection of phonological knowledge. Additionally, we introduce a phoneme-aware constrained decoding module (PACD) that can be integrated into autoregressive language models, demonstrating that this method generates good-quality tongue twisters both with and without fine-tuning the underlying model. We additionally implement a range of automatic metrics for the task of tongue twister generation that is phonologically motivated and captures the unique essence of tongue twisters, primarily based on phonemic edit distance.

4.1 Introduction

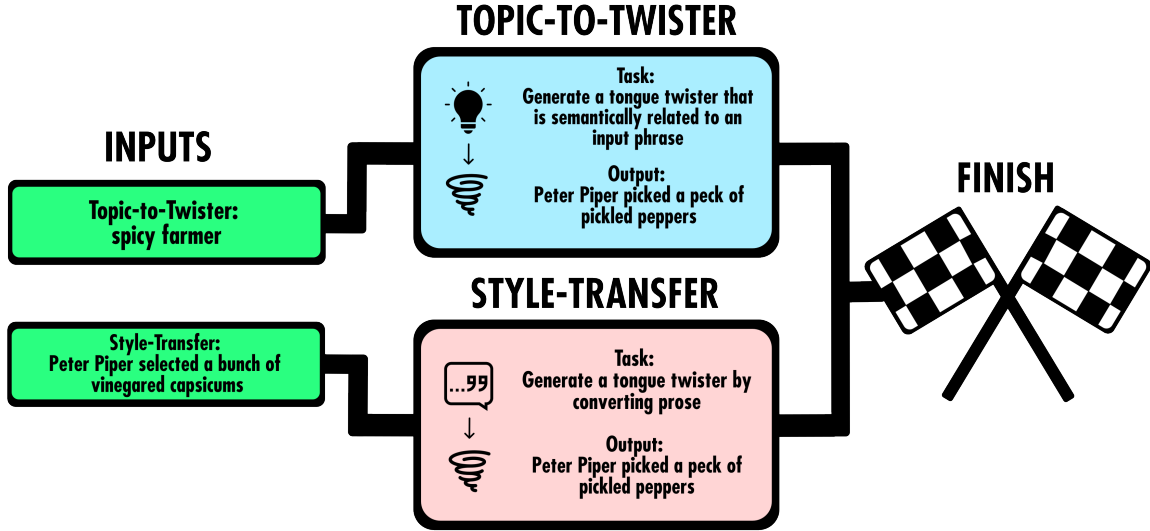


Figure 4.1: Outline of the generation pipeline for both the topic-to-twister and style-transfer tasks. Both settings use the same base models and hyper-parameters, differing only in the data used for fine-tuning.

Whilst the dawn of large language models (LLMs) such as OpenAI’s GPT-4 (OpenAI et al., 2023), Meta’s Llama 2 (Touvron et al., 2023) and Google’s Gemini (Anil et al., 2023) has brought unprecedented performance improvements in many natural language generation (NLG) tasks, these models are exceptionally resource-hungry in terms of data, compute, and API expenses. Consequently, many works have started to investigate the ways in which LLM capabilities and knowledge can be infused into smaller models, using the larger model for data enhancement via the generation of pseudo-labels (Tang et al., 2023a) or additional training examples (Yang et al., 2024a).

Additionally, LLMs are primarily designed to select the most probable continuation of a span of text based on their training data. Creative language, in direct opposition to non-creative language, is predominantly desired to be non-derivative, containing phrases and word sequences that are not ubiquitous in everyday language in order to evoke various emotions and engage a reader rather than purely convey information in a

linguistic form. As a result, creative language generation’s goals are at odds with the primary training paradigm of LLMs, as the goal is often *not* to select the most probable continuation, and instead surprise and engage readers (Roush et al., 2022).

4.1.1 Contributions

Towards the automatic generation of high-quality tongue twisters, we expand upon prior works to present TwisterLister, an LLM-based pipeline for the generation of unique, non-derivative tongue twisters to provide more extensive training data to enable the fine-tuning of smaller language models. TwisterLister employs semantic and phonological knowledge in the form of sentence embeddings and phonemic edit distance to restrict the candidate vocabulary to be passed to an LLM. In doing so, we create TwistList 2.0, the largest existing dataset of tongue twisters with over 17k examples, of which approximately 15k are derived directly from the proposed TwisterLister pipeline. We motivate the need for this extended wealth of training data by demonstrating the impact on both automatic metrics and human evaluation as a function of training-data volume by fine-tuning various smaller-scale language models on various splits of this dataset. We present these results as part of two different tongue twister generation approaches, topic-to-twister and, inspired by Keh et al. (2023), style-transfer (i.e., prose-to-twister). With the aforementioned real-world applications of tongue twisters, we motivate the topic-to-twister setting for applications such as language learning, where multiple outputs can be generated to test an individual’s articulatory abilities in a new language, whilst simultaneously expanding their vocabulary. On the other hand, the style-transfer setting is motivated by applications such as marketing, where a standard sentence conveying a desired meaning (e.g., a brand’s mission statement or a product’s features) can be reworded to have increased phonemic complexity to become a tongue twister, consequently engaging the reader and increasing memorability.

CHAPTER 4. TRAIN & CONSTRAIN

We additionally introduce PACD, a Phoneme Aware Constrained Decoding module, that can be used with any autoregressive language model to ensure token outputs meet phoneme-level criteria. Overall, our contributions may be summarised as follows:

- **TwisterLister** - A phonologically and semantically informed pipeline for generating English tongue twisters with large language models, both as a stand-alone generation method and as a data synthesis approach for additional training data.
- **TwistList 2.0** - The most extensive English tongue twister dataset at the time of publication, containing 17,000+ tongue twisters produced via TwisterLister ($\approx 15k$) and human authors ($\approx 2k$), including extensive quality control procedures, for use in training tongue twister generation models as well as presenting a resource for the study of this language form on a linguistic level.
- **PACD** - A phoneme-aware constrained decoding algorithm that applies hard lexical constraints on the outputs of autoregressive language models to achieve phoneme-level overlap and generate tongue twisters.
- **iPED/oPED** - Phonemic edit distance metrics for assessing the articulatory characteristics of tongue twisters on a word-initial and overall basis.
- A range of experiments training smaller language models to generate tongue twisters in topic-to-twister and style-transfer settings using TwistList 2.0, including extensive automatic and human evaluation.
- Extensive qualitative linguistic analysis of generations from different models fine-tuned on varying amounts of TwistList 2.0 data, with and without the constrained decoding PACD module.

4.2 Phonetics & Phonology

Due to the strong reliance on theory and ideas from the fields of phonetics and phonology in this article, we find it apt to begin with a short introduction to these domains. Firstly, phonetics refers to an area of linguistics that studies the production and perception of speech sounds present in spoken languages (Gick et al., 2013; Jessen, 2008; Ladefoged, 1996) whilst the related field of phonology focuses on the abstract mental representations of speech sounds and the development of feature-based taxonomies for the categorisation of related sounds (Clements and Ridouane, 2011; De Lacy, 2007; Klausenburger, 1970).

4.2.1 Place & Manner of Articulation

Figure 4.2 presents the primary pulmonic consonants present in human languages. For each consonant, three important pieces of information on their *phonetic* characteristics are interpretable. Firstly, the rows each represent a *manner* of articulation which refers to the physical process that occurs to produce a particular sound (such as a “plosive” involving the build-up and sudden release of air pressure within the mouth, whilst a “nasal” involves the nasal cavity through the lowering of the velum). On the other hand, each column refers to a *place* of articulation, which relates to the main location that articulators (e.g., tongue, teeth, and lips) make contact within the vocal tract (for example, “bilabial” sounds involve both lips and “labiodental” sounds involve the lips and teeth in their production). Finally, the last remaining detail is voicing, which refers to whether or not the glottal folds (also known as vocal folds or vocal cords) are vibrating during the sound’s production.

CHAPTER 4. TRAIN & CONSTRAIN

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b		t d			ʈ ɖ	c ɟ	k ɡ	q ɢ		ʔ
Nasal	m	ɱ	n			ɳ	ɲ	ŋ	ɴ		
Trill	ʙ		r						ʀ		
Tap or Flap		ⱱ	ɾ			ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative			ɬ ɮ								
Approximant		ʋ	ɹ			ɻ	j	ɰ			
Lateral approximant			l			ɭ	ʎ	ʟ			

Figure 4.2: *The International Phonetic Alphabet consonant chart. Symbols to the right of a cell are voiced (i.e., involve the vibration of the vocal folds) and sounds to the left are unvoiced. Grey boxes present place/manner combinations that have been deemed impossible for human production. This chart is reproduced from the International Phonetic Association, which distributes it under the CC-BY-SA license.*

4.2.2 Phonological Features

In addition to physical production-based characteristics, there are also further *phonological* features that can be used to explain the patterns and processes that phonemes undergo in human speech. An example of different phonological features is presented in Table 4.1. It is phonological features such as these that the phonemic edit distance (PED) metrics introduced in §4.4 and used throughout this article rely on to determine the similarity between two speech sounds (and therefore the likelihood of mispronunciation when transitioning between them).

4.2.3 Phonetic Transcription Standards

Multiple transcription standards exist for phonetic and phonological representations of text. In the majority of this article, the preferred standard is the International Phonetic Alphabet (IPA), where each speech sound is represented by a single symbol (e.g., /t, d, s, z/). The IPA is the standard transcription convention used within linguistic research

Phoneme	Consonantal	Sonorant	Voiced	Nasal	Continuant	Coronal
/t/	+	-	-	-	-	+
/d/	+	-	+	-	-	+
/s/	+	-	-	-	+	+
/l/	+	+	+	-	+	+
/i/	-	+	+	-	+	-

Table 4.1: A phonological feature chart. + means the feature is present for a given sound, whereas - means the feature is absent. Only a subset of possible sounds and phonological features are shown for demonstration purposes.

Standard Orthography	IPA	ARPABET	SAMPA
Hello World	/hɛləʊ wɜːld/	HH EH L OW W ER L D	hEl@U w@:ld

Table 4.2: Phonetic transcriptions of the phrase “Hello World” in the three most common standards. IPA refers to the International Phonetic Alphabet. The transcriptions reflect a non-rhotic Southern Standard British English accent. However, ARPABET is limited in its ability to represent non-rhotic vowels (hence “ER”).

and is used within Figure 4.2 and Table 4.1. We additionally make use of another transcription standard, ARPABET, which uses one or two characters to represent a given sound (or three where vowels have their stress marked). Table 4.2 presents an example of these standards, alongside SAMPA, another common standard.

4.3 Related Work

4.3.1 Constrained Generation

An inherent paradox often exists when using language models for creative language generation. Models trained to generate the most probable sequence of tokens are used to output text where the most *probable* continuation is sometimes one of the least *preferable*. In the simplest form, control over the output can be exercised through simple methods such as restricting the output vocabulary at decoding time (Hokamp and Liu, 2017;

Valitutti et al., 2013), or through the application of penalties for n -gram repetition, therefore encouraging diversity (Zhang et al., 2021; Foster and White, 2007). Several works have been performed in the direction of toolkits to aid in constrained language generation. Roush et al. (2022) present the Constrained Text Generation Studio (CTGS), an AI-writing assistant that in its most basic form applies a range of pre-made filters to the output of a probabilistic language model (such as “avoid the letter e ”, Wright, 2016), scanning the most probable next-word generations and only selecting those which fall in line with the selected constraints (of which numerous can be applied in parallel). However, due to the design of the constraints, CTGS struggles with the generation of well-formed outputs for models trained using the predominant paradigm of sub-word tokenisation. Similar approaches have also been utilised for the generation of other non-creative language types, such as machine translation and summarisation, where stylistic constraints and editorial decisions may result in a preferred output structure. For example, Yao et al. (2024) present COLLIE, a grammar-based framework for the application of advanced compositional constraints on the outputs of language models, in addition to a tool for generating example task instances from raw text. Additionally, Iso (2022) presents AutoTemplate, a method of formatting a task structure to realise lexically constrained text generation. However, whilst many constraint-based systems exist, most work on lexical constraints is focused on the inclusion of specific word choices within the output, and therefore, the desired candidate vocabulary has to be known a priori. Lu et al. (2022) propose to solve a common downfall of autoregressive decoding, where it is necessary to plan ahead, introducing NeuroLogic A*esque, a decoding approach that uses lookahead heuristics to more carefully consider future token generations. We build upon work in the area of constrained generation in §4.7 with PACD, a phoneme-aware constrained decoding approach that dynamically applies constraints on allowable tokens at each generation time step.

4.3.2 Knowledge Distillation

Alongside the advent of LLMs, so too arose the area of knowledge distillation (Gupta and Agrawal, 2022; Hinton et al., 2015; Buciluă et al., 2006) whereby the aim is to achieve the successful transfer of knowledge from a much larger model (often referred to as a teacher model) to one or more smaller models (often referred to as student models). In doing so, much more robust generalist models, such as GPT-3 (Brown et al., 2020) and GPT-4 (OpenAI et al., 2023) can have elements of their abilities passed down to smaller models that do not possess the same level of computing requirements (Yang et al., 2024b). Whilst there are numerous methods of achieving the desired distillation, perhaps the simplest and easiest of these approaches, particularly in domains where data is scarce, is to generate new synthetic training data using the larger models, either via the generation of completely new examples or via data augmentation and perturbation (Whitehouse et al., 2023; Askari et al., 2023). However, generating novel instances is not always straightforward, with the quality of generations often being dependent on the task type. For instance, Li et al. (2023b) find that more subjective tasks may result in lower quality synthetic data, such as not reflecting the same level of diversity as human-written equivalents, something that is key to creative language domains and that we aim to overcome in §4.4.1 for the creation of TwistList 2.0.

4.4 Dataset (TwistList 2.0)

Loakman et al. (2023b) presented TwistList 1.0, a dataset of 2.1k+ human-authored tongue twisters from various sources available on the web, including listicles and works of fiction. Whilst this allowed a high level of quality in the examples within the dataset (as all instances were automatically filtered, reviewed by a linguist, and then underwent quality control on a subset), the small size of this dataset (even whilst being the largest

dataset of tongue twisters we are aware of to-date) means that smaller models struggle to learn the key features of tongue twisters from such limited examples, specifically regarding the need to balance high levels of phonemic repetition alongside maintaining grammatical coherence.

To combat this shortfall, we extend TwistList 1.0 8-fold into TwistList 2.0, containing 17,000+ unique tongue twister examples. To achieve this feat, we note the near-human performance that was achieved by early versions of ChatGPT in our previous human evaluation studies (Loakman et al., 2023b), and opt to build a generation pipeline we name TwisterLister using GPT-3.5-Turbo to generate novel examples to facilitate the training of smaller parameter count models from the resulting synthetic dataset.¹

4.4.1 TwisterLister Pipeline

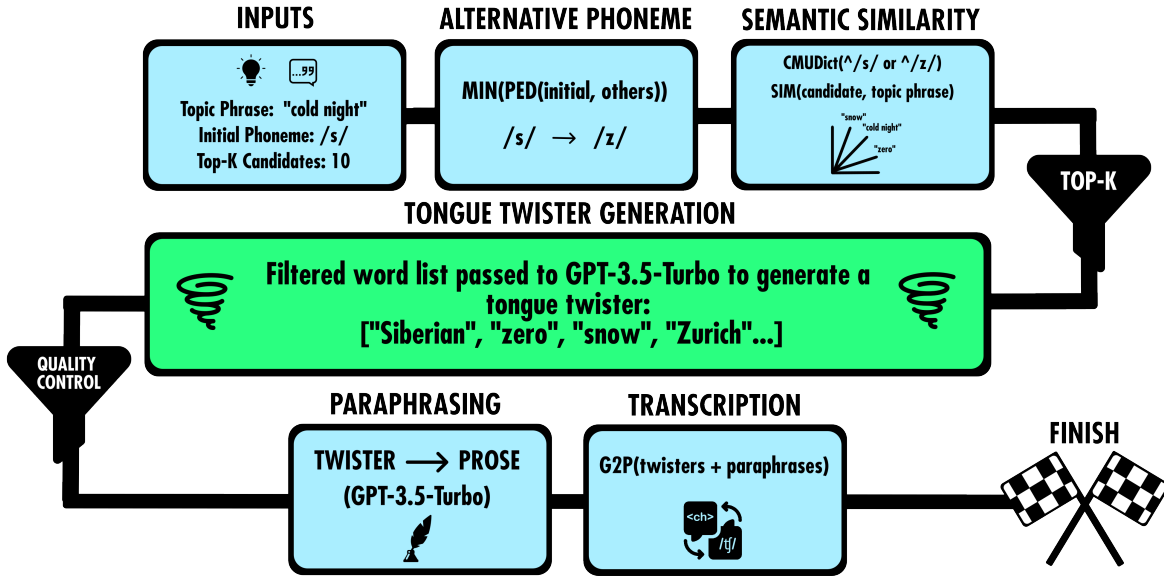


Figure 4.3: Illustrated outline of the TwisterLister pipeline used to create TwistList 2.0. PED refers to phonemic edit distance, SIM refers to cosine similarity, and G2P refers to grapheme-to-phoneme conversion.

¹In both Loakman et al. (2023b) and the present work, we access GPT-3.5-Turbo (i.e., ChatGPT) via the OpenAI API.

CHAPTER 4. TRAIN & CONSTRAIN

A key discovery in previous work (Loakman et al., 2023b) was the common reliance of ChatGPT on slightly modifying well-known existing tongue twisters that had been memorised from the training data when presented with a new topic (for example, “*silver shiny ships*” generated “*How much wood could a woodchuck chuck if a woodchuck could chuck silver shiny ships*”). To avoid this pitfall, we create a more constrained pipeline for tongue twister generation that promotes the generation of unique, non-derivative examples, illustrated in Figure 4.3. Initially, we generate the topics for the tongue twisters by building a set of topic phrases by combining a randomly sampled adverb or adjective with a noun to represent an abstract topic (using part-of-speech tags from NLTK’s Brown Corpus, Francis and Kucera, 1979). Following this, we then randomly select a phoneme on which to focus the tongue twister, restricting these choices to consonants (due to these being more commonly exploited in tongue twisters than vowels) and additionally removing any phonemes that are not legal in word-initial position in standard English phonotactics (such as the glottal stop /ʔ/) or phonemes with very few entries in our vocabulary pool (such as the voiced post-alveolar fricative /ʒ/). Following this, we search the CMU Pronouncing Dictionary (CMUDict)² for words starting with our preferred phoneme and calculate the cosine similarity between the SentenceBERT embedding of our NLTK-generated topic phrase with each candidate word retrieved from the CMUDict. Following this, the top N retrieved candidate words with the highest semantic similarity are kept, and the others are discarded (N = 5 or 10).

We then calculate the pairwise weighted phonemic edit distance (PED) between our initially selected phoneme and all other allowable phonemes, and select the lowest-scoring phoneme (i.e., most similar) to act as the best secondary phoneme.³ Where multiple share the same edit distance, the first reached when iterating over the list is selected. This is due to tongue twisters frequently relying on the reader mispronouncing a sound,

²Available at <https://github.com/Alexir/CMUDict/tree/master>.

³Phonemic edit distance is implemented with the PanPhon package (Mortensen et al., 2016).

CHAPTER 4. TRAIN & CONSTRAIN

Prompt A
“Generate a sensible and grammatical tongue twister using words from the following list: [word-list]. The output should be a single sentence and be grammatical and coherent.”
Prompt B
“Generate a tongue twister by primarily using words from the following list: [word-list]. The output should be grammatical and coherent.”

Table 4.3: *Prompts used in the TwisterLister pipeline for tongue twister generation. [word-list] refers to the conditioned vocabulary selected in the previous steps. Prompts were determined a posteriori, based on the empirically observed output characteristics.*

often due to confusion with a similar sound (e.g., “she sells sea shells” exploiting /ʃ/ and /s/). Consequently, allowing the generation process to select from two word banks that start with similar sounds means that we can more directly promote mispronunciations rather than solely relying on the repetition of a single sound. We then repeat the process of generating a candidate list of words that are semantically related to the topic and start with our desired secondary phoneme.

We then combine the word lists and shuffle their order to promote alternation between words with different, yet similar, initial phonemes, as experimentation showed that requesting an LLM to “alternate” between words from two different word banks, and presenting them separately, still resulted in the words being used largely in the same order that they were presented, and therefore not resulting in the desired alternation. The list is then fed into the LLM (GPT-3.5-Turbo, accessed August-September 2023 via the API) using the prompts presented in Table 4.3.

In Prompt A, where we specify single-sentence outputs, we see much more concise outputs that may suffer from coherence issues. On the other hand, with Prompt B, we see more coherent outputs, but which often resemble standard poetry more than tongue twisters (to which we apply further filtering as discussed in §4.4.3). Consequently, we use a combination of tongue twisters generated with either prompt to have a diverse range of styles. All generations are performed with *max_tokens* set to 1000 and a

temperature of 0.8 to facilitate creative completions. In total, we generated 17,500 example tongue twisters, of which 11,500 were generated with Prompt A, and 6,500 with Prompt B.⁴

4.4.2 Style-Transfer Paraphrase Generation

In line with [Keh et al. \(2023\)](#), we present an additional task setting alongside the topic-to-twister approach that utilises style-transfer. Whilst [Keh et al. \(2023\)](#) generate non-tongue-twister examples of their dataset entries via simple rule-based synonym replacement conditioned on part-of-speech tags, we leverage GPT-3.5-Turbo to paraphrase each entry in our tongue twister dataset. To achieve this, we pass the following system prompt to GPT-3.5-Turbo:

“In this task you will pretend that you’re an author who is rewriting existing works into a non-literary form that more resembles prose. You will be presented with a tongue twister and asked to rewrite it using synonym replacement so that there are no longer high levels of phonetic overlap and sound repetition. Example 1: INPUT = “She sells sea shells by the seashore.” OUTPUT = “The girl sells conches by the ocean.” Example 2: INPUT = “Peter Piper picks pickled peppers” OUTPUT = “Peter Piper selects preserved capsicums”.

We then present the following user message to GPT-3.5-Turbo for each dataset entry: “INPUT = [twister], OUTPUT = ”, where [twister] is a tongue twister from TwistList 2.0. This approach is superior to simple synonym substitution as the LLM can dynamically select new vocabulary in a way that further ensures the new paraphrases

⁴This imbalance is due to promoting the generation of “twisty” content, which is more prevalent with Prompt A. Consequently, we sample fewer Prompt B examples to contain a moderate amount of poetic text.

are fluent, rather than using raw synonym replacement that can result in semantic drift, as very few words are true synonyms (Chiang and Lee, 2023).

4.4.3 Output Refinement

To further refine the generated outputs, we process the resulting dataset in several stages. Firstly, all outputs generated using Prompt A (which promotes succinct tongue twisters that are not as coherent) are re-fed into GPT-3.5-Turbo in the prompt “**Improve the following tongue twister by editing it so that it makes more sense and is grammatical: [tongue twister]**”. Consequently, this step fixes errors that arise from the original word lists given in the prompt containing morphological variants that are difficult to turn into a coherent output (for example, all nouns being in possessive form, or verb tenses being mixed in a way that is detrimental to coherence). To further check that the remaining outputs are sensible, we calculate the perplexity (PPL) of the generations using a pre-trained language model (i.e., GPT-2) as a basic heuristic for well-formedness. We then compare these scores with the average PPL from the original TwistList 1.0 dataset of human-authored tongue twisters and remove any outputs from our new dataset that have perplexities that are higher than the original dataset’s mean plus one standard deviation. This stage removed 397 examples.

Additionally, as previously mentioned, Prompt B promoted the generation of longer, more poetic outputs, at the cost of not always resembling tongue twisters. To maintain only the best generations, we apply a metric for assessing the phonological characteristics of tongue twisters using weighted phonemic edit distance (PED) (Mortensen et al., 2016), which is outlined in further detail in §4.5.3. Again, we compare the results to the mean score from the original TwistList 1.0 dataset, and filter our new dataset by removing any examples that do not score lower (i.e., better) than the original dataset’s mean plus one standard deviation. However, no examples were caught in this

CHAPTER 4. TRAIN & CONSTRAIN

filtering stage, suggesting that the majority of the more poetic works still exhibited tongue-twister-esque phonological characteristics.

Next, to encourage diversity, we remove examples based on word-overlap. To achieve this, we apply a pairwise fuzzy-matching algorithm based on the token sort ratio (which is indifferent to word order) across the remaining tongue twisters and remove examples with more than a 60% overlap with an existing entry in the dataset.⁵ This step consequently led to the removal of 1,747 entries, therefore reducing repetition and increasing diversity in the remaining dataset.

We then filter out offensive examples that may have been created by poor topic and phoneme combinations that resulted in undesirable stereotypes or associations being expressed. These include words relating to the topics of racism, sexism, homophobia, transphobia, and additional terms that some may find offensive (including general expletives and references to various anatomy). We perform this by comparing the tongue twisters with a bank of offensive words and removing any twisters containing any of the examples, regardless of context. This stage removed a further 69 entries.

Finally, to ensure the diversity of input-output pairs when training in a topic-to-twister fashion, we remove any entries with duplicate NLTK topic phrases, as these cases would result in two tongue twister outputs for a single input (but with different phonological characteristics, as they otherwise would have been caught and filtered by previous stages). This stage removed a final 135 examples.

The final additions to the dataset comprise 15,151 examples, which, when combined with the existing TwistList 1.0, result in 17,278 unique tongue twisters in TwistList 2.0. We maintain a distinction between human and machine-authored tongue twisters in the final dataset so that different communities can make use of whichever is most pertinent to their application.

⁵We arbitrarily decide which to remove and continue until no remaining samples exhibit this level of overlap. We implement this using the RapidFuzz package: <https://github.com/rapidfuzz/RapidFuzz>.

4.4.4 Quality Control & Enhancement

Phonetic Transcription. As in Loakman et al. (2023b), we then enhance the dataset with the addition of phonetic transcription using the g2pE Python package.⁶ We experimented with other grapheme-to-phoneme solutions to see if an improvement was possible, such as SoundChoice (Ploujnikov and Ravanelli, 2022) which better accounts for correctly transcribing homographs, but opted to utilise g2pE again. This is primarily due to our TwisterLister pipeline relying on the CMUDict to retrieve vocabulary. g2pE queries the CMUDict for transcriptions before falling back on a trained neural network model to infer pronunciations for out-of-vocabulary tokens. Consequently, the majority of vocabulary will already have a gold standard transcription in CMUDict (in the General American accent, due to the limitations of this resource). We then convert these transcriptions into the International Phonetic Alphabet (IPA) to facilitate the use of our phoneme-based metrics, whilst also providing a transcription standard that is more common in linguistics.

Input Topics. Finally, whilst RAKE (Rose et al., 2010) was used in Loakman et al. (2023b) to extract keywords from the human-authored examples to represent the topic, we skip this stage and utilise the topic phrases from TwisterLister pipeline. As a result, unlike the keywords for the original tongue twister collection, those that are new to TwistList 2.0 have more abstract topics as the topic words themselves are not forced to appear in the output. We hypothesise that this may also help to reduce drawbacks seen in the original work, where our trained models often repeated the topic keywords numerous times to achieve a tongue twister, rather than learning a deeper representation of semantics.

⁶Available at: <https://github.com/Kyubyong/g2p/tree/master>.

Quality Control. Quality control on TwistList 2.0 was performed in multiple ways. Five human evaluators, who are native speakers of English, were provided with 50 sampled instances from the dataset from different conditions to rate the quality of the resulting tongue twisters (25 from Prompt A and 25 from Prompt B). Scores for the criteria were given on a scale from 1 (low quality) to 5 (excellent quality) for five criteria: (i) *Twister* (i.e., the overall assessed quality of the tongue twister and whether it exhibits the expected phonological characteristics), (ii) *Topic* (i.e., how well the semantics of the input topic are represented in the output), (iii) *Paraphrase Quality* (i.e., whether or not the paraphrase maintains meaningful and grammatical text), (iv) *Paraphrase Prosaic* (i.e., to what extent the paraphrase is believed to have successfully removed the sound overlap and tongue twister nature of the original input to better resemble standard prose), and (v) *Overall* (i.e., a holistic assessment of the dataset entry as a whole, considering the tongue twisters itself, in addition to its paraphrase and topic relevance). Table 4.4 presents the breakdown of the human evaluation results.

Overall, TwistList 2.0 is considered to be of good quality, particularly regarding the paraphrased versions and overall evaluation. When compared to other aspects, the ratings for the tongue twisters and topic relevance show lower scores but still indicate good quality (with 3 being a medium quality rating). This is particularly true when considering that tongue twister quality is highly subjective due to entertainment value being a fundamental component that people perceive differently. Additionally, topic relevance is lower due to not enforcing that the input terms are present in the output. On the other hand, extracting keywords from twisters and using these as the inputs would lead to an artificially inflated rating for the topic criteria, as the topic phrase would be guaranteed to be represented explicitly in the output. Overall, the dataset samples were given a mean rating of 3.778 on a 5-point scale. Further details of human participant recruitment are reported in §4.5.4.

TwistList 2.0 Quality Control				
Twister	Topic	Paraphrase Quality	Paraphrase Prosaic	Overall
3.138**	3.333*	4.421*	3.969**	3.778**

Table 4.4: Results of human quality assessments for TwistList 2.0. We sample only from the new machine-generated samples created with TwisterLister. The “Gold” samples evaluated in §4.6.2 are exclusively from the human-authored portion of TwistList 2.0. We calculate Fleiss’ κ for each metric, and mark the agreement fair* (0.21 - 0.40) and moderate** (0.41-0.60).

4.4.5 TwistList 2.0 Summary

Descriptive statistics regarding TwistList 2.0 can be seen in Table 3.1. Additionally, two example entries in the dataset are outlined in Table 4.6. As the examples demonstrate, the combination of adjective/adverb and noun used as an input topic during the TwisterLister generation pipeline can be directly reproduced as part of the final tongue twister or be represented solely by semantics. For example, in the top example (*TT_ID* 68), the input adjective “public” is replicated in the output, owing to the phoneme /p/ being randomly selected by the pipeline. Consequently, “public” itself, and its morphological variants, are very likely to be in the top-k most semantically relevant words to the topic word “public”, and therefore constitute part of the constrained candidate vocabulary. The choice of /b/ as the secondary phoneme can also be observed, which was selected via minimising phonemic edit distance (where /p/ and /b/ differ only in the presence/absence of voicing). This is observed through words such as “broadcaster” and the proper noun “Berman”. Similarly, in the second example (generated from the input “direct language”) we see “non-direct” in the output due to the selected initial or secondary phoneme being /n/ (an alveolar nasal), whilst “language” has been referenced less directly through the alternative phoneme /m/ (a bilabial nasal), using terms such as “monolingual”, “novelistic”, and “Mandarin”. The impact of the different prompt forms can also be seen, with the first example (using

CHAPTER 4. TRAIN & CONSTRAIN

	Train (A)	Train (B)	Full Train	Test	Total
<i>Dataset Totals</i>					
# Samples	9,653	5,471	15,124	2,124	17,248
# Vocabulary	60,306	61,160	93,517	10,336	98,703
<i>Sample Averages</i>					
Avg. Topic Words	2.00	2.00	2.00	3.16	2.14
Avg. Paraphrase Length (words)	26.63	57.51	37.80	17.19	35.26
Avg. Twister Length (words)	23.90	53.60	34.64	14.98	32.22

Table 4.5: *Descriptive statistics of TwistList 2.0. The validation set is obtained by randomly selecting the same volume of data as the test set from the shuffled training set. Test set input phrases were extracted from TwistList 1.0 entries using RAKE. This allows comparison to human-authored gold standards in our later testing. Train (A) and Train (B) refer to samples generated with Prompt A and Prompt B, respectively.*

Prompt A) being much more succinct than the second example (which used Prompt B). Finally, regarding the paraphrasing used for the style-transfer task, GPT-3.5-Turbo maintains the grammaticality and coherence of the original tongue twister, but replaces much of the vocabulary. However, some terms remain for semantics, such as “BBC” in the first example and “non-verbal” in the second.

Due to exercising limited control over the paraphrasing stage, some synonym replacements may result in maintaining similar levels of sound overlap (at least in the word-initial alliterative sense). For instance, “pollster” has been replaced by “poller” in the first example, and “methods” has been replaced by “means” in the bottom example. The reason for the former example is similar to why the input phrase *may* be in the output of the final tongue twister, as the closest synonym to many words is a derivative of the word itself (i.e., variants of “poll-”).

CHAPTER 4. TRAIN & CONSTRAIN

TT ID:	68
Topic:	“public commentator”
Source:	GPT-3.5-Turbo
Prompt:	A
Tongue Twister:	The public-spirited BBC broadcasts presented by persistent presenters perplexed the publicist Berman the broadcaster and the pollster profoundly.
Paraphrase:	The civic-minded BBC airs hosted by determined hosts confused the publicist Berman the broadcaster and the poller deeply.
Twister ARPA-BET:	DH AH0 P AH1 B L IH0 K S P IH1 R IH0 T AH0 D B IY2 B IY0 S IY1 B R AO1 D K AE2 S T S P R IY0 Z EH1 N T AH0 D B AY1 P ER0 S IH1 S T AH0 N T P R EH1 Z AH0 N T ER0 Z P ER0 P L EH1 K S T DH AH0 P AH1 B L IH0 S IH0 S T B ER1 M AH0 N DH AH0 B R AO1 D K AE2 S T ER0 AH0 N D DH AH0 P OW1 L S T ER0 P R OW0 F AW1 N D L IY0.
Paraphrase ARPABET:	DH AH0 S IH1 V Z IH0 K EY2 D M IH0 D B IY2 B IY0 S IY1 EH1 R Z HH OW1 S T IH0 D B AY1 D IH0 T ER1 M AH0 N D HH OW1 S T S K AH0 N F Y UW1 Z D DH AH0 P AH1 B L IH0 S IH0 S T B ER1 M AH0 N DH AH0 B R AO1 D K AE2 S T ER0 AH0 N D DH AH0 P AA1 L ER0 D IY1 P L IY0.
TT ID:	10397
Topic:	“direct language”
Source:	GPT-3.5-Turbo
Prompt:	B
Tongue Twister:	Non-direct, multilingual, and non-verbal, Monolingual speakers were nominal. Some could speak Mandarin natively, While others relied on mediated means. The novelistic twist began when mispronouncing, Novellus words of complex morphology. Non-native speakers struggled with Macromedia, And nonnatives found Mandarin challenging. Nonaccrual in Marathi, they preferred, As net learners of a morphological world. But despite the hurdles, they persevered, Finding beauty in languages unfurled.
Paraphrase:	Indirect, polyglot, and non-verbal, Single-language speakers were nominal. Some could converse in Mandarin from birth, While others depended on mediated methods. The narrative turn commenced when mispronouncing, Novellus terms of intricate structure. Non-indigenous speakers struggled with Macromedia, And non-natives found Mandarin demanding. Non-accumulation in Marathi, they favored, As online learners of a structural world. But despite the obstacles, they persisted, Discovering elegance in languages unfolded.

Table 4.6: Example entries from *TwistList 2.0*. All entries also contain IPA transcriptions. In the interest of space, no phonetic transcriptions are presented for the second example.

4.5 Fine-Tuning Methodology

4.5.1 Tasks

Inspired by [Keh et al. \(2023\)](#), we define two settings for the task of tongue twister generation, *topic-to-twister* and *style-transfer*. In the former topic-to-twister setting, for a given topic phrase, we aim to generate a tongue twister T , whereby T comprises a sequence of words $\{w_1, w_2, \dots w_n\}$. The generated output must satisfy the following constraints: (i) the output should be semantically related to the input topic, (ii) the output should show maximal levels of phonological overlap across tokens, and (iii) the output should be grammatically valid. On the other hand, for the style-transfer setting we aim to generate a tongue twister again, $T = \{w_1, w_2, \dots w_n\}$, but provide as input a paraphrased (i.e., non-“twisty”) tongue twister and aim to convert it via style-transfer back into a tongue twister through learning to replace vocabulary to promote phoneme overlap.

4.5.2 Models

To realise the goal of tongue twister generation in topic-to-twister and style-transfer settings, we fine-tune a range of popular language models with varying parameter counts on TwistList 2.0.

- **GPT-2 117M** (`openai-community/gpt2-base`) ([Radford et al., 2019](#)) - A popular transformer-based text generation model consisting of both an encoder and a decoder.
- **DialoGPT 117M** (`microsoft/DialoGPT-small`) ([Zhang et al., 2020c](#)) - A version of GPT-2 that has been fine-tuned on an extensive corpus of dialogue to enable better conversational performance (and therefore better instruction-following).

- **BART 139M** (facebook/bart-base) (Lewis et al., 2020) - A popular denoising autoencoder model consisting of a BERT-like encoder (Devlin et al., 2019) with a GPT-2-like decoder (Radford et al., 2019).
- **Flan-T5 250M** (google/flan-t5-base) (Chung et al., 2024a) - A further refined instruction tuned version of T5 (Raffel et al., 2020).
- **ByT5 582M** (google/byt5-base) (Xue et al., 2022) - A version of T5 trained with byte level tokenisation, rather than subwords.
- **Baichuan 7B** (Yang et al., 2023) - A large-scale open-source LLM trained on English and Mandarin, achieving SoTA performance on many tasks at the time of writing, for a model of its size. Due to the significant size of Baichuan, we perform fine-tuning with the help of Low-Rank Adaptation (LoRA) (Hu et al., 2022).⁷
- **ChatGPT** (GPT-3.5-Turbo) (Ouyang et al., 2022) - A large language model fine-tuned for chat-based interactions and instruction following, which excels in few and zero-shot tasks. Importantly, we use ChatGPT in a zero-shot manner and do not perform any fine-tuning.⁸

Training Splits. To investigate the benefit of fine-tuning data volume for tongue twister generation, we train all models on various volumes of fine-tuning data, including 2k, 4k, 8k, and 13k samples from TwistList 2.0. The exception is Baichuan, which, due to computational requirements, we train only on the largest 13k split. The test set consists of 2,124 samples, covering the entirety of TwistList 1.0. Due to using TwistList

⁷<https://github.com/baichuan-inc/Baichuan2>

⁸GPT-3.5-Turbo was accessed for this purpose during August-September 2023 via the Chat Completions API.

1.0 entries as our test set, all reference-based metrics are compared to human-authored outputs, and human evaluation scores can be directly compared to human performance.

Prompt. For all topic-to-twister settings, we use the prompt “Generate a tongue twister on the topic of “[TOPIC]”” where “[TOPIC]” refers to the input topic phrase from the test set, and for all the style-transfer settings we use the prompt “Generate a tongue twister by rewriting the following text: [PARAPHRASE]” where [PARAPHRASE] refers to the non-literary paraphrase of a tongue twister from the test set.

Full fine-tuning hyperparameters can be found in Appendix 4.12.2.

4.5.3 Automatic Metric Suite

We present extensive automatic evaluation on the following metrics: Perplexity (PPL), BLEU (B-1/B-2/B-3/B-4) (Papineni et al., 2002), ROUGE (Ro-1/Ro-2/Ro-L) (Lin, 2004), and BERTScore precision, recall, and F-score (Zhang et al., 2020b) (BS-P/BS-R/BS-F). PPL, BLEU, and ROUGE are standard reference-based metrics to assess generation quality, whilst BERTScore assesses semantic similarity to a gold reference. Due to the nature of the task, many potential “gold standard” tongue twisters exist for any given input. Consequently, as with all creative generation works, reference-based metrics should be interpreted with caution.

Readability Tongue twisters are known for their intricate phoneme-level patterns and linguistic complexity, making them challenging to articulate correctly. Readability metrics can be used to indirectly assess whether tongue twisters are using more complex vocabulary and syntax in order to meet phonological constraints, such as when a selected phoneme only has a few obscure words that are related to the input topic.

CHAPTER 4. TRAIN & CONSTRAIN

We present a range of readability metrics, incorporating the Dale-Chall Readability Index (Chall and Dale, 1995) (Re-D), Flesch–Kincaid Readability Score (Flesch, 1948) (Re-F), Gunning-Fog Index (Gunning, 1971) (Re-G), and ARI Index (Smith and Senter, 1967) (Re-A). These readability metrics relate to the necessary comprehension level of a text’s audience. They calculate a numerical score based on factors such as sentence length and word complexity, offering a quantitative measure of how difficult a text is to understand. These further complement our other metrics by analysing linguistic complexity via means other than phonology.⁹ Extremely high scores on such metrics can likewise indicate the nonsensical nature of any particularly poor generations, owing to considering factors such as sentence length and syllable counts, with high scores being given to indefinitely long sequences due to the absence of sentence-final punctuation, as well as convoluted and overly complex syntactic structures and lexical choices.

Phonetics & Phonology We further develop our tongue twister measures and present phonemic edit distance metrics (PEDs) relying on the *weighted phonemic edit distance* function from the *PanPhon* package (Mortensen et al., 2016). This allows a more direct analysis of the phonology of our generated outputs by taking into consideration the articulatory similarities and differences between different phonemes. For example, the previous metrics from Loakman et al. (2023b), phoneme overlap (PO) and initial phoneme overlap (IPO), treat all phonemes as equidistant in feature space, resulting in a transition from /s/ to /ʃ/ being viewed as the same quality as a transition from /s/ to /g/. In the former case, the transition is from a voiceless alveolar fricative to a voiceless *post*-alveolar fricative (therefore moving the tongue slightly further back in the mouth), whilst in the latter case, the transition is between a voiceless alveolar fricative and a voiced velar plosive, resulting in a change of voicing, position, and manner of articulation.

⁹All readability metrics were implemented from <https://pypi.org/project/py-readability-metrics/>.

CHAPTER 4. TRAIN & CONSTRAIN

Consequently, using phonemic edit distance does not overly punish transitions between similar phonemes, which are more likely to encourage mispronunciations. As with PO and IPO, weighted edit distance is applied on both a word-initial and overall level, with *oPED* taking the *overall* average edit distance between every phoneme transition in the tongue twister, whilst *iPED* calculates the edit distance between word-*initial* phonemes, therefore being a more accurate measure of soft alliteration.

Formally, for iPED and oPED, let X and Y be two phonemes, represented as feature vectors, where each vector contains binary values representing the presence or absence of a phonological feature. Therefore, the weighted feature edit distance between X and Y is defined as: $D_{\text{wfe}}(X, Y) = \min_{\text{alignments}} \sum_{i=1}^n w_i \cdot d_i$, where X and Y are sequences of phonological features represented as feature vectors, n is the length of the longest common subsequence of X and Y , d_i represents the distance between the i -th feature in the alignment, and w_i is the weight assigned to the i -th feature. The minimum is taken over all possible alignments of the sequences X and Y . Each alignment assigns a distance d_i between corresponding features in the sequences. The weighted sum of these distances is computed, where each distance is multiplied by its corresponding weight w_i . For iPED, we calculate the mean distance across sequences of word-initial phonemes, whilst for oPED we calculate the mean distance when comparing all adjacent phonemes across the tongue twister. This metric provides a flexible way to measure the similarity between sequences of phonological features. In the case of oPED and iPED, the lower the score, the better, with a score of 0 relating to 100% overlap of a single phoneme throughout the tongue twister.

In addition to our novel metrics discussed above (iPED/oPED), we additionally present two metrics from [Loakman et al. \(2023b\)](#), phoneme overlap (PO) and initial phoneme overlap (IPO). PO refers to the average overlap of all phonemes across words, whereas IPO is the ratio of unique word-initial phonemes to the number of words.

Whilst these metrics reward longer outputs, we argue that all other things being equal, a longer tongue twister is better than a shorter one as it provides more entertainment and more opportunities for mispronunciation. Perfect scores on PO and IPO can be achieved by the repetition of a single word. Whilst this does not lead to high-quality outputs, these metrics are intended exclusively to be indicators of sound characteristics, rather than an overall guide to quality. In both cases, higher levels of overlap result in lower (i.e., better) scores, and the highest (i.e., worst) achievable score is 1.

4.5.4 Human Evaluation Protocol

Due to the limitations of automatic evaluation metrics in a creative domain where articulation abilities are tested, we also perform human evaluation. In line with [Loakman et al. \(2023a\)](#), we aim to be transparent in our human evaluation protocol. In total, five evaluators were asked to rate 20 outputs from the best performing fine-tuned baselines, Flan-T5 and ByT5, in addition to Baichuan, ChatGPT, and gold examples from TwistList 2.0 on the following criteria: *Relevance* (i.e., how relevant the tongue twister is given the input), *fluency* (i.e., how grammatically valid the output is), *difficulty of articulation* (i.e., how difficult a tongue twister is to say), *Coherence* (i.e., how much grammatical and semantic sense the output makes), *entertainment value* (i.e., how entertaining the output is, considering sounds and semantics), and a holistic *overall* criteria. All ratings were given on a 5-point scale where 1 equates to “poor” and 5 equates to “excellent”. Importantly, we include Flan-T5 and ByT5 trained on both 2k and 13k samples, allowing the analysis of the effects of an increased volume of fine-tuning data.

Evaluator Recruitment and Demographics In total, we recruit five evaluators via internal notices and word of mouth. All evaluators have university-level education to a minimum of undergraduate level in a range of fields, including linguistics, computer science, engineering, and animation, therefore representing a wide range of academic backgrounds. All evaluators are native speakers of English and report no language processing issues. Evaluators were provided with a 55 GBP Amazon gift card for their combined work on dataset quality control and output evaluation, totalling approximately four hours of work.

Materials Provided to Evaluators Evaluation was performed on an online platform. Participants were provided with a page of detailed instructions on how to navigate the platform and the order in which to perform evaluation tasks. Importantly, all human evaluation responses were given on a 1-5 rating scale where 1 equates to “poor” and 5 equates to “excellent” for how well a criterion is met. For example “*The tongue twister can be considered logically and semantically coherent.*” for the criteria of coherence. Additionally, those with non-linguistic backgrounds had terms clarified, such as the meaning of “prosaic” in the style-transfer task. Each evaluator rated 450 samples across all settings, consisting of 20 samples from seven models (i.e., Flan-T5_{2k}, Flan-T5_{13k}, ByT5_{2k}, ByT5_{13k}, Baichuan_{13k}, ChatGPT, and the gold references) on two tasks (280 in total), 50 quality control examples, and 120 examples from constrained decoding algorithm PACD (20 of which used constrained base GPT-2, 20 of which came from our constrained fine-tuned GPT-2, and 20 of which came from unconstrained GPT-2 trained on 13k samples, with the same for Baichuan).

4.6 Fine-Tuning Results

4.6.1 Automatic Evaluation

Model	PPL↓	B-1↑	B-2↑	B-3↑	B-4↑	Ro-1↑	Ro-2↑	Ro-L↑	Length
GPT-2 _{2k}	25.04	0.0391	0.0165	0.0083	0.0043	5.1255	0.6704	4.8414	51.81
DialoGPT _{2k}	23.74	0.0453	0.0187	0.0091	0.0047	6.3366	<u>1.0349</u>	5.9843	57.72
BART _{2k}	3.98	0.0749	0.0256	0.0105	0.0040	<u>8.1503</u>	0.7608	<u>7.4855</u>	44.07
Flan-T5 _{2k}	2.86	<u>0.1055</u>	<u>0.0545</u>	<u>0.0325</u>	<u>0.0209</u>	11.2417	2.1528	10.3641	25.60
ByT5 _{2k}	-	0.1440	0.0918	0.0639	0.0477	6.2582	1.0329	5.9611	47.56
GPT-2 _{4k}	24.84	0.0423	0.0168	0.0076	0.0037	5.7569	0.6818	5.5077	53.27
DialoGPT _{4k}	24.83	0.0422	0.0161	0.0071	0.0032	5.9712	0.6915	5.6415	55.91
BART _{4k}	4.09	0.0685	0.0215	0.0082	0.0031	<u>7.6023</u>	<u>0.7364</u>	<u>6.9026</u>	47.30
Flan-T5 _{4k}	2.93	<u>0.1090</u>	<u>0.0558</u>	<u>0.0328</u>	<u>0.0207</u>	10.5343	1.7701	9.8546	24.25
ByT5 _{4k}	-	0.1397	0.0900	0.0634	0.0479	6.8637	0.9744	6.4977	48.18
GPT-2 _{8k}	28.58	0.0430	0.0143	0.0060	0.0026	5.6638	0.5385	5.2935	54.23
DialoGPT _{8k}	24.54	0.0458	0.0172	0.0076	0.0035	6.3555	<u>0.7622</u>	5.9901	56.59
BART _{8k}	4.33	0.0633	0.0195	0.0072	0.0029	<u>6.8434</u>	0.5874	<u>6.1127</u>	49.68
Flan-T5 _{8k}	3.11	<u>0.1194</u>	<u>0.0600</u>	<u>0.0346</u>	<u>0.0217</u>	10.7728	1.3758	10.0456	22.97
ByT5 _{8k}	-	0.1531	0.0951	0.0647	0.0475	6.2602	0.7142	6.0271	40.96
GPT-2 _{13k}	28.77	0.0440	0.0152	0.0061	0.0025	5.9412	0.6024	5.5890	56.59
DialoGPT _{13k}	25.15	0.0504	0.0187	0.0078	0.0034	<u>6.7274</u>	0.8170	<u>6.3148</u>	60.05
BART _{13k}	4.16	0.0595	0.0152	0.0050	0.0017	6.6284	0.5090	5.9523	49.79
Flan-T5 _{13k}	3.10	<u>0.1189</u>	<u>0.0571</u>	<u>0.0319</u>	<u>0.0192</u>	10.0983	<u>1.1632</u>	9.2976	23.87
ByT5 _{13k}	-	0.1609*	0.0988	0.0660	0.0476	6.4857	0.6073	6.2272	40.05
Baichuan+LoRA _{13k}	14.54	0.0463	0.0227	0.0131	0.0080	6.2215	1.3878	5.9212	51.76
ChatGPT	-	0.1577	0.1073*	0.0788*	0.0585*	26.2949*	13.2789*	23.7763*	24.46

Table 4.7: Results of automatic evaluation on general metrics for the topic-to-twister task setting. The mean length of tongue twisters is 35.65/34.14/14.98 words for the train/evaluation/test sets, respectively. Results in **bold** represent the best performance for a given training data quantity, and underlined presents the second-best performance. Results followed by an asterisk * denote the overall best performance. The perplexity of ByT5 is not applicable, as the tokens are characters rather than sub-words.

Topic-to-Twister We present the results for automatic evaluation in the topic-to-twister task setting in Table 4.7 and Table 4.8. Firstly, regarding the reference-based metrics (BLEU, ROUGE, and BERTScore) we see clear performance differences across all models. On average across most metrics, the performance from worst to best is ordered as GPT-2, DialoGPT, BART, ByT5, and Flan-T5 for our fine-tuned models. However, Baichuan’s performance is variable, mostly outperforming BART and underperforming Flan-T5 and ByT5. However, Baichuan performs worse than BART considering a

CHAPTER 4. TRAIN & CONSTRAIN

Model	BS-P↑	BS-R↑	BS-F1↑	IPO↓	PO↓	iPED↓	oPED↓	Re-D	Re-F	Re-G	Re-A
GPT-2 _{2k}	0.7543	0.8265	0.7883	0.0849	0.0617*	3.8717	5.9131	17.08	45.79	48.71	59.52
DialoGPT _{2k}	0.7680	0.8310	0.7978	<u>0.0954</u>	<u>0.0660</u>	<u>3.8308</u>	<u>5.9032</u>	12.79	16.52	18.54	21.21
BART _{2k}	<u>0.7938</u>	<u>0.8384</u>	<u>0.8153</u>	0.3183	0.1848	4.5168	5.9219	14.21	11.70	14.12	11.51
Flan-T5 _{2k}	0.8075	0.8440	0.8249	0.2129	0.1610	4.0828	5.9141	13.10	14.68	16.19	18.25
ByT5 _{2k}	0.7705	0.8301	0.7982	0.1207	0.1021	2.5398	5.9048	14.61	28.54	30.20	38.35
GPT-2 _{4k}	0.7648	0.8303	0.7958	0.0881	0.0683	4.1320	5.9440	13.51	23.08	25.14	28.89
DialoGPT _{4k}	0.7648	0.8300	0.7956	<u>0.0984</u>	<u>0.0701</u>	<u>3.7237</u>	5.9205	13.01	17.39	19.26	21.86
BART _{4k}	<u>0.7962</u>	<u>0.8363</u>	<u>0.8156</u>	0.2821	0.1620	4.7537	5.9693	12.77	12.46	14.11	13.69
Flan-T5 _{4k}	0.8071	0.8436	0.8244	0.2089	0.1657	3.8280	5.8899	13.29	16.63	17.76	20.28
ByT5 _{4k}	0.7636	0.8297	0.7941	0.1250	0.0987	2.2248	<u>5.9197</u>	16.0915	35.6761	36.9863	47.4017
GPT-2 _{8k}	0.7712	0.8315	0.7998	0.1224	0.0841	4.3355	5.9751	12.57	14.93	16.30	18.00
DialoGPT _{8k}	0.7745	0.8327	0.8022	0.1116	0.0775	<u>4.0111</u>	5.8952	11.99	13.42	15.14	16.83
BART _{8k}	<u>0.7928</u>	<u>0.8343</u>	<u>0.8129</u>	0.2831	0.1675	4.9463	5.9693	11.97	12.36	13.71	13.86
Flan-T5 _{8k}	0.8160	0.8462	0.8304	0.2378	0.1866	4.1542	5.8669	12.74	14.40	16.24	17.31
ByT5 _{8k}	0.7711	0.8306	0.7986	0.1341	0.1087	2.1024*	<u>5.8804</u>	16.37	31.98	33.88	42.47
GPT-2 _{13k}	0.7748	0.8322	0.8021	0.1261	0.0859	<u>4.2671</u>	5.8985	12.06	13.46	15.21	16.64
DialoGPT _{13k}	0.7795	<u>0.8336</u>	0.8053	<u>0.1249</u>	<u>0.0839</u>	4.4665	5.9407	11.30	12.85	14.95	15.55
BART _{13k}	<u>0.7925</u>	0.8319	<u>0.8115</u>	0.2662	0.1489	4.8871	5.8184*	12.14	11.75	13.57	13.83
Flan-T5 _{13k}	0.8181	0.8468	0.8319	0.2568	0.1996	4.3122	5.8932	12.31	12.50	14.03	14.97
ByT5 _{13k}	0.7742	0.8320	0.8009	0.1554	0.1217	2.5520	<u>5.8374</u>	15.77	31.22	34.00	39.33
Baichuan+LoRA _{13k}	0.7658	0.8259	0.7940	0.0689*	0.0657	2.6752	5.9951	16.41	34.37	31.73	44.59
ChatGPT	0.8401*	0.8613*	0.8503*	0.3477	0.2991	4.0547	5.9153	9.83	9.37	10.53	11.48
Brown Corpus Prose	-	-	-	0.4870	0.2275	5.1431	5.9561	11.09	13.15	15.56	16.60

Table 4.8: Results of automatic evaluation on tongue twister related metrics for the topic-to-twister setting. Results in **bold** represent the best performance for a given training data quantity, and underlined presents the second-best performance. Results followed by an asterisk * denote the overall best performance. Brown Corpus Prose presents the scores of standard prose from the Brown Corpus from NLTK (across 602 sentences with a minimum length of 25 words).

range of metrics (B-1, Ro-1, Ro-L, BS-P, BS-R, BS-F1) when considering the same training data amount (13k). Flan-T5 and ByT5 alternate in performance, with the latter byte-based model performing better on BLEU but worse on ROUGE. When specifically considering changes alongside an increase in training data, in Table 4.7 we see little to no improvement in reference-based metrics within the topic-to-twister setting, with performance decreasing as training data increases in some cases. However, reference-based metrics are imperfect for the task of creative language generation due to the one-to-many dilemma prevalent in many NLG tasks (Gupta et al., 2019), whereby there are numerous potential tongue twisters to generate from any given input topic. This is also particularly true when considering TwistList 2.0, where the topics are often only related to the input phrase in a high-level conceptual fashion, as we do not enforce

CHAPTER 4. TRAIN & CONSTRAIN

the generation of the topic phrase within the tongue twister itself. Finally, ChatGPT demonstrates significantly higher performance than any of our fine-tuned models.

Regarding the readability and phonemic metrics presented in Table 4.8, a range of patterns can be observed. Firstly, for naive PO and IPO, we see GPT-2 “outperform” BART and the T5-based models almost universally. However, these naive metrics do not take into account more sophisticated phonemic characteristics. When considering the new metrics based on phonemic edit distance (iPED and oPED) we see less distinction across any of the presented models. However, Baichuan performs significantly better than any of the other models trained on 13k samples in the word-initial IPO and iPED metrics, suggesting high levels of word-initial phoneme overlap across words compared to other models. Likewise, we see that the byte-based ByT5 model outperforms Flan-T5 by a significant margin on phoneme-based metrics, suggesting that the finer-grained tokenisation of ByT5 is preferable for identifying the sound patterns implicitly included in tongue twisters via grapheme combinations. ChatGPT, on the other hand, scores the highest on the IPO and iPED metrics (i.e., worst). This, however, is not in itself indicative of poor tongue twisters, but rather the model’s focus on producing comprehensible and grammatical text, being less likely to fall victim to degenerate patterns such as single-word repetition to achieve overlap. As with the reference-based metrics, sole reliance on phoneme-based metrics can also be misleading. For example, intuitively, prior works have demonstrated that high scores (i.e., lower values) can be achieved in word-initial-based metrics simply by repeating the same word, rather than producing a complex and entertaining tongue twister. Regarding readability scores, we see GPT-2 and Baichuan score highly (i.e., generate outputs that are difficult to read) compared to other models. However, high readability metrics can be indicative of numerous things, including desirable behaviour (e.g., using complex structures and sophisticated polysyllabic vocabulary) and undesirable behaviour (e.g., producing

CHAPTER 4. TRAIN & CONSTRAIN

Model	PPL↓	B-1↑	B-2↑	B-3↑	B-4↑	Ro-1↑	Ro-2↑	Ro-L↑	Length
GPT-2 _{2k}	15.59	0.1055	0.0738	0.0534	0.0397	15.2637	6.2852	14.8342	62.53
DialoGPT _{2k}	14.54	0.1001	0.0688	0.0491	0.0359	14.4003	5.7688	13.9574	67.65
BART _{2k}	1.96	0.2498	0.1797	0.1324	0.0997	25.4414	11.3534	24.8061	36.87
Flan-T5 _{2k}	1.64	<u>0.5597</u>	<u>0.4426</u>	<u>0.3591</u>	<u>0.2964</u>	48.7184	24.3008	47.9246	15.18
ByT5 _{2k}	-	0.6770	0.5847	0.5210	0.4731	<u>46.8991</u>	<u>22.3860</u>	<u>46.0446</u>	16.53
GPT-2 _{4k}	15.15	0.1161	0.0837	0.0621	0.0472	16.7069	7.4608	16.3085	62.70
DialoGPT _{4k}	13.96	0.1099	0.0773	0.0565	0.0425	15.8087	6.7687	15.3450	65.72
BART _{4k}	2.00	0.2561	0.1859	0.1378	0.1041	26.3945	12.0011	25.8563	37.72
Flan-T5 _{4k}	1.62	<u>0.5500</u>	<u>0.4351</u>	<u>0.3532</u>	<u>0.2917</u>	<u>48.8662</u>	<u>24.6492</u>	<u>48.1269</u>	14.99
ByT5 _{4k}	-	0.7177	0.6242	0.5596	0.5111	49.1321	24.7326	48.2396	15.90
GPT-2 _{8k}	14.26	0.1152	0.0833	0.0622	0.0476	16.7724	7.5582	16.3746	63.30
DialoGPT _{8k}	13.19	0.1096	0.0772	0.0572	0.0438	15.7687	7.0081	15.248	67.10
BART _{8k}	1.95	0.2448	0.1803	0.1358	0.1043	30.1371	14.4970	29.4553	37.50
Flan-T5 _{8k}	1.63	<u>0.5699</u>	<u>0.4571</u>	<u>0.3756</u>	<u>0.3135</u>	51.1949	26.9544	50.4617	14.91
ByT5 _{8k}	-	0.7124	0.6224	0.5605	0.5136	<u>50.4737</u>	<u>25.9246</u>	<u>49.5851</u>	15.85
GPT-2 _{13k}	13.55	0.1229	0.0901	0.0683	0.0529	17.9146	8.4624	17.4801	60.71
DialoGPT _{13k}	12.84	0.1114	0.0796	0.0597	0.0462	16.2061	7.3163	15.7076	66.90
BART _{13k}	1.82	0.2298	0.1717	0.1309	0.1014	34.0662	16.9713	33.3075	37.78
Flan-T5 _{13k}	1.61	0.5832	0.4704	0.3885	0.3258	<u>52.7077</u>	<u>28.6604</u>	<u>51.8808</u>	14.82
ByT5 _{13k}	-	0.7356*	0.6465*	0.5846*	0.5376*	51.9886	27.7015	51.1895	15.55
Baichuan+LoRA _{13k}	5.15	<u>0.6046</u>	<u>0.5006</u>	<u>0.4229</u>	<u>0.3618</u>	60.1989*	37.1423*	59.1040*	15.27
ChatGPT	-	0.3288	0.2167	0.1491	0.1057	39.3536	13.9654	34.9884	17.43

Table 4.9: Results of automatic evaluation on general metrics for the style-transfer setting. The mean length of the tongue twisters is 35.65/34.14/14.98 words for train/evaluation/test, respectively. Results in **bold** represent the best performance for a given training data quantity, and underlined present the second-best performance. Results followed by an asterisk * denote the overall best performance. The perplexity of ByT5 is not applicable, as the tokens are characters rather than sub-words.

gratuitously long sentences). It is for this reason that we exclude an indicator of the preferred metric direction in the case of our readability metrics, yet include the scores for completeness.

Style-Transfer We present the results for automatic evaluation in the style-transfer task setting in Table 4.9 and Table 4.10. Overall, much the same pattern is observed as with the topic-to-twister setting, with performance ordering of DialoGPT, GPT-2, BART, ByT5, and Flan-T5 across referenced metrics, with the T5-based models alternating in ranking. However, unlike in the previous setting, style-transfer appears to favour Baichuan, which presents the highest scores on ROUGE (when considering models also trained on 13k samples), whilst ByT5 performs the best on BLEU. Additionally,

CHAPTER 4. TRAIN & CONSTRAIN

Model	BS-P $\uparrow$	BS-R $\uparrow$	BS-F1 $\uparrow$	IPO $\downarrow$	PO $\downarrow$	iPED $\downarrow$	oPED $\downarrow$	Re-D	Re-F	Re-G	Re-A
GPT-2 _{2k}	0.8170	0.8662	0.8405	<u>0.1264</u>	<u>0.0830</u>	4.7213	5.9688	10.66	7.45	9.52	6.94
DialoGPT _{2k}	0.8172	0.8634	0.8394	0.1177*	0.0786*	4.8873	<u>5.9570</u>	9.92	7.66	9.76	7.74
BART _{2k}	0.8162	0.9031	0.8570	0.5112	0.2474	<u>4.5594</u>	5.8120	14.76	12.21	14.13	11.47
Flan-T5 _{2k}	0.9281	0.9275	0.9277	0.5465	0.4290	4.7362	5.9696	10.51	6.08	7.77	5.07
ByT5 _{2k}	<u>0.9166</u>	<u>0.9171</u>	<u>0.9166</u>	0.4392	0.3949	4.2451	5.9703	10.78	6.81	8.72	5.90
GPT-2 _{4k}	<u>0.8247</u>	<u>0.8726</u>	<u>0.8477</u>	0.1221	0.0834	4.7674	<u>5.9775</u>	10.88	7.02	8.93	6.15
DialoGPT _{4k}	0.8213	0.8668	0.8432	<u>0.1280</u>	<u>0.0847</u>	4.8948	5.9825	10.34	7.83	10.16	7.35
BART _{4k}	0.8157	0.9026	0.8565	0.4028	0.2267	3.8435	5.7562*	16.74	14.10	15.72	14.19
Flan-T5 _{4k}	0.9262	0.9266	0.9263	0.5362	0.4290	4.6793	5.9743	10.72	6.10	7.83	5.05
ByT5 _{4k}	<u>0.9198</u>	<u>0.9199</u>	<u>0.9197</u>	0.4436	0.4011	<u>4.2007</u>	<u>5.9553</u>	10.47	6.40	8.03	5.59
GPT-2 _{8k}	0.8231	0.8714	0.8463	0.1188	0.0821	4.7338	5.9779	10.83	7.30	9.23	6.58
DialoGPT _{8k}	0.8183	0.8638	0.8402	<u>0.1329</u>	<u>0.0876</u>	4.9483	5.9787	10.10	8.23	10.47	8.03
BART _{8k}	0.8073	<u>0.9073</u>	<u>0.8538</u>	0.4451	0.2620	5.1885	5.8138	14.46	13.17	13.46	11.00
Flan-T5 _{8k}	0.9289	0.9290	0.9288	0.5262	0.4289	<u>4.6310</u>	5.9707	10.67	5.95	7.67	4.99
ByT5 _{8k}	<u>0.9208</u>	<u>0.9219</u>	<u>0.9212</u>	0.4549	0.4063	4.2830	<u>5.9598</u>	10.68	6.52	8.04	5.85
GPT-2 _{13k}	<u>0.8258</u>	<u>0.8738</u>	<u>0.8489</u>	0.1256	0.0870	4.7119	5.9536	11.65	8.59	10.84	7.73
DialoGPT _{13k}	0.8216	0.8676	0.8437	<u>0.1348</u>	<u>0.0891</u>	5.0811	5.9733	10.51	9.01	11.03	8.79
BART _{13k}	0.7903	0.9141	0.8470	0.4201	0.3216	4.0013*	6.1303	17.97	13.25	16.91	9.82
Flan-T5 _{13k}	<u>0.9311</u>	<u>0.9309</u>	<u>0.9309</u>	0.5226	0.4288	4.5770	5.9681	10.64	5.98	7.63	5.06
ByT5 _{13k}	0.9236	0.9246	0.9240	0.4619	0.4128	<u>4.2982</u>	<u>5.9634</u>	10.73	6.35	7.93	5.67
Baichuan+LoRA _{13k}	0.9442*	0.9411*	0.9425*	0.4908	0.4236	4.4789	5.9771	9.98	5.60	7.15	4.62
ChatGPT	0.8851	0.8898	0.8873	0.5495	0.3948	4.7725	5.9445	10.37	7.67	9.06	8.25
Brown Corpus Prose	-	-	-	0.4870	0.2275	5.1431	5.9561	11.09	13.15	15.56	16.60

Table 4.10: Results of automatic evaluation on tongue-twister-related metrics for the style-transfer setting. Results in **bold** represent the best performance for a given training data quantity, and underlined presents the second-best performance. Results followed by an asterisk * denote the overall best performance. Brown Corpus Prose presents the scores of standard prose from the Brown Corpus from NLTK (across 602 sentences with a minimum length of 25 words).

scores are universally higher than in the topic-to-twister setting. This is to be expected, as the style-transfer task provides a structure for the generated tongue twister based on the length and word choices present in the starting paraphrase. Consequently, increased amounts of overlap between the desired output and the gold reference are expected because not all words require modification. In contrast to the topic-to-twister setting, ChatGPT does not outperform all other models; instead being beaten by Baichuan and Flan-T5 in most cases. In terms of the impact of fine-tuning data volume, a clearer performance increase as data volume increases can be observed for the reference-based metrics. Regarding the readability and phonemic metrics presented in Table 4.10, we again see GPT-2 scoring well on the naive phoneme-based metrics (IPO/PO), but all models perform similarly on the informed iPED/oPED metrics. Regarding readability,

scores are seen to be lower overall than in the topic-to-twister setting, suggesting more legible text. On one hand, this may indicate that the style has failed to transfer, with the paraphrase representing a well-written standard non-literary text (and therefore the model has resorted to reconstructing the input). On the other hand, this may also be an artefact of following the original structure of the non-literary paraphrase, therefore avoiding unnaturally long sentences and nonsensical outputs.

4.6.2 Human Evaluation

Topic-to-Twister. Firstly, for the topic-to-twister setting in Table 4.11, the highest scores on all criteria go to the human-authored gold samples, or those generated with ChatGPT. With a rating of 3 being the midpoint for the given criteria, it is evident that the fine-tuned unconstrained generation models struggle with the open-ended topic-to-twister task setting. However, when investigating the fine-tuned model performance, patterns emerge that indicate the benefit of having such an extensive dataset as TwistList 2.0. For instance, Flan-T5 benefits from additional training samples, particularly on the metrics of fluency and coherence, and moderately in entertainment and the holistic overall rating. These findings for fluency and coherence are intuitive, as additional training samples increase the likelihood of generating grammatical and semantically coherent outputs due to the increased training data through which to learn. On the other hand, articulation and entertainment refer to metrics that are more abstract and consequently difficult to learn from training data. Relevance is the only metric shown to decrease when moving from 2k training samples to 13k, and we hypothesise that this is due to the increased volume of training data with more abstract links between the input topic and the output, therefore decreasing the likelihood of input keywords being directly present in the output (which is a straightforward way of performing well on relevance metrics). Overall, ByT5 is preferable to Flan-T5, outperforming it in human

Topic-to-Twister (1 to 5)	Model						
	Flan-T5 _{2k}	Flan-T5 _{13k}	ByT5 _{2k}	ByT5 _{13k}	Baichuan _{13k}	ChatGPT	Gold
Relevance	2.077***	1.625***	2.180**	2.070**	1.688**	4.824**	<u>4.647**</u>
Articulation	1.800**	1.882**	3.050**	<u>3.290***</u>	1.176**	2.667**	3.375*
Fluency	2.000**	3.462**	2.290**	3.380**	1.450**	<u>4.632**</u>	4.944**
Coherence	1.200**	2.133**	1.610**	1.930**	1.118**	<u>4.333**</u>	4.444***
Entertainment	1.200*	1.833*	1.620**	2.070**	1.000*	3.267*	<u>3.077**</u>
Overall	1.063*	1.888*	1.800**	2.300**	1.316**	<u>3.538**</u>	3.909**

Table 4.11: Results of human evaluation in the topic-to-twister task setting. The best scores are in **bold**, and the second-best are underlined. We calculate Fleiss’ κ for each metric, and mark the agreement fair* (0.21 - 0.40), moderate** (0.41-0.60), and substantial*** (0.61 - 0.80).

evaluation on the criteria of relevance, articulation, entertainment, and overall, but underperforming in regard to fluency and coherence in the 13k instance. This sheds light on how humans perceive quality tongue twisters, with articulation difficulty being more integral to a tongue twister than grammatical validity and semantic coherence. Finally, we see Baichuan perform very poorly in the topic-to-twister setting (which is explored further in §4.9), frequently opting to repeat the input topic phrase continuously, therefore artificially increasing relevance scores, but performing poorly on all other metrics.

Style-Transfer On the other hand, performance is better overall for the style-transfer setting as facilitated by the additional high-quality paraphrases included in TwistList 2.0. We hypothesise that due to style-transfer requiring access to a well-formed input, this acts as an extended, prescriptive form of an input topic, where the entire tongue twister is predefined in structure and semantics. As a result, Baichuan outperforms ChatGPT on the criteria of entertainment and relevance, whilst Flan-T5 trained on the 13k split, outperforms ChatGPT regarding the semantic coherence of the output. Interestingly, ByT5 trained on only 2k examples performs more closely to the 13k version than is seen in Flan-T5, suggesting the alternative tokenisation approach allows ByT5 to learn the relevant patterns from fewer examples. Importantly, we see Baichuan

Style-Transfer (1 to 5)	Model						
	Flan-T5 _{2k}	Flan-T5 _{13k}	ByT5 _{2k}	ByT5 _{13k}	Baichuan _{13k}	ChatGPT	Gold
Relevance	4.467*	4.714**	4.090*	4.160*	<u>4.882</u> **	4.692**	5.000*
Articulation	1.462**	2.231**	3.710**	3.520*	2.250**	<u>2.471</u> **	3.313**
Fluency	4.611**	4.895***	4.150**	4.270**	4.800*	5.000 **	<u>4.950</u> **
Coherence	4.188**	<u>4.733</u> **	3.500**	3.550**	4.375**	3.929**	4.786*
Entertainment	1.733*	2.000*	3.400**	3.240*	<u>2.846</u> *	2.583*	3.455 **
Overall	2.308*	3.000*	3.500**	3.510*	3.333*	<u>3.500</u> *	3.941 **

Table 4.12: Results of human evaluation in the style-transfer task setting. The best scores are in **bold**, and the second-best are underlined. We calculate Fleiss’ κ for each metric, and mark the agreement fair* (0.21 - 0.40), moderate** (0.41-0.60), and substantial*** (0.61 - 0.80).

perform much better in the style-transfer task than in the topic-to-twister setting, suggesting that Baichuan requires much more explicit instruction to understand a given prompt and generate high-quality outputs.

4.6.3 Human vs. Automatic Metrics

Spearman’s ρ	Automatic Metrics							
	Re-D	Re-F	Re-G	Re-A	IPO	PO	iPED	oPED
Relevance	-.395	-.442	-.423	-.494	.394	.496	-.004	.104
Articulation	-.218	-.271	-.224	-.218	-.063	-.087	-.226	.110
Fluency	-.400	-.590	-.564	-.630	.516	.674	-.001	-.013
Coherence	-.320	-.482	-.551	.482	-.486	.622	-.003	.064
Entertainment	.301	-.287	-.267	-.278	.151	.262	-.077	-.106
Overall	.342	.386	.372	.386	.211	.361	-.129	.055

Table 4.13: Spearman correlation coefficients for each human evaluation criterion with reference-free automatic metrics. We take the mean score across evaluators for the combined unconstrained topic-to-twister and style-transfer task settings.

Human-Automatic Correlation. To see the effectiveness of the selected automatic metrics, Spearman correlations are calculated between our eight reference-free metrics, including the readability (i.e., Re-D, Re-F, Re-G, and Re-A) and phonemic metrics (i.e., IPO, PO, iPED, and oPED), against the human evaluation ratings for all criteria (i.e., relevance, articulation, fluency, coherence, entertainment, and overall). Correlations

CHAPTER 4. TRAIN & CONSTRAIN

are presented in Table 4.13. To evaluate the predictive power of the automatic metrics for human ratings, a standard multiple linear regression model was developed for each criterion, defined as follows:

$$y_i = \beta_0 + \sum_{j=1}^p \beta_j X_{ij} + \epsilon_i \quad (4.1)$$

where y_i represents the human score, β_0 is the intercept, β_j are the coefficients for the p automatic metrics, and ϵ_i is the error term. No regularisation was applied to the model. The R^2 values, which indicate the proportion of variance explained by the automatic metrics, were validated through 5-fold cross-validation to ensure the stability and significance of the predictions. The average R^2 values across the folds were as follows: Fluency ($R^2 = .553$), coherence ($R^2 = .496$), relevance ($R^2 = .393$), overall ($R^2 = .260$), articulation ($R^2 = .241$), and entertainment ($R^2 = .152$). Furthermore, all coefficients in the models were found to be statistically significant, with p-values below 0.01 ($\alpha = 0.01$), indicating that each of the automatic metrics significantly contributes to the prediction of human ratings.

Overall, we intuitively see the entertainment criterion being the hardest to predict due to inherent subjectivity. On the other hand, the naive phonemic metrics (IPO and PO) demonstrate moderate correlations with relevance, fluency, and coherence. This is due to the high relevance of tongue twisters often being seen in examples where the input topic is simply repeated, which results in high levels of phoneme overlap. Similarly, high fluency scores are given to sentences that better reflect standard prosaic text, which score lower on IPO and PO. Moreover, the informed phonemic metrics (iPED/oPED) show little correlation with human results on these criteria, but iPED demonstrates moderate correlation with articulatory difficulty. However, articulatory difficulty still remains challenging to predict. We hypothesise that this may be related

to the “visual tongue twister” effect (McCutchen and Perfetti, 1982). This is due to human evaluation being performed online and asynchronously, where we cannot force participants to speak aloud each sample. Consequently, naive metrics may correlate better with human judgments, as human ratings were confounded by the visual impact of the tongue twister (for example, seeing a particular grapheme repeated, representing the same sound). Furthermore, there are additional effects from the influence of factors such as fluency and coherence affecting articulation due to violating expectations and reducing legibility. Consequently, iPED/oPED should be used as indicators of text resembling a tongue twister, but not as a holistic metric for overall quality. It is clear from comparison with prosaic text that the phonemic metrics can differentiate the specific characteristics of tongue twisters from those of standard text, but phonological complexity is not the sole contributor to the perception of articulatory difficulty.

LLM-as-a-Judge. We additionally perform evaluation on the same samples presented to human evaluators in the unconstrained topic-to-twister and style-transfer settings, using GPT-4o. To do this, the model is prompted with the same rubric presented to human evaluators (see Appendix 4.12.1).¹⁰ Overall, we see moderate-to-high correlations between model scores and human-assigned scores for all criteria: relevance ($\rho = .671$), articulation ($\rho = .653$), fluency ($\rho = .768$), coherence ($\rho = .658$), entertainment ($\rho = .716$), and overall ($\rho = .776$).¹¹ This demonstrates that the human evaluation rubric is well defined and can be effectively followed by state-of-the-art LLMs.

¹⁰Specifically, `gpt-4o-2024-05-13` via the API.

¹¹All significant at $\alpha = 0.01$.

4.7 Constrained Decoding Methodology

In the following section, we present a constrained decoding-based approach to tongue twister generation. In contrast to §4.5, we focus exclusively on the topic-to-twister setting due to the text-continuation nature of our decoding approach. The benefit of this algorithm, in contrast to the fine-tuned models presented in §4.5, is that constrained decoding *guarantees* that only desirable tokens appear in the output due to the layering of hard phoneme-based constraints. Additionally, due to the interaction with language model token predictions, this process can be applied to any autoregressive language model, including both pre-trained base models and those that have been fine-tuned.

4.7.1 Task

For a given input, we aim to generate a tongue twister T , whereby T comprises a sequence of words $\{w_1, w_2, \dots, w_n\}$. In contrast with the previous section, T is a continuation of the input that we generate token by token, evaluating the model’s next token predictions at each step. As per §4.5, the generated output must satisfy the following constraints: (i) the output should be semantically related to the input topic phrase; (ii) the output should show maximal levels of phonemic overlap across tokens; and (iii) the output should be grammatically valid.

4.7.2 Phoneme-Aware Constrained Decoding (PACD)

Algorithm 1 Phoneme-Aware Constrained Decoding (PACD)

```

1: for each  $s$  in  $S$  do
2:    $ph_1 = \text{G2P}(\text{topic in } s)[0]$ 
3:    $ph_2 = \arg \min(PED(ph_1, WIP \setminus \{ph_1\}))$ 
4:   while  $\text{len}(s^*) < \text{max\_length}$  do
5:     Retrieve next word probabilities  $P = \{p_1, \dots, p_n\}$  from  $LM(s^*)$ 
6:     for each  $rank, p$  in  $\text{ENUMERATE}(P)$  do
7:       if  $p$  in  $F$  and  $rank \leq \text{function\_window}$  then
8:         append  $p$  to  $s^*$ 
9:       break
10:    end if
11:     $\text{candidates} = []$ 
12:    if  $\text{len}(p) > \text{min\_stem\_length}$  and  $\text{G2P}(p)[0] \in \{ph_1, ph_2\}$  then
13:      append  $p$  to  $\text{candidates}$ 
14:       $\text{temp\_prompt} = s^* + p$ 
15:      for  $i$  in  $\text{range}(4)$  do
16:         $\text{next\_token} = LM(\text{temp\_prompt})$ 
17:        if  $\text{next\_token.isalpha}()$  and  $\text{next\_token}[0] \neq " "$  then
18:           $\text{longest} = "" \cdot \text{join}(\text{candidates})$ 
19:           $\text{longest} += \text{next\_token}$ 
20:          append  $\text{longest}$  to  $\text{candidates}$ 
21:        end if
22:        if  $\text{next\_token}[-1] == " "$  then
23:          break
24:        end if
25:      end for
26:       $\text{temp\_prompt} = ""$ 
27:      for  $\text{candidate}$  in  $\text{candidates.sort(longest-to-shortest)}$  do
28:        if  $\text{candidate} \in D$  and  $\text{COUNT}(\text{candidate in } s^*) < \text{max\_repetition}$ 
then
29:          append  $\text{candidate}$  to  $s^*$ 
30:        end if
31:      end for
32:    end if
33:  end for
34: end while
35: end for

```

CHAPTER 4. TRAIN & CONSTRAIN

We present an outline of phoneme-aware constrained decoding (PACD) in Algorithm 1.¹² To summarise, for every starting prompt s in our test set S , we firstly perform grapheme-to-phoneme conversion $G2P$ with g2pE and extract the initial phoneme of the first word in the topic phrase of s to increase the likelihood of retrieving a semantically related output (with the phoneme denoted as ph_1). However, where the selected phoneme is not a valid consonant, we randomly select a phoneme from a list of phonotactically legal word-initial consonant phonemes for English, WIP . Following this, we calculate the weighted phonemic edit distance (PED) between ph_1 and all other legal word-initial phonemes, selecting the lowest scoring (i.e., most similar) as the secondary phoneme ph_2 (analogous to the system in §4.4 for TwisterLister). Following this, we autoregressively generate new tokens up to the limit defined by max_length based on numerous criteria. To do this, we feed the starting prompt s into the language model of choice, LM , and retrieve the next token probabilities P . Then, in descending order (i.e., most-probable to least-probable) we iterate through predictions $p \in P$ until specific criteria are met. Firstly, to increase the likelihood of generating grammatical output, if a function word (i.e., article, pronoun, conjunction, preposition, or auxiliary verb) from a predefined list F is within the range defined by $function_window$, it is allowed to generate. For example, when generating the first word, if $function_window$ is set to 3, and “The” is the token with the 2nd highest probability, it is allowed to generate due to being within the top 3.

To account for subword tokenisation in non-function words, we check that the predicted token is longer than the limit defined by min_stem_length . Not limiting this results in a reliance on outputting the grapheme that most closely corresponds to a desired phoneme (rather than a standard morpheme), significantly increasing inference time and decreasing output quality. The phoneme constraints are then

¹²PACD is intended to be pronounced as “packed”.

CHAPTER 4. TRAIN & CONSTRAIN

enforced by feeding the predicted word stem p into the grapheme-to-phoneme model $G2P$ and comparing the first phoneme to ph_1 and ph_2 , continuing if it matches either. Consequently, in this stage, we have ensured that generated words are either closed-class grammatical function words or words that start with one of two similar sounds. Following this, we optionally engage the subword loop (lines 15-25 in Algorithm 1). Within this loop, we temporarily append the candidate stem to the current prompt s^* to create $temp_prompt$. Following this, the $temp_prompt$ is fed to the language model LM , where the highest probability token is taken, $next_token$. We then check that $next_token$ is alphabetical and does not start with whitespace (as this would indicate a new word). If this is the case, it is appended to $temp_prompt$ and the loop is repeated (up to four times, allowing words that consist of one to five subwords). Within this loop, a list of potential words is created, $candidates$, by appending the concatenated subwords (e.g., [“anti”, “antidis”, “antidisestablish”, “antidisestablishment”, “antidisestablishmentarian”]). The loop is terminated early if a predicted $next_token$ ends with whitespace, as this indicates that the model has predicted the end of the current word. Once we have the list of $candidates$, we iterate through in descending length order, assessing the following criteria.

Firstly, we check that each $candidate \in candidates$ is longer (in characters) or equal to the length defined by min_word_length and that $candidate \in D$, where D is the English dictionary as defined by the Enchant Python package.¹³ Once this check is complete, we ensure that we have not already generated this specific $candidate$ more times than permitted by $max_repetition$, in order to avoid falling into the perpetual loop of repeating the same words (see §4.9). s^* denotes the starting prompt s with newly generated tokens appended, which is to say that $s^* - s$ equals only the LM-generated words. Finally, if no $candidate$ meets the criteria, we increase $rank$ by looking at

¹³Available at <https://pypi.org/project/pyenchant/>.

CHAPTER 4. TRAIN & CONSTRAIN

the next-best prediction in P until the vocabulary is exhausted. When no suitable candidates exist in the vocabulary, we move on to the next s .

To illustrate the algorithm with an example, consider S to consist of two input topics s : [“fun”, “sadness”]. For the first example (“fun”), we select ph_1 by performing G2P on the topic, returning /fʌn/, and select the word-initial /f/ as ph_1 . We then decide ph_2 by selecting the phoneme that has the lowest phonemic edit distance to /f/, returning /v/ as ph_2 . Following this, we feed the full prompt “*Generate a tongue twister on the topic of 'fun'.*” to the language model, and retrieve the next-token probabilities P . For example, P could be {1: “The”, 2: “It”, 3: “A”...}, where the set is the length of the decoded vocabulary. We then iterate through the predictions until a word meets the given criteria. For instance, the most likely continuation, “The”, is in the function word list F and within the *function_window* due to being at *rank* 1, so we append it to the prompt and now have “*Generate a tongue twister on the topic of 'Fun'. The*”, which we now denote s^* . We then feed this extended prompt into the language model to retrieve the second word, where P may be {1: “grey”, 2: “big”, 3: “fun”}. Here, options in *rank* 1 and 2 (“grey” and “big”) do not start with ph_1 or ph_2 and are also not function words in F . However, the word in rank 3, “fun” is transcribed as /fʌn/, where the initial phoneme /f/ matches ph_1 , so we enter the subword loop and find the candidate “funniest”, allow it to generate, and append it to the prompt, resulting in “*Generate a tongue twister on the topic of 'Fun'. The funniest*”. This is repeated until *max_length* is reached, where the process is started again for the remaining topic in S , “sadness”. We further ensure that we do not generate the same word more than once, as determined by *max_repetition*, and that we do not generate words shorter than *min_word_length*.

4.7.3 Models

To demonstrate the effectiveness of our decoding module, we utilise two decoder-only autoregressive language models as our *LM* to which PACD is applied: GPT-2 (Radford et al., 2019) and Baichuan (Yang et al., 2023). We investigate to what extent fine-tuning towards tongue twister generation is beneficial by additionally using fine-tuned GPT-2 and Baichuan models from §4.5, referred to as GPT-2_{13k} and Baichuan_{13k}. Importantly, we assess only these models fine-tuned on the largest amount of data (13k).

Regarding the other settings for PACD, we set *max_length* to 30 (as a sensible midpoint generation length ascertained from Table 4.7), *function_window* to 1, and implement *F* as the NLTK stopwords list with all punctuation removed. Additionally, we set *min_stem_length* to 2 and *min_word_length* to 3 (as all standard one or two-letter words such as “I”, “a”, “I’m”, “am”, “at”, “in”, “up”, and “on” are in *F*). Additionally, *max_repetition* is set to 1, effectively banning repetition (though allowing plural/singular forms and case variants). We use the *g2pE* package for our *G2P* model, as in the creation of TwisterLister (§4.4.4). Finally, we only decode the top 2500 predictions at each time step, rather than the entire vocabulary, in order to significantly speed up inference, as it is rare to select tokens ranked below this point. We load Baichuan using 8-bit quantisation to make inference possible. For GPT-2, PACD takes approximately 5-10 seconds to generate a 30-word tongue twister (with or without subword generation), whilst Baichuan takes 10-15 seconds when generating full words only, and 30-100+ seconds when allowing subwords on a consumer CPU (i5 9600k).

4.8 Constrained Decoding Results

Automatic evaluation is performed in the same manner as §4.6.1, on both reference-based (i.e., BLEU, ROUGE, and BERTScore) and unreferenced metrics (i.e., IPO/PO/iPED/oPED

and the readability metric suite). Human evaluation is likewise performed in the same manner as §4.6.2 and with the same evaluators, following the protocol for the topic-to-twister task setting. Each evaluator is presented with 20 examples (using the same inputs as in §4.6.2) from base GPT-2 and Baichuan with the addition of PACD, and fine-tuned GPT-2_{13k} and Baichuan_{13k} with and without PACD.

4.8.1 Automatic Evaluation

Model	B-1↑	B-2↑	B-3↑	B-4↑	Ro-1↑	Ro-2↑	Ro-L↑
GPT-2 _{13k} -w/o	0.0440	0.0152	<u>0.0061</u>	<u>0.0025</u>	5.9412	0.6024	5.5890
GPT-2 _{13k} -w	0.0832	<u>0.0170</u>	0.0046	0.0013	8.6673	<u>0.9067</u>	7.5011
GPT-2 _{13k} -ws	0.0584	0.0045	0.0003	0.0000	0.0089	0.0000	0.0792
GPT-2 -w	<u>0.0759</u>	0.0142	0.0032	0.0010	<u>8.0608</u>	0.6403	6.8765
GPT-2 -ws	0.0594	0.0059	0.0009	0.0001	0.0940	0.0121	0.0790
Baichuan _{13k} -w/o	<u>0.0463</u>	0.0227	0.0131	0.0080	<u>6.2215</u>	1.3878	<u>5.9212</u>
Baichuan _{13k} -w	0.0493	0.0041	0.0004	0.0000	0.0846	0.0081	0.0677
Baichuan _{13k} -ws	0.0525	0.0051	0.0006	0.0001	0.0940	0.0106	0.0742
Baichuan -w	0.0547	0.0054	0.0009	0.0002	0.0885	0.0109	0.0750
Baichuan -ws	0.0600	0.0080	0.0019	0.0005	0.1010	0.0175	0.0846

Table 4.14: Results of automatic evaluation on model outputs generated using PACD. Results in **bold** represent the best performance, and underlined results show the second-best performance. -w/o denotes without constraints, -w denotes the addition of PACD, and -ws denotes the PACD with subword generation enabled.

Automatic evaluation results for PACD are presented in Table 4.14 for referenced metrics and Table 4.15 for unreferenced metrics. Firstly, for GPT-2, regarding the reference-based metrics, fine-tuned GPT-2 with PACD (GPT-2_{13k} -w) outperforms the standard fine-tuned model (GPT-2_{13k} -w/o) on B-1 and B-2, in addition to all ROUGE-based metrics (Ro-1, Ro-2, and Ro-L). The exception is for the higher-order BLEU measures, B-3 and B-4, where the unconstrained model achieves higher overlap. Base GPT-2 also benefits from the addition of PACD (GPT-2 -w), outperforming the unconstrained fine-tuned model on recall-based ROUGE. Observing the unreferenced

CHAPTER 4. TRAIN & CONSTRAIN

Model	BS-P↑	BS-R↑	BS-F1↑	IPO↓	PO↓	iPED↓	oPED↓	Re-D	Re-F	Re-G	Re-A
GPT-2 _{13k} -w/o	0.7748	<u>0.8322</u>	0.8021	0.1261	0.0859	4.2671	5.8985	12.06	13.46	15.21	16.64
GPT-2 _{13k} -w	0.8056	0.8207	0.8129	0.2305	0.2317	3.0376	5.8930	13.88	14.58	17.58	18.77
GPT-2 _{13k} -ws	0.8018	0.8267	0.8139	0.1724	0.2287	1.7252	5.7686	11.60	12.24	15.59	17.20
GPT-2 -w	0.7986	0.8208	0.8094	0.1689	<u>0.2110</u>	1.6529	5.7699	11.36	12.66	15.76	15.91
GPT-2 -ws	<u>0.8119</u>	0.8258	<u>0.8186</u>	0.2469	0.2614	3.1642	5.8991	9.02	10.45	14.16	13.65
Baichuan _{13k} -w/o	0.9442	0.9411	0.9425	0.4908	0.4236	4.4789	5.9771	9.98	5.60	7.15	4.62
Baichuan _{13k} -w	0.7915	0.8202	0.8053	0.1629	0.2416	<u>1.4324</u>	<u>5.6876</u>	10.72	11.18	13.52	15.18
Baichuan _{13k} -ws	0.7887	0.8215	0.0742	<u>0.1512</u>	0.2194	1.2898	5.6708	11.95	12.59	14.82	17.21
Baichuan -w	0.8014	0.8204	0.8106	0.2102	0.2461	2.4449	5.8094	9.08	10.55	14.26	14.25
Baichuan -ws	0.7984	0.8217	0.8096	0.1992	0.2339	2.3284	5.8016	9.89	11.37	14.86	15.38
Brown Corpus Prose	-	-	-	0.4870	0.2275	5.1431	5.9561	11.09	13.15	15.56	16.60

Table 4.15: *Results of automatic evaluation on tongue-twister-related metrics for the constrained decoding approach to tongue twister generation in the topic-to-twister task setting. Results in **bold** represent the best performance, and underlined presents the second-best performance. -w/o denotes without constraints, -w denotes with the addition of our PACD module, and -ws denotes the addition of PACD with subword generation enabled. Brown Corpus Prose presents the scores of standard prose from the Brown Corpus from NLTK (across 602 sentences with a minimum length of 25 words).*

results in Table 4.15, performance ordering varies on the BERTScore semantic metrics (BS-P, BS-R, and BS-F1), with the unconstrained fine-tuned model (GPT-2_{13k} -w/o) trading places with the constrained equivalents (GPT-2_{13k} -w and -ws) for the most performant. However, base GPT-2 with PACD (GPT-2 -w) consistently achieves second place. Furthermore, we see the original fine-tuned, unconstrained model perform the best on the original naive phoneme-based metrics IPO and PO, whilst the non-fine-tuned, constrained model, GPT-2 -w, places second. It is here that the less naive newly presented phoneme-based metrics, iPED and oPED demonstrate their usefulness. For instance, whilst fine-tuned GPT-2 without any constraints (GPT-2_{13k} -w/o) performed best on the naive metrics (IPO/PO), it performed the worst on the more linguistically informed metrics (iPED/oPED). This is because good tongue twisters exploit the relationships between similar sounds, whilst IPO/PO penalise transitions between similar phonemes. In contrast, the informed metrics reflect favourably, penalising such transitions less than transitions between weakly related phonemes.

CHAPTER 4. TRAIN & CONSTRAIN

When looking at Baichuan, we see that the unconstrained fine-tuned model scores highest overall on B-2, B-3, B-4, and Ro-2, as well as all BERTScore measures. Interestingly, when analysing the impact of the addition of PACD, we see non-fine-tuned Baichuan with subword generation enabled (Baichuan -ws) to outperform all other versions of Baichuan that contain PACD, even without being fine-tuned on tongue twisters (however, scores are low overall). Regarding the phonemic metrics, we again see the benefit of PACD in reducing the scores on all phoneme-based metrics, successfully increasing sound overlap. Additionally, we see fine-tuned Baichuan with PACD (Baichuan_{13k} -w and -ws) outperform the non-fine-tuned models on the phoneme-based metrics. Finally, the addition of PACD to Baichuan led to a significant increase in readability scores (i.e., an increase in reading difficulty). However, these scores remain below the formal non-literary text of the Brown Corpus.

These results demonstrate the effectiveness of our constrained decoding approach that takes into consideration which word-initial phonemes would be the best to centre generation around to encourage mispronunciation. Overall, we see similar performance between our models using full-word and subword versions of PACD.

4.8.2 Human Evaluation

Score (1 to 5)	Constrained Topic-to-Twister					
	GPT-2 _{13k} -w/o	GPT-2 _{13k} -w	GPT-2 -w	Baichuan _{13k} -w/o	Baichuan _{13k} -w	Baichuan -w
Relevance	2.16***	<u>2.44</u> ***	2.42***	3.04 **	2.39**	2.41**
Articulation	2.80***	3.85 ***	<u>3.69</u> ***	3.21**	3.58***	3.28**
Fluency	<u>2.66</u> ***	2.39**	3.13 ***	2.24**	1.84**	1.88**
Coherence	<u>2.17</u> **	2.00**	3.41 ***	2.09**	1.68**	1.64**
Entertainment	2.00***	<u>2.14</u> **	2.97 **	1.81**	1.67**	1.64**
Overall	1.72***	2.00**	2.91 ***	<u>2.09</u> **	1.98**	1.89**

Table 4.16: Results of human evaluation in the topic-to-twister task setting using constrained decoding. -w denotes to models using our decoding algorithm, and -w/o denotes fine-tuned models without this additional module. The best scores are in **bold**, and the second-best are underlined. We calculate Fleiss’ Kappa for each metric, and we mark the extent of agreement with the following markings: fair* (0.21-0.40), moderate** (0.41-0.60), and substantial*** (0.61-0.80).

Human evaluation was performed in the same manner and with the same evaluators as §4.6.2. Due to enforcing a 30-word output length, we ask that evaluators not penalise generations on the criteria of fluency for being cut off prematurely. Additionally, due to the similarity of outputs generated with and without sub-word generation, we only perform human evaluation on outputs with subword generation disabled, to minimise evaluator fatigue and potential acquiescence bias. The results of the human evaluation of PACD are presented in Table 4.16. We additionally perform human evaluation on GPT-2 and Baichuan trained on 13k samples (the former of which was excluded from §4.6.2) in order to have a point of comparison without the addition of PACD (referred to as GPT-2_{13k} -w/o and Baichuan_{13k} -w/o, respectively).

A clear benefit from the addition of PACD can be seen for GPT-2. Firstly, across all evaluation criteria, either base (i.e., vanilla) or fine-tuned GPT-2 receives the highest scores when PACD is applied. However, fine-tuned GPT-2 without the addition of PACD (GPT-2_{13k} -w/o) outperforms the equivalent model on the criteria of fluency and coherence with PACD enabled. Fine-tuned GPT-2 has previously been shown to perform poorly in the topic-to-twister setting when considering coherence and grammaticality

(see also §4.9). Consequently, the addition of even more restrictive decoding rules introduced by PACD is further detrimental to the general output quality. On the other hand, non-fine-tuned language models are primarily designed to output standard prose that is grammatical and coherent. As a result, the addition of PACD does not damage the overall readability of the outputs too severely. Interestingly, however, base GPT-2 with PACD (GPT-2 -w) outperforms the fine-tuned models either with or without PACD, suggesting that the most desirable approach may be to train *or* constrain, rather than train *and* constrain, as this allows phoneme-level control without sacrificing grammaticality.

When considering Baichuan, human ratings decrease when moving from stand-alone fine-tuned Baichuan (Baichuan_{13k} -w/o) to the versions containing PACD. This, however, is not the case for the difficulty of articulation criterion, which increases in both PACD versions of Baichuan (i.e., Baichuan_{13k} -w and -ws). This suggests a possible trade-off regarding the addition of constrained decoding to a language model, with larger models suffering from the additional control being exerted over their outputs, whilst smaller models are better able to be guided towards desirable characteristics by complementing their reduced knowledge capacity during fine-tuning and zero-shot inference.

A significant finding from the human evaluation results is that PACD results in a substantial increase in the difficulty of articulation criteria. This indicates the success of the constraint cascade in selecting phoneme pairs that contribute to maximal levels of phoneme-level confusion and therefore mispronunciation.

CHAPTER 4. TRAIN & CONSTRAIN

Input:	"rural brewery"
Gold:	"Rory the warrior and Roger the worrier were reared wrongly in a rural brewery."
GPT-2_{2k}	"In the bustling city of Chattanooga, where Chattanooga's vibrant craft beer scene thrives, The Chattanooga's craft beer scene thrives, [With Chattanooga's craft beer scene thriving, The Chattanooga's craft beer scene thrives,]*3 With Chattanooga's craft bee "
GPT-2_{13k}	"In the bustling town of Chattanooga, where the townspeople reside, The Chattanooga's taproom is filled with tapestries of flavors. Chattanooga's taproom is filled with the finest tapestry, From the tapestry of Chattanooga's charm, it's a sight to see. Chattanooga's taproom is filled with the finest tapestry, From the tapestry of Chattanooga's charm, it's a sight to see. Chattanooga's taproom is filled with the finest tapes"
DialoGPT_{2k}	"In the bustling city of Bethlehem, a bustling brewery was bustling with craftsmanship and craftsmanship was its main attraction. It was also the place to find the best beer in the city. It was also the place to find the best breweries in the city. It was also the place to find the best breweries in the country. In the bustling city, a bustling brewery was bustling with craftsmanship and craftsmanship. It was also the place to find the best beer in the city. It was als"
DialoGPT_{13k}	"In the bustling city of Tharrington, where the city's charm is known, There's a thriving tavern called Tharrington's, Where the taverns thrive with a vibrant atmosphere. The tavern's owner, a skilled craftsman, built a taproom that was truly grand, With a taproom full of flavors, it was truly a sight. The tavern's atmosphere was so grand, it was like a vibrant symphony, With taverns and taverns, it was truly a sigh"
BART_{2k}	"In the bustling city of New Zealand, a New Zealander named New Zealanders gathered at the New Zealand's New Zealand brewery, where New Zealandians gathered to discuss the New Zealander's New Zealanders' New Zealand heritage." \ " \ " \ " " " " suscept " suscept ' ' char non"
BART_{13k}	"In the village of Vineland, where Vineland's vineyard thrived, Vineland and Vineland thrived. Vineland flourished, as Vineland joined the scene, As Vineland added to the scene. Vineland's vineyards, a vibrant scene, Where Vineland was the scene." " ' " ' " thwart " thwart thwarted"
Flan-T_{52k}	"In the township of Westfield, a brewer named Westman brewed a brew in the townhouse, whilst Westfield's breweries brewed brews in the townhouses of Westfield."
Flan-T_{13k}	"In the quaint town of Greenland, where gin and lager are the mainstays, Craftsmen from Greenland and Greenland gather to create a gin and lager masterpiece. The gin and gin are renowned for their exceptional craftsmanship, While the gin and gin are renowned for their exceptional taste. Greenland's gin and gin are renowned for their exceptional"
ByT_{52k}	"In the village's villages, the village's villagers visit the village's villages, whilst Victorian villagers visit the village's villages. The village's villagers visit the village's villages, whilst Victorian villagers visit the village's villages. The village's villagers visit the village's villages, and the farmhouses of Victorian villages and villages in Victorian villages. The village's villagers visit the village's villages, whilst Victorian villagers visit the village's villages. The village's villagers visit the village's villages, all from Victorian villages, and Victorian villages are"
ByT_{13k}	"In the vicinity of the village, the farmers from the farmhouse filled the villages with their farmhouses and farmsteads, whilst the farmers from the farmhouse farms filled the villages with their farmhouses."
Baichuan_{13k}	"[Breweries brew beer,]*18
ChatGPT	"In a rural brewery, they brewed a brew so unique, Where hops happily hopped and barley barreled sleek. With malty melodies and yeasty jigs, they thrived, But rural brewery's ruckus rhythm rarely arrived. The rural brewery's brewmaster was a friendly fellow, Fermenting fizzy flavors, flowing with yellow. With each sip, the patrons picked passionately, Pouring pints, pretty pleased, with plenty of glee. But with the twist of tongue and tricky tales, The rural brewery's name became like snails. Try as they might, lips tangled and twisted, Rural brewery remained quite disjointed and resisted. The rural brewery's brand built on rustic charm, But their tongue twister seemed to cause alarm. With each attempt, patrons tangled like vines, Rural brewery's name spinning in their minds. So, remember to take your time and don't hurry, When uttering words about a rural brewery. Though the tongue may twist and the speech may waver, The taste of their brew will make you a believer!"

Table 4.17: Example outputs from models where different volumes of fine-tuning data were used (2k or 13k), in addition to Baichuan with 13k training examples and zero-shot ChatGPT in the topic-to-twister task setting. Text within square brackets indicates repeated spans, with the number of repetitions indicated next to it.

4.9 Qualitative Analysis

4.9.1 Topic-to-Twister (Fine-Tuned)

Table 4.17 presents example generations using 2k and 13k training samples to qualitatively investigate the effects of variable volumes of fine-tuning data. We additionally include Baichuan trained on 13k samples, and ChatGPT. For these examples, we randomly select an input topic from the first 100 samples in the test set.

GPT-2. Firstly, when considering generations from GPT-2, the 2k training sample model does not demonstrate any clear phoneme-level patterns, displaying no prominent sound repetition. However, some orthographic repetition is observed, with “Chattanooga”, “craft”, and “scene” presenting three ways in which the graphemes <c> and <s> appear together (as <c> is frequently realised as /s/, like <s>, in words such as “celery”). On the other hand, the 13k training sample generation presents a much better tongue twister, demonstrating repetition of /t/ in “*town of Chattanooga[...] Chattanooga’s taproom is filled with tapestries*”. This is additionally complemented by the repetition of the affricate /tʃ/ in the phrase “*Chattanooga’s charm*”. Whilst these tongue twisters appear successful, in both instances the models have resorted to the repetition of very similar words. However, all in all, GPT-2_{13k} has successfully generated a tongue twister that is largely grammatically coherent, demonstrates phoneme overlap, and is semantically related to the input (even if the rural nature of Chattanooga, a city in Tennessee, may be debatable). Consequently, this first instance lends support to the proposed benefit of extended volumes of fine-tuning data for the task of tongue twister generation, as provided by the extension of TwistList 1.0 into TwistList 2.0, due to GPT-2_{13k} demonstrating better performance than GPT-2_{2k}. This finding also supports

our claims that the proposed TwisterLister pipeline creates high-quality tongue twisters from which to learn.

DialoGPT. Regarding DialoGPT in the topic-to-twister setting, we see high levels of repetition in the 2k training sample output, such as “*bustling with craftsmanship and craftsmanship*”. However, some elements are clearly tongue twister-esque, such as the initial “*In the bustling city of Bethlehem, a bustling brewery was bustling*”, exploiting the voiced bilabial plosive, /b/. In the 13k training sample output, “Bethlehem” is swapped for “Tharrington”, and the exploitation of the voiceless dental fricative /θ/ also found in “thrive” and “thriving”. Overall, DialoGPT demonstrates a more meaningful narrative-like generation alongside a higher volume of fine-tuning data, in addition to a change in primary phonemes (potentially arising from the distributional properties of different phonemes across the larger training split).

BART. BART, on the other hand, demonstrates less desirable traits in the 2k setting, using extreme levels of repetition for “New Zealand” (and morphological variants), resulting in a poor quality output that only clearly represents the input topic through the inclusion of the word “breweries”. Additionally, BART resorts to generating nonsense output towards the end, producing myriad punctuation and random words. In the case of 13k fine-tuning samples, the output is less overtly repetitive and is more coherent, though still far from perfect, exploiting /v/ and /θ/ (a voiced labiodental fricative and a voiceless dental fricative). For example, “*In the village of Vineland, where Vineland’s vineyard thrived, Vineland and Vineland thrived*.” However, the generation does exhibit significant breakdown towards the end, once again producing nonsensical output.

Flan-T5. Flan-T5 generates the shortest output seen across the models for this input. The generated tongue twister in the example only appears to engage with the repetition

CHAPTER 4. TRAIN & CONSTRAIN

of /b/ (rather than alongside a similar phoneme), but does so successfully, such as “[...] *a brewer named Westman brewed a brew in the townhouse, whilst Westfield’s breweries brewed brews[...]”]. However, the semantic coherence of this output is lacking, with the discussion of two parallel events. Additionally, the repetition of /b/ is achieved exclusively through the exploitation of morphological variants of “brew”, suggesting the signal that a morphological signal was learnt during training, rather than a graphemic or phonemic signal. With an increased volume of training data, the generation length changes significantly for Flan-T5. In the generation from Flan-T5_{13k}, there is clear repetition of the voiced velar plosive /g/ and the voiceless counterpart /k/, such as “*Craftsmen from Greenland [...] gather to create a gin and lager”. However, unlike in the 2k example, the output exhibits significant semantic redundancy as seen in phrases such as “*The gin and gin [...] whilst the gin and gin*”.**

ByT5. ByT5, the largest of the models we fine-tune on varying volumes of data (at 582M parameters), performs poorly when fine-tuned on 2k samples, with the output degrading into repetition. However, the repetition consists of full phrases and sentences rather than a single noun and maintains grammaticality throughout (albeit lacking in coherence). As for the version fine-tuned on 13k samples, this remedies the repetition issue. Interestingly, this version more clearly demonstrates a phonemic pattern in the outputs, switching between /v/ in “*villager*” (and words with the same root) and /f/ in “*farmers*” (and words with the same root), which are voiceless and voiced counterparts, and therefore an ideal pattern for a tongue twister to exploit. This is also demonstrated in the 2k version, but only weakly, with one instance of an /f/-initial word, “*farmhouses*”.

Baichuan. Baichuan, as the largest model we fine-tune (exclusively on the 13k training data split) with 7B parameters, performs rather poorly in the topic-to-twister setting. The example consists exclusively of a single phrase “Breweries brew beer” repeated 18

CHAPTER 4. TRAIN & CONSTRAIN

times. Whilst this is a valid tongue twister in terms of relevance to the input, Baichuan is shown to very quickly loop, degrading the quality of the overall output. This pattern is seen frequently across other outputs from Baichuan.

ChatGPT. Finally, ChatGPT presents the longest tongue twister of the examples presented. However, this is an exception for this randomly selected generation, rather than an overall tendency, as seen in the automatic evaluation results, where the average generation length for ChatGPT was 17.43 words. In terms of quality, ChatGPT excels at generating well-formed grammatical text. Additionally, the generation exploits numerous different phonemic patterns, including repetition of /h/ “... hops happily hopped”, /b/ “In a rural brewery they brewed a brew[...]”, and /ɹ/ “But rural brewery’s ruckus rhythm rarely arrived”. Additionally, one demonstrated ability of ChatGPT that is not exploited by any of the other fine-tuned models is the incorporation of additional sound-based literary techniques such as rhyme (e.g., “thrived”/“arrived” and “tails”/“snails”). In addition, the coherence level of the ChatGPT output is very high, with the output demonstrating a clear narrative that would engage readers. Consequently, we can see why ChatGPT may score poorly on phoneme-based metrics (cf., Table 4.8) due to incorporating many words to enhance grammar, as well as exhibiting low-level local repetition of a particular sound in a phrase, rather than maintaining focus on the same sound for the entire output.

Overall, most models appear to benefit from the presence of increased levels of training data in this example, particularly regarding producing more grammatical and coherent output. The only model for which this observation does not hold is Flan-T5, which does not have a clear change in generation quality from 2k to 13k training samples. This is in some respects to be expected, as larger models are often able to abstract away

language patterns more easily than smaller models from the same amount of training data (therefore resulting in a law of diminishing returns between training data quantity and model size). However, the smaller models demonstrate the utility of our TwistList 2.0 dataset in presenting a sufficiently large increase in training data to result in tongue twister generations of demonstrably improved quality. This also shows that there is room to improve generated tongue twisters with an increase in training data exclusively, before the requirement for more complex training paradigms is necessary.

4.9.2 Topic-to-Twister (Constrained Decoding)

In contrast with the previous approach, the constrained decoding algorithm PACD involves the random selection of an initial phoneme and a nearest neighbour, through which to condition generation. Consequently, numerous tongue twisters can be generated for a single input, as long as the selected phoneme pairs differ each time. For this reason, we present an additional case study of model outputs using PACD. Please note, we additionally reproduce the examples from fine-tuned GPT-2 and Baichuan without the addition of PACD.

GPT-2_{13k} -w/-ws. Fine-tuned GPT-2 with the additional constraints imposed by PACD demonstrates several characteristics. Firstly, as is the case with the non-fine-tuned model, our constrained generations are limited to 30 words, making all generations of equal length, and consequently shorter than the output of GPT-2 without these constraints. Regarding the content, each example is related to the input keywords, but varies between direct and abstract references. For example, the first generation conditioned on /b/ and /p/ demonstrates a direct reference to the input word “brewery”, something that has been afforded to the model as the selected initial phonemes match one of the word-initial phonemes of the input. The generations exploiting /ɪ/ and

CHAPTER 4. TRAIN & CONSTRAIN

Input:	“rural brewery”
Gold:	“Rory the warrior and Roger the worrier were reared wrongly in a rural brewery.”
GPT-2_{13k}	“In the bustling town of Chattanooga, where the townspeople reside, The Chattanooga’s taproom is filled with tapestries of flavors. Chattanooga’s taproom is filled with the finest tapestry, From the tapestry of Chattanooga’s charm, it’s a sight to see. Chattanooga’s taproom is filled with the finest tapestry, From the tapestry of Chattanooga’s charm, it’s a sight to see. Chattanooga’s taproom is filled with the finest tapes”
GPT-2_{13k} -w	“Brewers brew beer in the barn brewing process before brewers bring their beers to the bar boys breweries prior to being brewed beforehand but before being bottled previously because of”
GPT-2_{13k} -ws	“With their rust wood working wonders with their water works whilst watering wheat wine wines within their win ware warehouses where they are waiting for the weather to warm up”
GPT-2 -w	“Brewers brew beer in the breweries belonging to both baker boys and bar patrons but the bartender prefers the beers brewed by bud brothers and bro pal pit bulls before”
GPT-2 -ws	“With a rust wood wind ward rip weed from the woods whilst watering the water works well within the walls of the warehouse where the work room is wall rolling”
GPT-2 -w	“But before you do begin building your own brewery please bear in place that you are not a brewer but a person who has been brewing for a bit because “
GPT-2 -ws	“What is the reason for this one with the word rural rather than regional where it is written when it is referring to the region of the world where the”
GPT-2 -ws	“But I believe that the best place to begin is with the basics before you proceed beyond that point because the more you practice brewing beer properly both in the”
GPT-2 -ws	“What is the reason for this request with the word rural when you are referring to a region that is not a regionally relevant one where there is a real”
Baichuan_{13k}	“[Breweries brew beer]”*18
Baichuan_{13k} -w	“Try to drink tea in the town during dinner time drinking delicious drinks from the tap tasting different types of tasty treats trying to taste the differ den between two”
Baichuan_{13k} -ws	“What is the relations with rural roads running through resident rial ways within regional regions where rivers run through rocky rid rig road way runs through region rich river routes”
Baichuan_{13k} -ws	“Try to drink tea in the township of taverna towns during the daytime drinking tour dates downtown drinker drinks dry den dutch town tavern a day time tasting tours tourist”
Baichuan_{13k} -ws	“What is the relationship between the rural region and the regional railway route running through rut rum runners runs through rocky regions with rich resources whilst retaining its rustic roots”
Baichuan -w	“Try to twist the tongue as tight together as time does during the day today due to the temper ter differ ten times too different degrees depending on the direction”
Baichuan -ws	“Write a one word response to each of the words below with a relevant rural reference within the reply which is not a repeated ref rain water recycling wind renewable”
Baichuan -ws	“Try to twist the tongue as tightly as desired during the time of delivery trying to deliver the twistiest tong toe track ter day till the target is delivered tight”
Baichuan -ws	“Write a one word response to each of the words below with a relevant rural reference within the reply which is not a repeated ref rain water recycling rainwater runoff”

Table 4.18: *Example outputs using PACD. -w denotes models using PACD for full words only, whilst -ws indicates the subword version of PACD. We present two generations for each model and PACD variant. The example exploit the phonemes /b, p/ and /ɹ, w/, respectively. To achieve this, the initial phoneme selection process of ph_1 in PACD is overwritten, in order to present diverse outputs from the same input (as /b, p/ would be otherwise be selected exclusively).*

CHAPTER 4. TRAIN & CONSTRAIN

/w/ are similarly abstract, referring to relevant words such as “water”, “wine”, and “wood working” (the latter activity being more expected in a rural locale). Regarding grammaticality, the /b, p/ example is largely correct (if “boys” had a possessive apostrophe, creating “*the bar boys’ breweries*” as a noun phrase) but is cut off at 30 tokens, resulting in an incomplete sentence. However, the semantics are difficult to comprehend due to the frequency of temporal adverbs. Likewise, the /ɪ, w/ generation is mostly grammatical, with the presence of numerous relative clauses leading to difficulty in understanding (however, this does require liberty in accepting “*wheat wine wines*” and “*win ware warehouses*” as compound nouns). With the subword loop enabled for GPT-2 using PACD, we see similar outputs in style to the full-word version. However, one difference is a slight reduction in grammaticality towards the end of the /b,p/ example, with the phrase “*and bro pal put bulls before*”, which is difficult to parse. Similar effects can be seen with the /ɪ, w/ version, where the outputs are similar in word choice, and the overall result is perhaps slightly more difficult to parse.

GPT-2 -w/-ws. Regarding non-fine-tuned base GPT-2 with the addition of PACD, all generations demonstrate similar levels of grammaticality to those seen from the fine-tuned model. On the other hand, the generations are less literary, with the /b, p/ example suggesting the model is generating instructional text in “*But before you do begin building your own brewery[...]*”. Similarly to previously discussed generations with PACD, all of these generations are hindered by the 30-word limit, with examples ending in “because”, “*his desire to try*” (which suggests the verb “try” will take an additional argument) and “the”. Overall, similar outputs are seen with the subword loop enabled.

Baichuan_{13k} -w/-ws. With the addition of full-word PACD to fine-tuned Baichuan, repetition is avoided due to constraints enforced at decoding time. For the first example, the enforcement of /t, d/ is evident in grammatical phrases such as “*Try to drink the*

tea in the town during dinner time...”, and is continued throughout. Overall, the first generation is grammatical, only demonstrating a clear degradation towards the end with the generation of “*differ den*” which hinders fluency and coherence. On the other hand, in the /ɪ, w/ example, the sound overlap is demonstrated, but the coherence and fluency of the output are shown to suffer much earlier, with sequences such as “*rivers run through rocky rid rig road ways runs through region rich river routes*”. Overall, generation quality starts to drop after a single suboptimal token is selected due to the detrimental impact it has on the following token’s prediction. When comparing to the subword-enabled PACD (-ws) we see the initial difference in the 7th word of the /t, d/ example, where “town” has been extended to “township”, resulting in the remainder of the generation diverging from that of the full-word version. This is further reflected in the /ɪ, w/ example as “relation” becomes “relationship”. Overall, both present sequences that are difficult to articulate, with minimal quality difference between subword and full-word versions.

Baichuan -w/-ws. Regarding base Baichuan with the addition of PACD, there is a lack of topical relevance in the /t, d/ examples for both full-word and subword versions, with the input of “rural brewery” not being reflected in the semantics of the output (although the generation has noted the request for a tongue twister in the starting prompt, and references it directly). This is also seen in the /ɪ,w/ versions but to a lesser extent, as the topic word “rural” starts with an allowable phoneme, /ɪ/, causing it to generate in the phrase “*rural reference*”. In this example, therefore, fine-tuned Baichuan is better able to understand the desired topic, whilst the base model produces overtly generic text, but with repetitive phonology.

Overall, the examples demonstrate the effectiveness of the simple constrained decoding module that constitutes PACD. In the generated examples, the ability to control token repetition means the pitfalls of standard fine-tuned models are avoided, such as frequent reliance on phrasal repetition, ensuring that tongue twisters generated with PACD use a wide range of vocabulary. Additionally, due to the phoneme-based constraints, control can be exercised over the included sounds. To this end, similar to TwisterLister (§4.4.1), PACD facilitates phoneme overlap within the output, rather than relying on this signal being learnt during fine-tuning. The case studies additionally demonstrate that rule-based decoding of a pre-trained language model can result in tongue twisters that are comparable to those of small fine-tuned models, and that future work in the area should make extensive use of constrained decoding approaches such as PACD, rather than exclusively fine-tuning on input-output pairs.

4.9.3 Style-Transfer (Fine-Tuned)

We now qualitatively investigate generations from the style-transfer setting. Table 4.19 presents example generations from models fine-tuned on 2k and 13k samples. We further include Baichuan trained on 13k samples, and zero-shot ChatGPT.

As demonstrated in the *Input* and *Gold Standard* fields, the generated paraphrase is not easily identifiable as a tongue twister. Therefore, this is a prime example of a successful paraphrase to enable style-transfer, as the twisted style has been removed in order to have parallel data to learn the mapping from prose to tongue twister. The gold standard tongue twister exploits the common articulation difficulty and confusion between the rhotic /ɹ/ and lateral /l/ (Georgievska-Jancheska, 2019), encouraging the accidental transposition of speech sounds across words (i.e., metathesis).

CHAPTER 4. TRAIN & CONSTRAIN

Input:	“The small crimson truck traveled along Limuru street.”
Gold:	“The little red lorry went down Limuru road.”
GPT-2_{2k}	“The small carriages of the small carriages, with their small wheels, were moving along Limuru’s road.”The small yellow truck traveled along Limuru’s road. The yellow cars of the yellow cars, with their yellow wheels, were traveling along Limuru’s road.”The yellow yellow truck traveled along Limuru’s road.” The yellow cars of the yellow cars, with their yellow wheels, were traveling along Limuru’s road.”The yellow yellow truck traveled along Limuru’s road”
GPT-2_{13k}	“The small red truck, with its small red tires, was a sight to see. The small red truck, with its small red tires, was a sight to see. The small red truck, with its small red tires, was a sight to see. The small red truck, with its small red tires, was a sight to see. The small red truck, with its small red tires, was a sight to see. The small red truck, with its small red tires, was a sight to see. The small red truck, with its small red tires, was a sight to see. The small red truck, with its small red tires, was a sight to see.”
DialoGPT_{2k}	“It passed through the town of Limburg, where the small town of Limburg’s inhabitants were located. It passed through the town of Limburg, where the town of Limburg’s inhabitants were located. It passed through the town of Limburg, where the town of Limburg’s inhabitants were located. It passed through the town of Limburg, where the town of Limburg’s inhabitants were located. It passed through the town of Limburg, where the town of Limburg’s inhabitants were located.”
DialoGPT_{13k}	“It passed by the small houses and the small houses’ small houses, whilst the small houses’ small houses and the small houses’ small houses were observed.”The minuscule minivan passed by the mansions and mansions’ mansions, whilst the mansions’ mansions and the mansions’ mansions were seen.” Improved version: The minuscule minivan passed by the mansions and mansions’ mansions, whilst the mansions’ mansions and the mansions”
BART_{2k}	“The limo journeyed along Limuru thoroughfare.” “ churn “ churned B hollow f un ” umbled inc ops op edy - standard r unilaterally worsened thumb pl thrust res ed churn”
BART_{13k}	“The little red limousine traveled down Limuru thoroughfare.” “The little red limousine traveled down Limuru thoroughfare.” “The little red limousine traveled down Limuru thoroughfare.” “The little red limousine traveled down Limuru thoroughfare.” “The little red limousine traveled down Limuru thoroughfare.” “The little red limousine traveled down Limuru thoroughfare.” “The little red limousine traveled down Limuru thoroughfare.” “The little red limousine traveled down Limuru thoroughfare.”
Flan-T5_{2k}	“The yellow limo limped along Limuru limoway.”
Flan-T5_{13k}	“The little lilac lorry limped down Limuru lanes.”
ByT5_{2k}	“The little limuru truck looked along Limuru street.”
ByT5_{13k}	“The little red truck wandered along Limuru thoroughfare.”
Baichuan_{13k}	“Limuru’s little red truck drove along Limuru road.”
ChatGPT	“Truck crimson, small, along Limuru street did travel.”

Table 4.19: Example outputs from all fine-tuned models where different training data amounts were used (2k v. 13k) in addition to Baichuan with 13k training examples, and zero-shot ChatGPT in the style-transfer setting. The majority of ChatGPT outputs resemble tongue twisters, whilst the randomly selected input topic generated twisted syntax in this particular case.

GPT-2. Firstly, the main notable characteristic of generations from GPT-2 is the excessive length. This is particularly interesting when considering the nature of style-transfer, where the ideal target would be of similar length to the input. However, upon further inspection, it is clear that GPT-2 resorts to redundant repetition to achieve the difference in length. Consequently, if only the first sentence of the 2k version is taken as the output “*The small carriages of small carriages, with their small wheels, were moving along Limuru’s road.*”, the output can more clearly be seen as a style-transferred version of the input. However, this generation is still poor, with little phoneme-level repetition (excluding repetition of the same word, such as “small” in the quoted passage) and much semantic redundancy. Regarding the 13k training example output, GPT-2 again exhibits large levels of repetition, but in this case being a verbatim repetition of the initial sentence numerous times. In addition, the output is free of any grammatical errors. In this example, there is clear use of the sibilant /s/ in “*The small red truck with its small red tires, was a sight to see*”. Additionally, the input semantics of a small truck have been maintained, but the setting of Limuru, a town in Kenya, has been lost.

DialoGPT. On the other hand, the output from the 2k version of DialoGPT lacks semantic coherence due to using deixis in the form of “it”, which is not resolved as a cataphoric reference later in the text, causing ambiguity. A potential attempt at phoneme repetition is present with “Limburg” and “located”, but this is tenuous. Overall, the generation with 2k fine-tuning examples does not clearly present a tongue twister. On the other hand, the extended 13k example demonstrably resembles a tongue twister, achieving repetition of /s/ in “*Small houses and the small houses’ small houses*”, and /m/ in “*The minuscule minivan passed by the mansion and the mansions’ mansions*”. However, this generation also exhibits a strange structure, including “improved version” as part of the output. If only a subset of this generation is considered, “*The minuscule*

minivan passed by the mansion and the mansions’ mansions”, the output constitutes a high-quality tongue twister. Whilst it may be semantically bizarre to discuss the ownership of mansions by other mansions, it is not grammatically invalid, and human-authored tongue twisters also frequently convey unusual semantics.

BART. Unlike the GPT-2 models, BART produces style-transferred versions of the input that are much closer in length to the original input, without needing to exclude sentence-level repetition. In the 2k example, the initial sentence “*The limo journeyed along Limuru thoroughfare*” is a grammatically and semantically valid output, but does not resemble a tongue twister outside of the lateral /l/ in both “limo” and “Limuru”. However, the sentence following this consists primarily of noise. On the other hand, alongside the increase in fine-tuning data comes an improvement in tongue twister quality, with “*The little red limousine traveled down Limuru thoroughfare*” appearing to exploit the articulatory similarity between /l/ and /ɹ/ without relying on single-word repetition. Again, however, the output devolves into noise towards the end.

Flan-T5. Flan-T5, alongside ByT5, presents the most clear paraphrases of the desired output, containing no redundant repetition or noise. In the 2k example, Flan-T5 repeats similar phonemes in “*The yellow limo limped along Limuru limoway.*”, exploiting the similar categories of laterals, /l/, and glides, /j/. On the other hand, from more fine-tuning examples, Flan-T5 presents a paraphrase that is more faithful to the input, relying on /l/ and /ɹ/: “*The little lilac lorry limped down Limuru lanes.*” (where “lorry” is more similar to “truck” than “limousine”).

ByT5. ByT5 shows equivalent performance to Flan-T5 in this example, opting to alliterate different parts of speech to Flan-T5, but having an overall similar effect in the 2k example. On the other hand, the 13k version does not present alliteration as clearly

as Flan-T5, but does alternate between the related phonemes of /l, ɹ, w/ in “*little red truck wandered*”.

Baichuan. Baichuan performs much better in the style-transfer setting than the previous topic-to-twister setting. Similar to other generations, Baichuan exploits the similarity between /ɹ/ and /l/ with “*Limuru’s little red truck drove along Limuru road*”, where every word of the output contains one of these sounds, and alternation is frequent.

ChatGPT. Finally, ChatGPT in a zero-shot setting appears to misinterpret the prompt, outputting non-standard syntax such as swapping the order of an adjective and noun in “truck crimson”, reminiscent of the speaking style of Star Wars’ Yoda. Additionally, the ChatGPT generation does not appear to reflect a tongue twister in any clear sense. This is not a common pattern with ChatGPT outputs, but it is the case for the randomly selected case study example.

Overall, as with the topic-to-twister setting, a clear benefit from additional fine-tuning data is observed in the style-transfer setting, with all models producing better quality outputs both in terms of phoneme patterns, grammatical validity, and semantic coherence. An overarching theme, however, is that several models misinterpret the desire to paraphrase the input as a tongue twister, instead producing outputs that are much longer. However, if edited in post, often simply by taking the first sentence, these generations are often still valid, if not perfect.

4.10 Conclusion

In this work, we have presented multiple novel contributions towards the development of more phonologically-aware models for tongue twister generation. We presented a pipeline (TwisterLister) to generate tongue twisters at scale using large language models, encouraging non-derivative examples through the careful selection of candidate vocabulary to develop a large dataset of machine-generated tongue twisters (TwistList 2.0). We then fine-tuned a series of smaller language models on TwistList 2.0 and observed different characteristics between the topic-to-twister and style-transfer task settings when considering variable volumes of fine-tuning data. These findings demonstrated a fundamental difference in the requirements of the two approaches and further motivated the need for additional fine-tuning data as supplied by TwistList 2.0. We additionally presented a novel algorithm (PACD) that implements hard lexical constraints based on the phonemic characteristics of words that can be used to realise tongue twisters from an autoregressive language model by accessing the next token predictions and applying a cascade of filters. We then extensively evaluated the generations from our proposed approaches, presenting both automatic and human evaluations. In the former, we additionally presented two novel metrics for measuring the sound complexity of a generated tongue twister based on the concept of phoneme edit distance (iPED and oPED). With these fine-tuned models and constrained decoding module, we then provided an in-depth exploration of the generation characteristics through case studies investigating the propensity of each model to generate high-quality tongue twisters, as well as the differences between models trained on variable volumes of fine-tuning data. Overall, we find that fine-tuning language models for tongue twister generation still has substantial room for improvement to meet human-authored standards. However, we additionally demonstrate that simple constrained decoding approaches are able to

generate better tongue twisters than fine-tuned models alone, due to phoneme-level awareness.

4.11 Limitations

As this work aimed to investigate the ability of pre-trained language models to learn phonetic and phonological patterns from orthographic text alone, no models were trained on phonetic representations such as the IPA. Furthermore, the PACD constrained decoding algorithm was implemented to demonstrate how phonological constraints could be imposed on model outputs without the need for fine-tuning data. However, constrained decoding remains a computationally expensive approach to language generation, and the presented implementation was not optimised for efficiency. Finally, all tongue twisters were phonetically transcribed into the General American (GenAm) accent. As a result, the implemented phoneme edit distance metrics may not fully align with the experience of speakers of other varieties of English with different phoneme inventories and distributions. Furthermore, the presented finetuning set up remains similar across all trained models. Whilst this allowed for consistency in training, it also means that individual models were not fine-tuned explicitly to best match each model’s intended paradigm (e.g., text completion, dialogue). However, inputs were formatted to match general expectations (e.g., task prefixes for T5). Statistical significance testing was omitted from cross-model evaluations due to the main contribution of this work being the identification of a general capability gap between language models and humans in the task of tongue twister generation. Whilst inter-model performance differences are discussed, the purpose of this work is primarily exploratory and to provide a novel dataset rather than to suggest a specific model for the task of tongue twister generation or present a state-of-the-art architectural solution.

4.12 Appendix

4.12.1 Evaluation Rubric

Below is the evaluation rubric presented to human evaluators. The prompts have the following format: “[criterion description] \n Instruction: Generate a tongue twister relating to [input] \n Response: [output] \n Rate the response from 1 to 5:\n [rubric]” Where [criterion description] refers to the first line of the examples below (e.g., “Is the model proficient in...”), [input] is either a topic in the topic-to-twister setting, or standard non-literary text in the style-transfer setting, [output] is a model’s response to a given input, and [rubric] is the 5-point rating system as outlined below.

Relevance

Is the model proficient in generating text that is relevant to the input topic?

1. The model completely ignores the input topic and generates irrelevant text.
2. The model generates text that is mostly irrelevant to the input topic, with minimal and unclear association.
3. The model generates text that is partially relevant to the input topic, but the association is weak and inconsistent.
4. The model generates text that is mostly relevant to the input topic, with clear association but occasional lapses.
5. The model excels in generating relevant text, where the responses are consistently on topic and the association is clear.

Difficulty of Articulation

Is the model proficient in generating text that is difficult to pronounce due to alliteration

CHAPTER 4. TRAIN & CONSTRAIN

and phonetic complexity?

1. The model generates text that is no harder to pronounce and articulate than standard writing.
2. The model generates text that is slightly more challenging to say than standard writing, demonstrating some simple techniques such as alliteration.
3. The model generates text that is somewhat difficult to say but clearly exhibits techniques such as alliteration.
4. The model generates text that is generally difficult to say, with techniques such as alliteration, but also alternating between similar sounds.
5. The model generates text that is highly phonetically complex and difficult to say, consistently exploiting repetition of closely related sounds and alliteration.

Fluency

Is the model proficient in generating grammatical and well-formed text?

1. The model produces text with no grammatical phrases or spans.
2. The model generates text that is largely ungrammatical but with some grammatically valid sequences.
3. The model generates text that contains grammatically valid sequences but overall is difficult to parse.
4. The model generates text that is generally grammatical and well-formed, with minor errors or awkward phrasing.
5. The model produces text that is grammatically correct and well-formed, demonstrating a strong command of the English language.

Coherence

Is the model proficient in generating semantically coherent and logical text?

CHAPTER 4. TRAIN & CONSTRAIN

1. The model neglects to generate semantically coherent text, producing text that is nonsensical in meaning.
2. The model generates text that is mostly incoherent, with only occasional hints of logical meaning.
3. The model generates text that is partially coherent, but the text lacks logical structure and consistency.
4. The model generates text that is generally coherent, with a logical structure and clear meaning, though minor inconsistencies may be present.
5. The model excels in generating semantically coherent text, where the text is logically structured and maintains a clear and consistent meaning.

Entertainment

Is the model proficient in generating text that a human reader would find entertaining or amusing?

1. The model demonstrates no creativity, either in the content or the structure of the text, resulting in uninteresting or unamusing outputs.
2. The model generates text with minimal creativity, resulting in outputs that are only slightly interesting or amusing.
3. The model generates text that is somewhat entertaining or amusing, but the creativity in content and structure is limited.
4. The model generates text that is generally entertaining and amusing, showing noticeable creativity in content and structure, though some outputs may be less engaging.
5. The model excels in creating entertaining and amusing text, demonstrating creativity in both content and structure and consistently producing engaging and enjoyable outputs.

Overall Quality

Is the model proficient in generating high quality English tongue twisters?

1. The model fails to generate high quality text that is recognisable as a tongue twister.
2. The model generates text that slightly resembles a tongue twister.
3. The model generates text that is recognisable as a tongue twister, but is lacking in fluency, coherence, or entertainment value.
4. The model generates text that is easily recognisable as a tongue twister, but may fall short of high quality.
5. The model excels in generating texts that resemble high-quality tongue twisters that are difficult to pronounce, make sense, and are entertaining.

4.12.2 Implementation Details

Table 4.20 presents the hyperparameters used for model fine-tuning. All other hyperparameters remained at default.

Hyperparameter	Setting
Epochs	10
Batch Size	16*
Optimiser	Adam
Learning Rate	1e-5
Maximum Generation Length	100
Maximum Source Length	512
Hardware	Nvidia A40

Table 4.20: *Fine-tuning hyperparameters. *8 for Baichuan w/LoRa.*

Part IV

Humour Explanation

Chapter 5

Comparing Apples to Oranges: Analysing LLM Understanding from Traditional Puns to Topical Humour

Findings of EMNLP ([Loakman et al., 2025c](#))

Tyler Loakman^a, William Thorne^a and Chenghua Lin^b

^aDepartment of Computer Science, University of Sheffield, UK

^bDepartment of Computer Science, The University of Manchester, UK

Contextual Statement

Transitioning away from the humour generation discussed in Chapter 3 and Chapter 4 through tongue twisters, this chapter addresses RQ4, investigating the extent to which large language models (LLMs) are able to explain different types of humour, from semantics-based homographic puns and sound-based heterographic puns, to esoteric topical jokes. This chapter continues the focus on subjective language through humour, as explored in previous chapters. However, additional emphasis is placed on the reasoning abilities of LLMs and their ability to retrieve relevant information based on subtle and ambiguous references to contemporary topics. The impact of phonetics and phonology continues to be addressed through the challenges posed by heterographic puns, which primarily require models to understand humour where the pun word relies on its pronunciation similarity to an alternative word, leading to phonetic hallucinations where models detect non-existent similarities between unrelated words.

Abstract

Humour, as a complex, communicative, and often esoteric language form, can be derived from myriad aspects of life. However, existing work on computational humour has focused almost exclusively on short, self-contained puns. Humour explanation presents significant reasoning-based challenges for computational models. In this work, we investigate whether the ability of large language models (LLMs) to explain humour depends on joke type. Specifically, we compare models on simple puns that rely on common knowledge to more complex contemporary topical humour that requires knowledge of real-world entities and events. To do so, we curate a set of 600 jokes from different domains, including standard heterographic and homographic puns, as well as contemporary internet humour and topical jokes, where adequate understanding is reliant on reasoning beyond “common sense”, rooted instead in world knowledge regarding news events, politics, pop culture, and more. Using this dataset, we compare the zero-shot abilities of a range of open- and closed-source LLMs of various sizes to accurately and comprehensively explain jokes of different types, identifying key research gaps in the task of humour explanation. We find that few models reliably generate adequate explanations of all joke forms, further highlighting the narrow focus of most work in computational humour.

5.1 Introduction

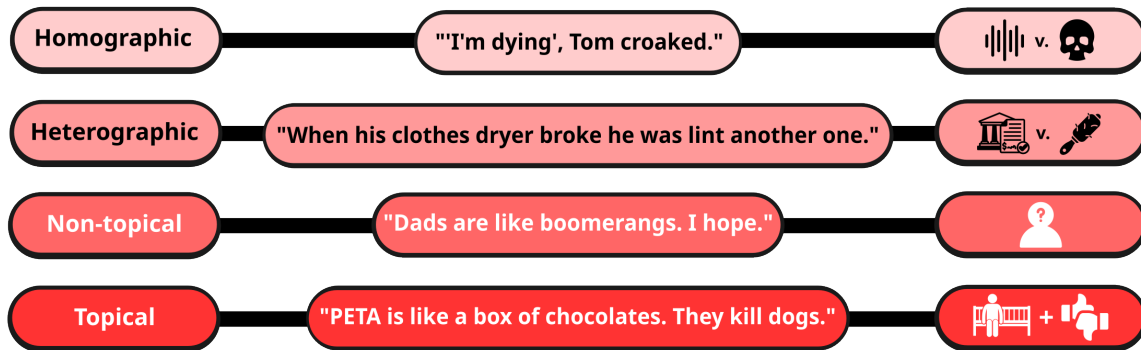


Figure 5.1: *Examples of the joke taxonomy.*

The perception and understanding of humour is essential to human interaction and entertainment. Consequently, the ability to correctly process jokes is paramount to achieving a human-like understanding of ambiguity and social context in language. However, most existing work on computational humour has narrowly focused on the simpler task of humour detection (Loakman et al., 2023a; Meaney et al., 2021; Miller et al., 2017), which allows models to learn from confounding signals to detect certain joke types (Lima Inácio et al., 2023). On the other hand, humour *explanation* tasks a model with the more challenging objective of explicitly outlining why a presented text is humorous (Hessel et al., 2023).

In addition to the general focus on joke detection, existing works on humour processing have mostly focused on short-form puns such as “*Yesterday I accidentally swallowed some food colouring. The doctor says I’m OK, but I feel like I’ve dyed a little inside*” (He et al., 2019), where humour arises from the interpretation of the punning word (“dyed”) with a phonetically similar word that has a different meaning (“died”) in the case of heterographic puns, or polysemy for homographic puns (Attardo, 2008). Whilst puns are commonplace, the humour is primarily self-contained, requiring only common sense (e.g., that food colouring “dyes” things) and

CHAPTER 5. COMPARING APPLES TO ORANGES

an understanding of the semantics of dictionary words. On the other hand, much of the humour encountered online, on TV panel shows, and in stand-up comedy routines, is based on contemporary topical knowledge of evolving pop culture phenomena and news events, where a full appreciation of the humour relies on potentially esoteric knowledge, rather than common sense (Highfield, 2015; Laineste, 2002), requiring highly complex reasoning and knowledge retrieval.

We present examples of each joke type that we investigate in Figure 5.1. The homographic pun exploits the dual meaning of “croaked” as both a style of speech and a euphemism for dying; the heterographic pun relies on the phonetic similarity between “lint” and “leant”; the non-topical joke plays on the trope of an absentee father (not) returning like a boomerang; and the topical joke refers to the animal welfare organisation PETA’s high euthanasia rates and the movie *Forrest Gump*. Each joke type plays on polysemy, phonetics, social constructs, and esoteric knowledge, respectively.

To address this research gap, we assess the ability of large language models (LLMs) to explain this range of joke forms in a zero-shot setting. As a result, we present the first work on assessing whether state-of-the-art LLMs are equally able to explain different joke formats, ranging from those that require an understanding of basic semantics and phonetics to those that require esoteric knowledge of pop culture and news events, therefore assessing the ability to perform complex reasoning and knowledge retrieval.

In this work, we address the primary research questions of: (i) whether different joke formats have an impact on the ability of LLMs to explain humour, and by extension, (ii) whether findings from existing work on computational humour are generalisable to the different jokes found in everyday life.

We summarise our main contributions below:

- We present a novel, balanced dataset of 600 jokes, categorised into homographic and heterographic puns, non-topical Reddit jokes, and topical Reddit jokes. Each joke is paired with a high-quality, succinct, human-authored reference explanation, enabling an empirical assessment of how joke format affects the ability of LLMs to explain humour.
- We assess the accuracy and completeness of 4800 joke explanations across eight state-of-the-art open- and closed-source LLMs of various sizes, including distilled reasoning models (600 jokes processed by eight models). We analyse explanations using human evaluation and automatic metrics, including an LLM judge.
- We qualitatively evaluate the models’ abilities to understand complex humour through a case study of generated explanations for a topical online joke.

5.2 Related Works

Humour Generation. Early humour generation efforts relied on simple template-based systems to create puns (Valitutti et al., 2013; Ritchie, 2005), whilst contemporary works fine-tune deep learning models on extensive datasets of existing jokes (Garimella et al., 2020; Weller et al., 2020; He et al., 2019; Petrović and Matthews, 2013). More recently, it has become the primary paradigm to use different prompting strategies for generating jokes (Chen et al., 2024; Mittal et al., 2022). Furthermore, Sun et al. (2022b) presented CUP, which turned the focus from stand-alone joke generation to context-situated generation. Work has also been performed on humour-adjacent language forms such as tongue twisters (Loakman et al., 2025b, 2023b; Keh et al., 2023).

Humour Detection. Similarly, humour detection has evolved from explicitly using lexical and syntactic features (van den Beukel and Aroyo, 2018) to training language model classifiers (Meaney et al., 2021; Wang et al., 2020; Weller and Seppi, 2019). Ao et al. (2022b) additionally present work using humour to aid in the detection of political parody, whilst Meaney (2020) and Loakman et al. (2023a) investigated the impact of demographic variables on humour perception. Furthermore, some works have approached humour detection from a multimodal standpoint, using elements such as audience laughter to aid in detection (Hasan et al., 2019).

Humour Explanation. Lastly, humour explanation provides an account of why a text is funny. Consequently, this task requires models to exhibit extensive reasoning capabilities. Whilst explaining stand-alone puns usually only requires a model to identify the pun word and its alternative interpretations in context, more complex jokes require an understanding of common sense, social constructs, and esoteric knowledge to identify key incongruities. Unlike detection, explanation requires models not only to identify humour, but also “understand” it, providing a natural language explanation. Miller et al. (2017) present a shared task on computational humour, in which subtask three pertains to humour interpretation, tasking models with assigning the correct WordNet sense keys to instances of pun words. Lima Inácio et al. (2023) tackle the humour explanation task indirectly by investigating the specific features learnt by humour classifiers. Regarding the production of natural language explanations, Hessel et al. (2023) perform the task of humour explanation on multimodal data from caption contests in The New Yorker, generating explanations as to why a particular best-rated caption is humorous. Finally, Wang et al. (2025b) investigate the ability of LLMs to explain a wide range of linguistic phenomena that require non-linear reasoning, such as tautologies and sarcasm. Our work empirically investigates the challenges posed by

different forms of humour, resulting in a performance gap. Most notably, we reveal the challenges posed by contemporary topical jokes.

5.3 Dataset

5.3.1 Existing Datasets

Many existing datasets focus on size, rather than annotation depth and consistency. For example, r/Jokes (Weller and Seppi, 2020) presents 550k jokes taken from Reddit with metadata such as upvote count, whilst SemEval 2017 Task 7 (Miller et al., 2017) consists of 3387 simple puns. Whilst Miller et al. (2017) add semantic annotation through WordNet senses, there are no natural language explanations. Sun et al. (2022a) improve upon the SemEval dataset by crowdsourcing textual explanations for a subset of the aforementioned puns. Furthermore, Hessel et al. (2023) present a dataset of multimodal humour from The New Yorker cartoon caption contest, providing written explanations and Wikipedia links to required world knowledge. Finally, whilst *ExplainTheJoke.com*¹ contains detailed joke explanations, the majority are simple puns and the explanations frequently lack objectivity, being highly editorialised for entertainment purposes.

5.3.2 Our Dataset

In the following section, we outline the process of creating our own humour analysis dataset with a larger range of joke types and concise, objective explanations. Our dataset contains 600 jokes, consisting of 150 of each type: homographic puns, heterographic puns, non-topical Reddit humour, and topical Reddit humour. Figure 5.2 presents the average joke lengths within each subset, whilst Figure 5.1 presents examples of each subtype. See Appendix 5.12.3 for further examples.

¹<https://explainthejoke.com/>

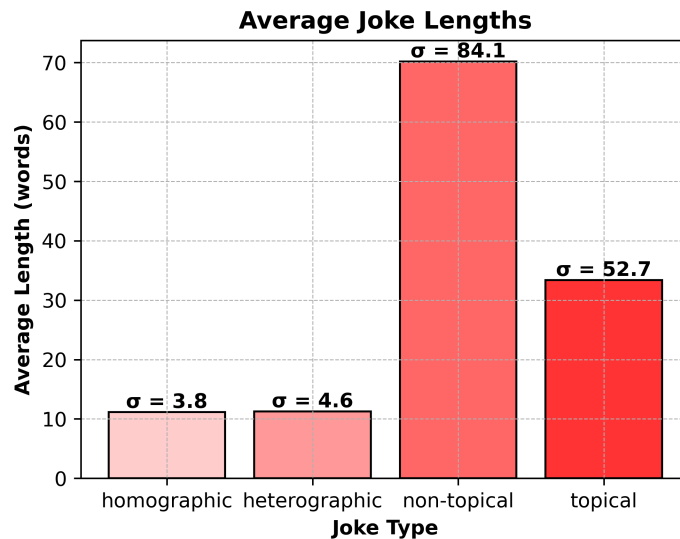


Figure 5.2: Average joke lengths. Standard deviations are presented on top of each bar.

Topical Humour. To identify jokes that require knowledge of real-world entities and events, we use a subset of the r/Jokes dataset (Weller and Seppi, 2020). Whilst Reddit contains potentially unsavoury material, the presence of strong opinions on topics such as politics and pop culture results in a rich array of topical humour. Whilst r/Jokes contains a plethora of different topics, not all joke instances rely on contemporary world knowledge. We therefore perform filtering based on upvotes and the presence of named entities (see Appendix 6.10.1 for more details). Of the filtered jokes, we take the 150 with the most upvotes. We remove jokes that do not require external knowledge, in addition to meta-level humour that relies on knowing that the joke came from Reddit (where this is not explicitly mentioned in the joke itself).

Non-topical Humour. We compare the results of topical humour explanation with other common forms of internet humour that do not require esoteric world knowledge. These jokes often present different forms of humour than the simple puns that have been the focus of most existing works on automatic joke explanation. Consequently, we select

another 150 top-scoring jokes from the r/Jokes dataset that do not require knowledge beyond what may be considered “common sense” (e.g., cultural norms and stereotypes). Both the topical and non-topical subsets are checked by another linguist, who is not an author of this work, to ensure their correct categorisation. In effect, the 300 jokes that form the topical and non-topical categories are the top 300 highest-scoring jokes based on upvote-downvote ratio, categorised by their content.

Traditional Puns. We source 300 simple puns from 2017 Task 7 (Miller et al., 2017), which has acted as the source data for many works on computational humour. Specifically, we take the first 150 jokes from both the homographic and heterographic training data for Subtask 3 as the basis of the homographic and heterographic joke subsets.

The jokes contained within the topical and non-topical subsets from Reddit may rely on the features of heterographic or homographic puns. However, the interpretation of these puns requires additional real-world knowledge besides the meaning of the pun words. For example, “*New Teslas don’t come with a new car smell. They come with an Elon Musk*”, requires knowledge regarding billionaire Elon Musk’s relationship to the car manufacturer Tesla.

5.3.3 Ground Truth Joke Explanations

As a central contribution, we write ground-truth explanations for all jokes. Due to being written by an author of this work, we perform quality assurance using 3 native English-speaking computer science students. Each evaluator assessed 200 explanations (50 per type), each seeing different jokes.² We evaluate these explanations using the

²Participants were given a 30 GBP Amazon voucher for approximately 1.5 hours of evaluation.

same evaluation criteria used for human evaluation in §5.5 (with the rubric presented in §5.4.3), using a 0-5 Likert scale on the criteria of accuracy and completeness. All evaluators gave all human-authored explanations the maximum score of 5 on both criteria or commented on required improvements to be implemented in revisions. Minor explanation post-editing was required for less than 1% of explanations. Further details are presented in Appendix 5.12.1.

5.3.4 Dataset Enhancement

We also present URLs to webpages containing the knowledge required to understand jokes from the topical subset, including Wikipedia pages, news articles, and blog posts. Additionally, we include phonetic transcriptions of all jokes after identifying model hallucinations stemming from a lack of phonetic knowledge (see details in Appendix 5.12.2). Neither of these fields is used in the following analysis.

5.4 Methodology

In the following section, we present our hypotheses relating to the research questions posed in §5.1, in addition to the model selection and explanation scoring criteria.

5.4.1 Hypotheses

In relation to our research questions, we pose the following hypotheses:

- **H1** - Traditional puns (homographic and heterographic) are easier for LLMs to explain than high-rated jokes from Reddit, owing to the former’s frequent reliance on the semantics and phonetics of common words.

- **H2** - Heterographic puns are more difficult for models to explain than homographic puns due to the former’s reliance on phonetic similarity, which is not explicitly encoded in orthographic text.
- **H3** - Topical humour is more difficult for models to explain than non-topical Reddit humour, due to the former’s reliance on subtle references to contemporary pop culture and events, rather than common sense reasoning and general knowledge.
- **H4** - Larger model variants perform better than smaller variants, particularly for topical humour, due to being able to store larger amounts of information in their parameters regarding specific events and individuals, whilst smaller models focus on retaining common knowledge.

5.4.2 Models

To assess the ability of state-of-the-art LLMs in this task, we select the following suite of open- and closed-source models: Llama 3.1 (8B and 70B) (Dubey et al., 2024), Gemini 1.5 (Pro and Flash) (Georgiev et al., 2024), and GPT-4o (standard and Mini) (OpenAI et al., 2023). These models represent a range of sizes, using both the flagship large variants and their lightweight counterparts, allowing investigation into the effects of model size. We additionally use variants of DeepSeek-R1 to identify the abilities of reasoning-specific models. Specifically, we use distilled 8B and 70B variants based upon Llama 3.1 and 3.3, respectively.

Importantly, all selected models were trained on data pertaining to events up to at least 2023, with all jokes in the non-topical and topical subsets being posted to Reddit between 2008 and 2019. Therefore, we are fairly assessing the ability of these models to explain jokes that reference events contained within their training data, investigating

the models’ ability to associate this knowledge with the abstract references within topical jokes.

All models were tasked with producing an explanation of approximately 100 words, based on the lengths of the gold standard explanations. Additional details, including the full prompt and specific models used, are presented in Appendix 5.12.1.

5.4.3 Evaluation Criteria

To assess the quality of joke explanations, we propose a scoring rubric with two core criteria, accuracy and completeness, each evaluated on a 6-point scale (0-5). Focusing on such criteria allows the application of the scoring rubric to all assessed joke types, such as real-world references in the topical subset, or pun word definitions in traditional puns, constituting completeness. The rubric is presented below:

Explanation Accuracy

0. The explanation is entirely incorrect, failing to identify the humour or context correctly.
1. The explanation is mostly incorrect, with minimal recognition of the joke’s intent.
2. The explanation is largely inaccurate, with only a few correct points amidst significant errors.
3. The explanation contains mostly correct elements but includes noticeable errors that may hinder understanding.
4. The explanation is mostly accurate with minor errors or omissions that do not significantly impact understanding.

CHAPTER 5. COMPARING APPLES TO ORANGES

5. The explanation is fully accurate, correctly identifying the joke’s main humour and context without errors.

Explanation Completeness

0. The explanation is entirely superficial or irrelevant, failing to address any key components.
1. The explanation is minimal, covering only a small part of the joke.
2. The explanation is incomplete, addressing only basic aspects whilst missing significant details.
3. The explanation addresses the central humorous element but omits some important details that maximise understanding.
4. The explanation covers most aspects but misses minor details.
5. The explanation is comprehensive, covering all relevant aspects and nuances.

Accuracy concerns whether the explanation contains correct material, assessing the presence and absence of hallucinations and misunderstandings. On the other hand, Completeness concerns whether the fundamental elements of a joke have been addressed. Consequently, a joke may be fully explained (i.e., completeness = 5), but additional information has been included which is incorrect (i.e., accuracy ≈ 3). Conversely, an explanation may identify that a joke is a pun based on a specific word (i.e., accuracy = 5), but omit explanations of the pun’s meanings or other required information for understanding (i.e., completeness ≈ 2).

Due to the significant time requirements of rating 4800 explanations (≈ 40 hours), one author (who is a native speaker of English and holds degrees in Linguistics) rated all explanations. However, a subset of 320 explanations (10 jokes from four joke types,

explained by eight models) was re-annotated by two third-party computer science students (who are also native speakers of English).³ These annotations were used exclusively to verify the primary evaluator’s reliability. The results showed moderate levels of agreement for both accuracy ($\alpha = .574$) and completeness ($\alpha = .553$) using Krippendorff’s alpha (Krippendorff, 2019). We additionally observed high positive Pearson’s correlations (Pearson, 1895) between the scores assigned by the primary human evaluator and third-party evaluators for accuracy ($r = .765$) and completeness ($r = .795$), both with $p < .001$. Overall, 82% of accuracy scores and 84% of completeness scores from the primary evaluator and the third-party evaluators differed by 0 or 1 points, resulting in an 80.5% agreement between the main evaluator and third-party evaluators when converted to a binary pass or fail (as in §5.5.1). Across the whole dataset, we additionally leverage the LLM-as-a-judge paradigm using Qwen 2.5 (72B Instruct), finding agreement for accuracy ($\alpha = .565$) and completeness ($\alpha = .519$) to be substantial. This is additionally supported by high Pearson’s correlations for accuracy ($r = .641$) and completeness ($r = .602$), both with $p < .001$. See Appendix 5.12.4 for details on the LLM-as-a-judge set up, and Appendix 5.12.5 for a deeper analysis of the ratings from the third-party evaluators and LLM judge.

5.5 Human Evaluation

Average explanation lengths, accuracy, and completeness scores are presented in Figure 5.3. Firstly, explanation length remains consistent within models regardless of joke type, suggesting that joke length does not necessarily impact the required explanation length. Notably, the Gemini-based models produce the shortest explanations across all joke types, whilst GPT-4o-based models consistently generate the longest explanations,

³Participants were given a 60 GBP Amazon voucher for approximately four hours of evaluation.

CHAPTER 5. COMPARING APPLES TO ORANGES

opting for more verbose outputs. However, the shorter explanations produced by the Gemini models do not negatively impact their accuracy or completeness scores. Despite being more concise, the Gemini models appear to be efficient, focusing on the essential elements of an explanation without adding superfluous detail.

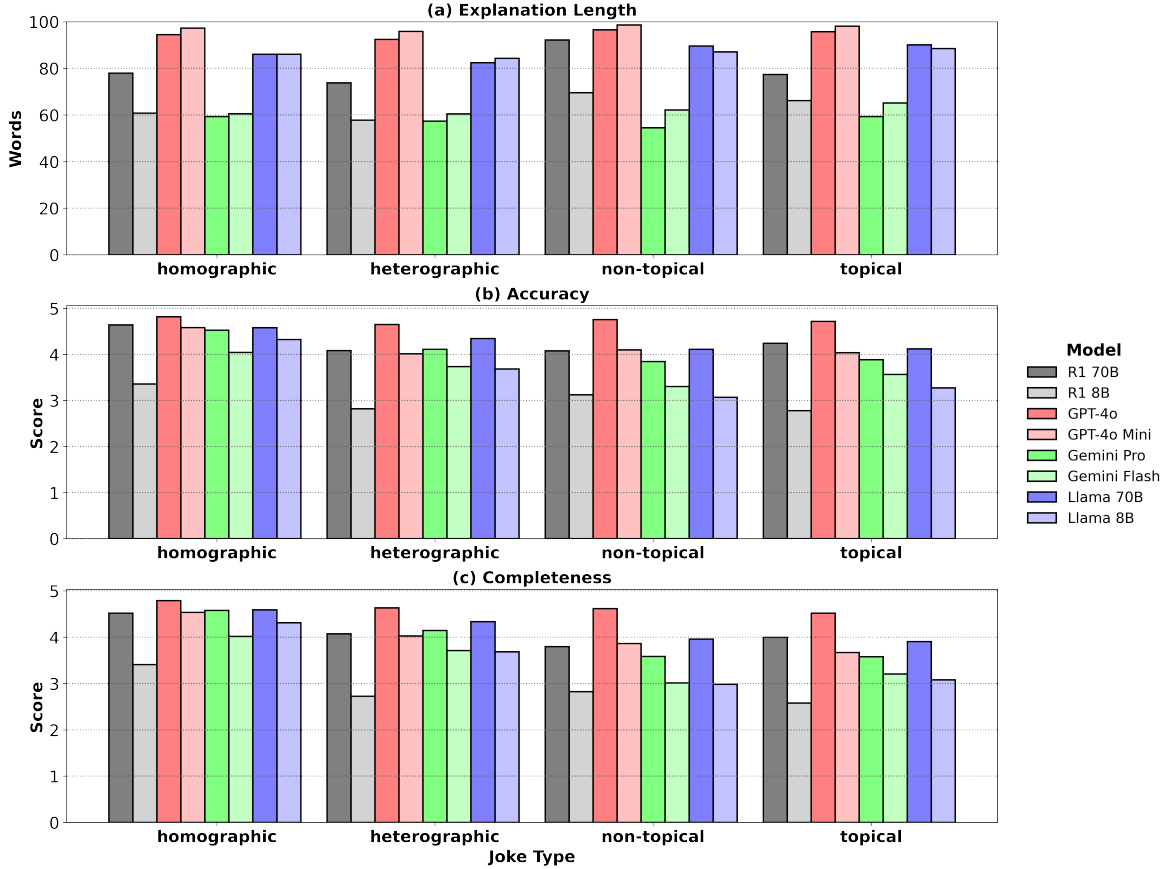


Figure 5.3: Average explanation length, accuracy score, and completeness score, by model and joke type.

In terms of overall performance, completeness scores are generally lower than accuracy across all models. This difference indicates that the models are more likely to miss key details in their explanations than to hallucinate incorrect information or misinterpret jokes entirely. This trend suggests that whilst the LLMs may recognise the core components of a joke, they often fail to elaborate on all relevant aspects that contribute to humour.

When comparing model performance, GPT-4o consistently outperforms all others, demonstrating the highest accuracy and completeness scores. Interestingly, GPT-4o Mini outperforms Gemini Pro in some specific cases, such as explanation accuracy for homographic puns, as well as in both accuracy and completeness for non-topical and topical jokes. Additionally, open-source Llama 70B outperforms most models except GPT-4o, whilst the R1-based reasoning models underperform most non-reasoning models.

5.5.1 Explanation Success Rate

For an explanation to be useful, it must be both accurate and complete, rather than scoring highly on only one criterion. In Figure 5.4 we present a binary categorisation of explanations, treating scores of 4 or above on both criteria as being “good” quality explanations (i.e., containing only minimal, non-intrusive errors, and/or omitting minor details). This facilitates a clearer comparison of different joke types than raw scores.

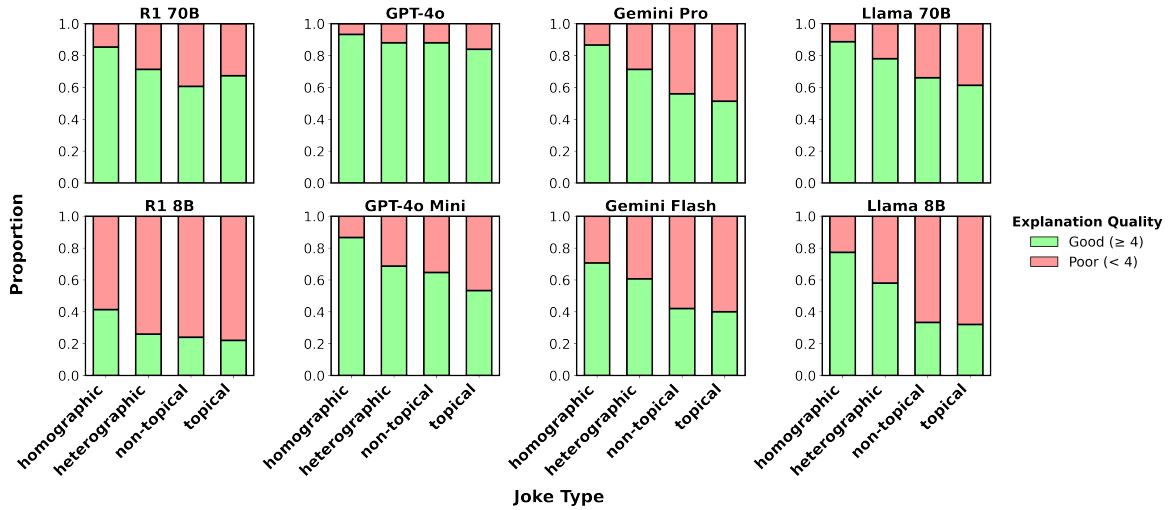


Figure 5.4: Proportion of “good” and “poor” joke explanations by model and joke type. A “good” explanation is defined as one that scores ≥ 4 on both accuracy and completeness.

CHAPTER 5. COMPARING APPLES TO ORANGES

As illustrated in Figure 5.4, explanation success rate varies significantly across joke types. Homographic puns, where a word has multiple meanings but identical spelling, consistently yield the highest proportion of successful explanations across most models. This suggests that it is easier to navigate the ambiguity of homographs, likely due to language models being aware of the polysemy of the pun words.

In contrast, heterographic puns, which involve words that sound the same but are spelt differently, pose more difficulty for all models. GPT-4o is the only model that maintains a relatively high success rate in this category. The challenge likely stems from the additional complexity introduced by needing to recognise the phonetic similarity of the pun word whilst being trained on graphemes.

Non-topical jokes also show varying levels of performance. The non-topical subset consists of jokes from Reddit, including puns, incongruities and situational irony that cover a wide range of joke forms that have been rated highly by the Reddit user base, causing the LLMs to struggle more severely than traditional puns.

Lastly, topical jokes present the most challenge, requiring contextual awareness of specific events. Whilst GPT-4o generates a relatively high proportion of “good” explanations, other models are substantially less successful. This is likely due to the subtle references to pop culture and real-world entities in the topical subset presenting a significant challenge, requiring both high levels of reasoning and advanced information retrieval based on ambiguous references.

Regarding the DeepSeek reasoning models, whilst R1 70B performs well overall, its reasoning capabilities do not see it outperform any other tested model. On the other hand, the 8B variant is the worst-performing model. We hypothesise that this is due to the focus on mathematical and logical “reasoning”, rather than common sense. Owing to the reasoning process, R1 8B may begin with a moderately good answer (in line with the other tested models), whilst each reasoning step introduces additional uncertainty.

causing hallucination. This results in incorrect reasoning before presenting the final output, therefore becoming progressively worse.

We perform a logistic regression analysis, which further supports our hypotheses. Model size has a strong and significant impact on explanation quality, with larger variants being far more likely to generate “good” explanations ($\beta = 1.707$, $p < 0.001$). This reinforces our finding that larger models significantly outperform smaller ones. Regarding joke difficulty, the results align with the predicted ordering. Homographic jokes are the easiest to explain ($\beta = 0.583$, $p < 0.001$), whilst non-topical ($\beta = -0.511$, $p < 0.001$) and topical jokes ($\beta = -0.574$, $p < 0.001$) are significantly harder. This confirms that jokes requiring external context pose a greater challenge to LLMs than those based on simple semantics and phonetics alone. Overall, the model is highly significant ($p < 0.001$) and explains a meaningful portion of the variance (pseudo $R^2 = 0.138$).

5.6 Automatic Evaluation

In addition to the previously presented human evaluation, we provide the results of automatic metrics in Table 5.1. Whilst existing reference-based automatic metrics are not well designed for this task, we do confirm our originally hypothesised ordering of joke difficulties, with the explanations of homographic/heterographic puns scoring higher than the non-topical joke subset, which in turn scores higher than the topical jokes subset on most metrics.

	Homographic	Heterographic	Non-Topical	Topical
SacreBLEU	10.15	8.53	7.86	7.09
ROUGE-1	0.43	0.41	0.41	0.40
ROUGE-2	0.15	0.12	0.13	0.12
ROUGE-L	0.28	0.25	0.25	0.23
METEOR	0.39	0.37	0.32	0.32
BERTScore	0.89	0.88	0.88	0.87

Table 5.1: *Automatic evaluation of LLM explanations. Human-authored gold explanations are used as the reference. A further breakdown by model and joke type can be found in Appendix 5.12.6.*

5.7 Case Study

Example explanations of a randomly selected joke from the topical subset are shown in Figure 5.5. The joke refers to the 2018 internet phenomenon of the “Tide Pod Challenge”, where teenagers dared each other to consume toxic laundry pods from the brand Tide. The central element to the joke’s explanation is the suggestion that this new “*teenage snack food market*” increased sales, ergo generating enough revenue for the company to afford notoriously expensive advertising at the Super Bowl American football event. Interestingly, all of the full-size models inferred the reference to this challenge in their explanations. However, the smaller counterparts consistently omitted this information or presented misinterpretations in their explanations. GPT-4o Mini, for example, explains the joke by suggesting that “*teenagers are known for being messy eaters*”, causing increased sales. On the other hand, Llama 8B and Gemini Flash focus on the idea that humour stems from “absurdity”, without explicitly mentioning the real-world phenomenon, thereby failing to inform the user of important context. A further discussion of explanation characteristics can be found in Appendix 5.12.3.

CHAPTER 5. COMPARING APPLES TO ORANGES

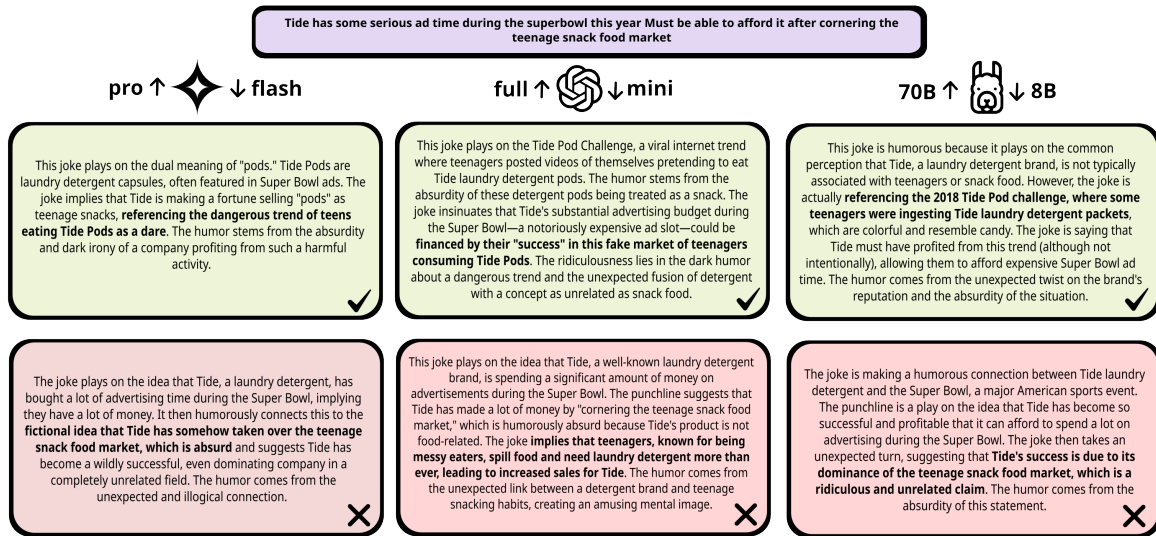


Figure 5.5: Case study of LLM explanations of a topical joke surrounding the “Tide Pod Challenge”. From left to right: Gemini 1.5, GPT-4o, and Llama 3.1. Larger models are along the top, with smaller variants below.

5.8 Discussion

H1 predicted that traditional puns would be easier for models to explain than high-rated Reddit jokes. This was found to be true, with the non-topical and topical subsets significantly impacting explanation success rates.

Similarly, H2 proposed that homographic puns, which rely on polysemy, would be easier to explain than heterographic puns. This was true in our testing. We observed that heterographic puns more frequently lead to hallucination (i.e., lower accuracy scores), owing to their reliance on phonetic characteristics, whilst LLMs have only implicit knowledge of how words are pronounced.

H3 posited that the topical subset would present the most difficulty for LLMs. Our testing revealed a statistically significant decrease in explanation success rates for topical humour compared to all other joke formats, with many explanations falling short of correctly identifying contemporary real-world references.

CHAPTER 5. COMPARING APPLES TO ORANGES

Finally, H4 stated that the larger models would outperform smaller variants. Our results showed that this was indeed the case, with R1 70B outperforming R1 8B, GPT-4o outperforming Mini, Gemini Pro outperforming Flash, and Llama 70B outperforming Llama 8B.

As predicted, the topical subset presented the most difficulty for LLMs. Whilst these jokes were significantly shorter than the non-topical subset (see Figure 5.2), their reliance on subtle references to esoteric entities and events required the models to identify information that was not explicitly stated. This creates an environment ripe for hallucinations, as the models attempt to generate explanations without correctly identifying the location of the punchline. We believe that larger models struggled less than smaller models with topical jokes due to the additional parameters allowing the storage of more specific knowledge that does not appear in the training data as frequently as common sense concepts.

5.9 Conclusion

We presented the first empirical assessment of how well LLMs are able to explain the humour of jokes of various types, including simple puns and humour that relies on contemporary topical knowledge. Our findings demonstrate that the primary focus on puns in prior work is not representative of the abilities of LLMs to accurately and comprehensively explain a broader array of joke types. Furthermore, we observe that none of the tested state-of-the-art models can consistently explain jokes of any type.

5.10 Limitations

In this work, we investigated the extent to which joke complexity affects the ability of LLMs to generate accurate and complete explanations. However, topical humour continuously evolves, with new jokes being created that relate to developing news stories and pop culture events. Consequently, our work focuses only on a subset of possible jokes. We do not assess the ability of models to explain jokes where there is unlikely to be any knowledge contained in the model’s pre-training (as the volume of reporting and discussion on a given topic grows over time). Additionally, we performed a comparatively small-scale analysis of 600 jokes (150 per type). Whilst this is necessary to enable manual human assessment of the generated explanations and authoring gold-standard reference explanations, a wider range of jokes would allow deeper insight. Due to the presented jokes being from secondary sources (i.e., not novel jokes written for this work), it is likely that most jokes had been seen previously by the LLMs. However, explanations relating to these jokes are much rarer. For the Reddit jokes in particular, the presence of explanations in the training data relies on individual Reddit users asking for an explanation within the relevant thread, and someone replying. We sorted the r/Jokes dataset by upvotes, and it is unlikely that r/Jokes users require explanations of examples from popular on r/Jokes itself, owing to being part of the same online community. We manually verified a subset and found readily available explanations rare. Whilst one author performed the majority of the evaluation, we robustly ensured reliability through comparisons with third-party annotators and an LLM-judge.

5.11 Ethics Statement

Whilst the dataset element of our benchmark contains potentially offensive materials (reflecting the often controversial nature of humour found on Reddit), this resource aims to offer explanations of potentially esoteric jokes in order to clarify where particular parties may be perceiving humour. We do not create novel jokes, nor present the included dataset as appropriate for training humour generation models. Whilst detoxifying and debiasing language models is of growing interest in the NLP community, humans are fundamentally opinionated and biased in ways that others may take offence at. Whilst models should not perpetuate these biases, the authors of this work believe that language models should be able to provide impartial, objective explanations of phenomena such as humour. We therefore do not view the creation of an *explanation* as equal to an *endorsement* of said joke, but rather as an ideologically neutral task. We additionally write our humour explanations objectively, not lending unjust credence to unverified scandals or presenting opinions as facts.

5.12 Appendix

5.12.1 Implementation Details

r/Jokes Filtering. To select the topical subset, we sort r/Jokes by the “score” calculated by [Weller and Seppi \(2020\)](#) based on the upvote-downvote ratio to find the most popular jokes. Following this, we performed named entity recognition using SpaCy, keeping jokes that contained the following entity types: ORG (organisation), FAC (facility), WORK_OF_ART, LAW, ORG (organisations), PROD (products), EVENT, and PERSON. This filtering allows us to easily identify r/Jokes entries that are likely to require esoteric world knowledge. This process reduced the dataset to just over 153k samples, of which the 150 highest scoring are taken following manual verification.

Prompt. For all models, we used the following prompt: “Explain the following joke (presented in square brackets) in approximately 100 words.\n\n[JOKE]\n\n”. This is based on good practice for clearly marking the boundaries between the task and the joke instance, whilst the 100-word length is based on the average length of the human-authored explanations.

Models. We specifically made use of the following model variants: GPT-4o (`gpt-4o-2024-05-13`), GPT-4o Mini (`gpt-4o-mini-2024-07-18`), Gemini 1.5 Pro (`gemini-1.5-pro-001`), Gemini 1.5 Flash (`gemini-1.5-flash-001`), Llama 3.1 8B (`meta-llama/Meta-Llama-3.1-8B-Instruct`), Llama 3.1 70B (`meta-llama/Meta-Llama-3.1-70B-Instruct`), Deepseek R1 8B (`deepseek-ai/DeepSeek-R1-Distill-Llama-8B`), and DeepSeek R1 70B (`deepseek-ai/DeepSeek-R1-Distill-Llama-8B`). For GPT-4o and Gemini, these refer to the stable releases available at the time (August 2024). All model hyperparameters were at the default. We use the Google API with all user-accessible safety limits disabled due to the potentially offensive content of some jokes.

Inference. OpenAI models were prompted via the OpenAI API, whilst the Google models used the Google GenerativeAI API. Llama 8B was run on a desktop PC with an Nvidia RTX 4080 Super, whilst Llama 70B, both R1 models, and the LLM Judge were run on an HPC node using two Nvidia A100s.

Participants. We recruited two third-party participants to evaluate a subset of 240 explanations. In compensation, they received a 60 GBP Amazon voucher for approximately three hours of work. Three participants also evaluated the gold standard reference explanations (200 each), for which they received a 30 GBP Amazon voucher. The third-party checks of joke categorisation did not involve monetary compensation.

5.12.2 Dataset Enhancement Details

Phonetic transcriptions in the International Phonetic Alphabet (IPA) are provided for all jokes in the dataset using Anthropic’s Claude Sonnet 3.5 (v.20241022), owing to advanced handling of factors such as abbreviations over traditional grapheme-to-phoneme (G2P) models. A subset of these transcriptions were manually verified. An additional test phrase of “*The pikachu weighs 45 lbs and is 5ft 9”. He works at the CIA but previously worked at NASA*“, was presented to verify the model’s handling of non-dictionary terms, abbreviations, initialisms and acronyms, which it did successfully. Transcriptions were requested in the General American (GenAm) accent, owing to the primary user base of Reddit being North Americans.

We provided a clear, canonical task description and requested the model to construct its own prompt. Prior work has shown this to be an effective strategy (Zhou et al., 2023b). The resulting prompt is as follows:

“You are a linguistic transcription tool. For any text samples provided:\n\n1. Convert EVERY sample into IPA format without exception - no samples

CHAPTER 5. COMPARING APPLES TO ORANGES

should be skipped or partially transcribed\n 2. Use General American (GenAm) pronunciation standards consistently\n 3. If uncertain about any transcription, ask for clarification rather than making assumptions\n 4. Do not proceed without transcribing ALL text samples provided\n\n Example input:\n 'Hello world'\n\n Expected output:\n hələv wəˈld\n\n Please proceed with the transcription.”

5.12.3 Additional Qualitative Analysis

Non-topical Joke Examples	
1	“What do you call a book club that’s been stuck on one book for years? Church.”
2	“TIL unvaccinated children are less likely to be autistic. Because they are more likely to be dead.”
3	“Set your wifi password to 2444666668888888. So when someone asks tell them it’s 12345678.”
4	“My girlfriend is like the square root of -100. A solid 10, but also imaginary.”

Table 5.2: *Examples of jokes from the non-topical subset requiring different types of common sense knowledge.*

Topical Joke Examples	
1	“If America is storming Area 51 then the Europeans can storm the Vatican. We’ll take the aliens, you get the predators.”
2	“For anyone attending Stan Lee’s funeral... Make sure you stay after the ceremony is finished.”
3	“R Kelly is really changing the rap game. He takes the art out of rap artist.”
4	“Obama smoked weed growing up, and now look where he is today. Unemployed with two kids and recently evicted.”

Table 5.3: *Examples of jokes from the topical subset that require different types of world knowledge.*

Table 5.2 and Table 5.3 present examples of non-topical and topical jokes from Reddit, respectively. Regarding the non-topical jokes, the first example plays on common

CHAPTER 5. COMPARING APPLES TO ORANGES

knowledge regarding religious practices and the focus of churches on the bible. The second joke is a form of deadpan humour that refers to the higher mortality rates of children who are not vaccinated against preventable diseases. The third joke presents phonetic wordplay, requiring the reader to interpret “12345678” as “one 2, three 4...” and so on. Finally, the last joke plays on the mathematical concept of imaginary numbers.

Regarding the topical jokes in Table 5.3, the first joke makes reference to the media franchise of *Alien v. Predator*, whilst referring to the 2019 news phenomenon where a Facebook event was created to convince people to storm Area 51 to find aliens. The second joke refers to the Marvel film franchise and their penchant for extended post-credits scenes, in addition to Stan Lee’s real-life death. On the other hand, joke three makes light of a celebrity scandal, involving wordplay that requires the sequence “art” to be removed from “rap artist” to identify accusations pertaining to R Kelly. Finally, the last joke refers to Barack Obama losing his presidency to Donald Trump, and therefore being “unemployed” and “recently evicted” from the White House.

Explanation Characteristics. During the process of evaluating LLM responses, several characteristics were observed. Firstly, the Google API refused to allow the Gemini models to produce explanations for some jokes, likely due to triggering content safeguards (five explanations across three jokes, where in one instance the request was only blocked by Flash). Such instances were not notably any more offensive than other examples. In some instances, models provided textual responses regarding their refusal to address a particular joke (however, both cases only account for $\approx 1\%$ of all jokes). Furthermore, in some instances, we observed that explanations contained subjective judgments as to the appropriateness of a particular joke, with the humour being deemed “unacceptable”, without acknowledging the nuance and subjectivity of

such a claim. Importantly, potentially offensive content was not exclusively found in the non-topical and topical Reddit joke subsets, with many of the simple puns containing sexual innuendo and derogatory terms. Furthermore, whilst all of the models tested are capable of producing high-quality coherent text, we observed some instances of the Llama models resorting to degenerate patterns, such as being stuck in repetitive loops. Again, these account for a very small number of jokes (<1%). Regarding the heterographic pun subset, we noticed many instances of hallucination whereby the LLMs would explain that a punning word is phonetically similar to a word that shares little similarity. This is demonstrative of the challenge that heterographic puns present to LLMs, relying on phonetic characteristics, whilst such models are trained on graphemes and therefore have limited phonetic knowledge.

5.12.4 LLM-as-a-Judge Protocol

When selecting our LLM judge, we required that candidates not be from the same model family as any of those under evaluation. Candidates included Prometheus 2 8x7B (prometheus-eval/prometheus-8x7b-v2.0) (Kim et al., 2024), which was specifically developed for evaluation, Qwen 72B Instruct (Qwen/Qwen2-72B-Instruct) (Qwen et al., 2025), and Gemma 2 9B Instruct (google/gemma-2-9b-it) (Riviere et al., 2024). These were selected due to being the best performing models in the Hugging Face Judge Arena⁴ (with preliminary results disabled). At the time of selection (04/02/2025), Qwen was ranked 6th with 1203 ELO and Gemma 13th with 1157. We present results from only Qwen 72B Instruct, due to demonstrating the best alignment with our human evaluation, rather than a jury ensemble or any other individual judge.

Judge hyperparameters were set to model defaults, with the exception of temperature, which was fixed to 0.1 for all models, based on prior work (Sanh et al., 2022). The

⁴<https://huggingface.co/spaces/AtlaAI/judge-arena>

CHAPTER 5. COMPARING APPLES TO ORANGES

scoring criteria were the same as given to the human annotators, except for the ordering of score categories, which were listed in ascending order rather than descending in order to better align with the training prompt of Prometheus. To obtain scores from each judge, we used the following prompt:

```
“You are an expert judge evaluating explanations of jokes. You will be
given a joke, a reference explanation (human-authored), and a model-generated
explanation to evaluate.\n\n You must evaluate the model’s explanation on
the matter of {criteria} using only the following scoring criteria:\n\n
{scoring_criteria}\n\n The joke:\n {joke}\n\n Reference explanation: \n
{reference_explanation}\n \n Model explanation to evaluate:\n {explanation}
\n \n Provide a score for this joke using only the scoring criteria provided.
You must only respond with a single number between 0 and 5. You must produce
no other output.”
```

5.12.5 Additional Annotator & LLM-as-a-Judge Results

Figure 5.6 and Figure 5.7 present the results of the evaluation from the two third-party evaluators we recruit to analyse a subset of joke explanations. Overall, the same trend that was viewed in Figure 5.3 and Figure 5.4 can be seen, with the order of joke difficulty being the same for these evaluators on the tested subset. Owing to the smaller subsets (10 jokes per type), some minor differences can be seen in the ordering of performance for some model/joke type combinations. The same is provided in Figure 5.8 and Figure 5.9 for our use of Qwen 72B as the LLM judge.

CHAPTER 5. COMPARING APPLES TO ORANGES

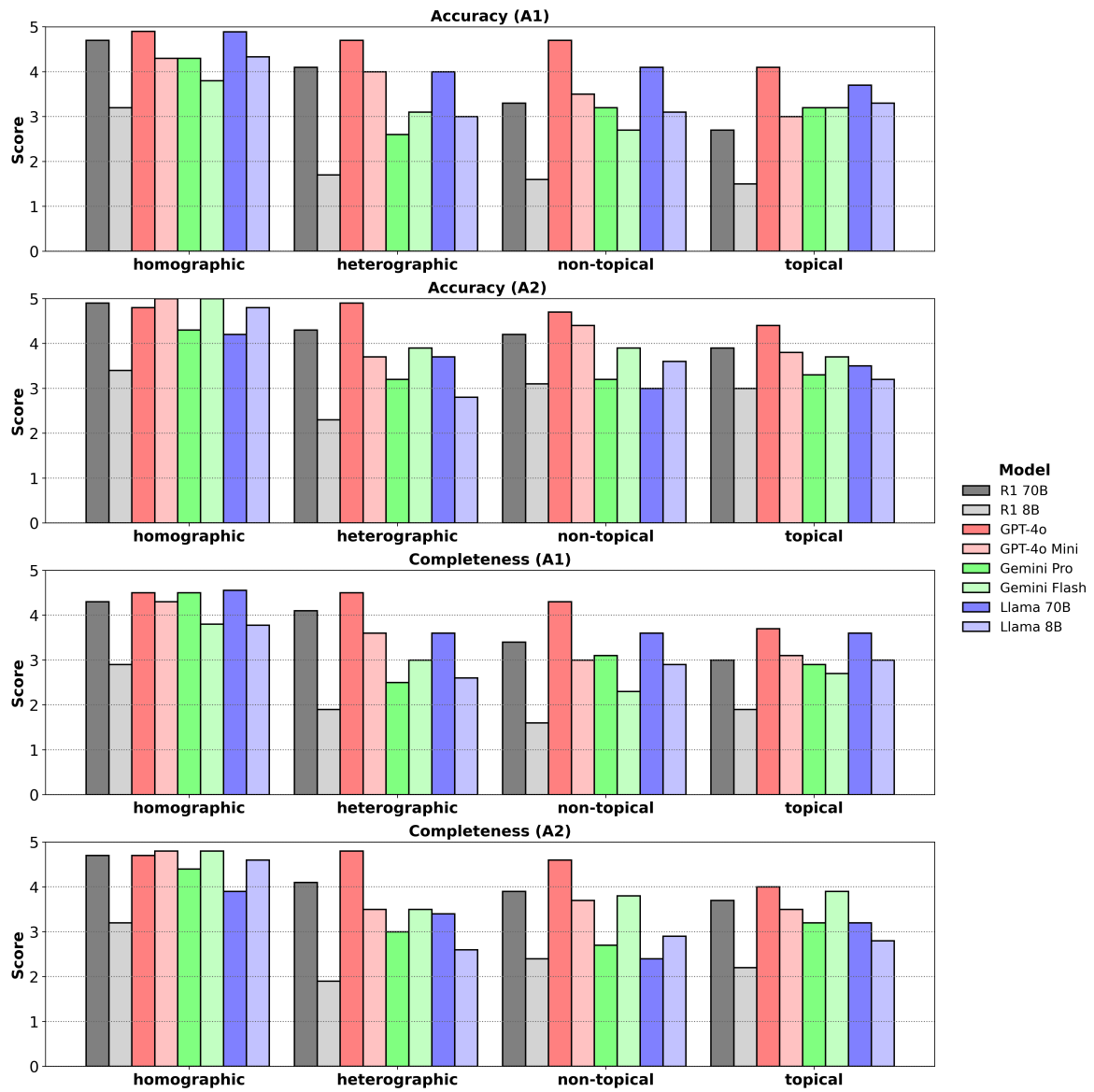


Figure 5.6: Average accuracy and completeness scores as assigned by third-party evaluators, by model and joke type.

CHAPTER 5. COMPARING APPLES TO ORANGES

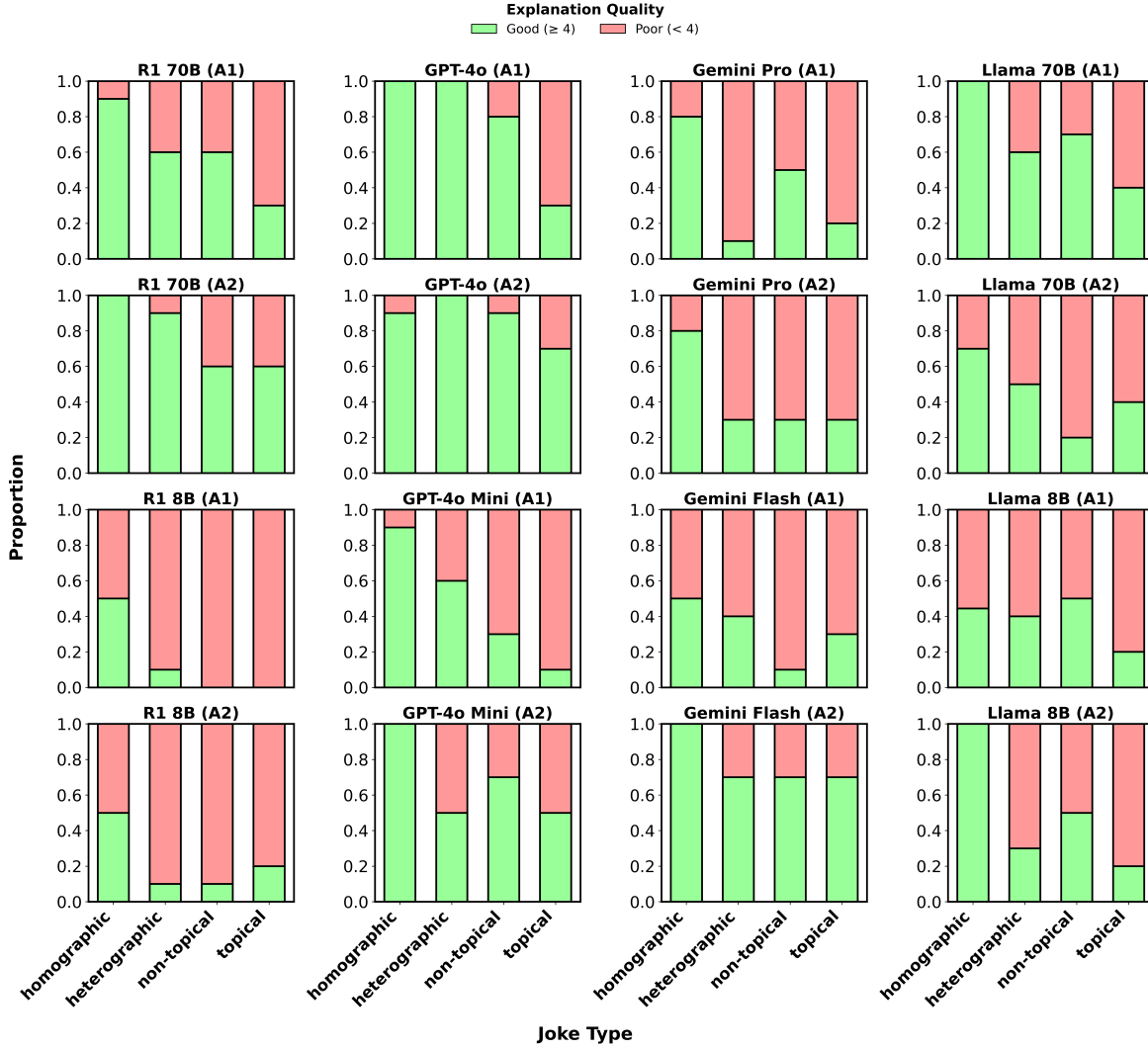


Figure 5.7: Binary evaluation results on a subset of 320 explanations (10 jokes across four types, by eight models). A1 and A2 refer to two different annotators.

CHAPTER 5. COMPARING APPLES TO ORANGES

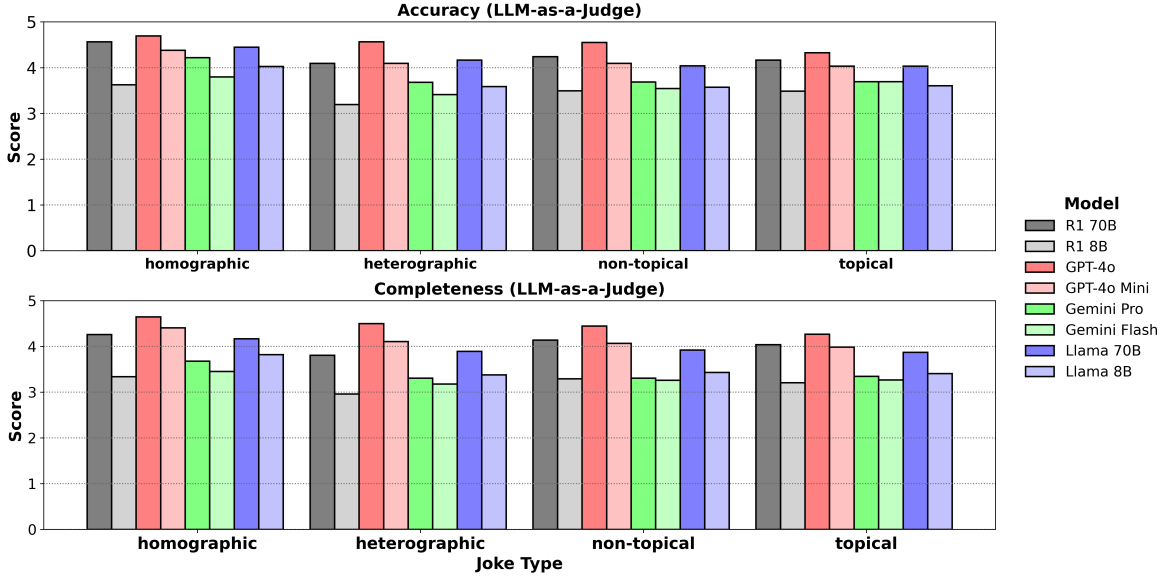


Figure 5.8: LLM-as-a-judge results on accuracy and completeness.

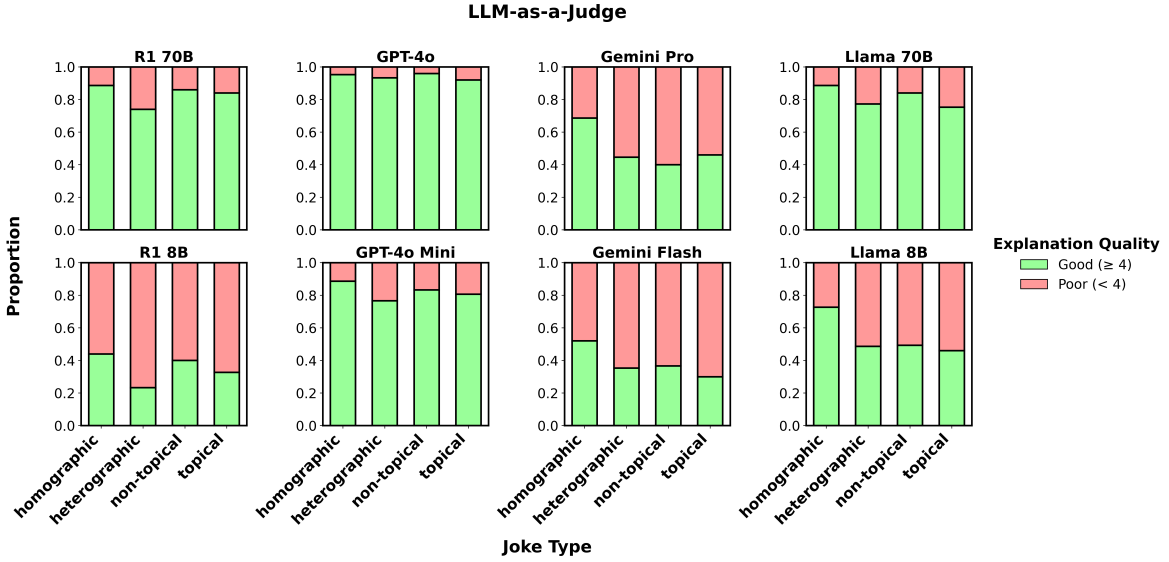


Figure 5.9: Proportion of “good” and “poor” joke explanations as determined by an LLM judge.

5.12.6 Additional Automatic Evaluation

Full results of automatic evaluation can be seen in Table 5.4.

CHAPTER 5. COMPARING APPLES TO ORANGES

Model	Metric	Homographic	Heterographic	Non-Topical	Topical
GPT-4o	SacreBLEU	8.51	10.15	7.85	7.09
	ROUGE-1	0.41	0.43	0.41	0.40
	ROUGE-2	0.12	0.15	0.13	0.12
	ROUGE-L	0.25	0.28	0.25	0.23
	METEOR	0.37	0.39	0.32	0.32
	BERTScore	0.88	0.89	0.88	0.87
GPT-4o Mini	SacreBLEU	7.06	9.27	6.12	5.30
	ROUGE-1	0.38	0.42	0.38	0.37
	ROUGE-2	0.10	0.14	0.11	0.10
	ROUGE-L	0.23	0.27	0.23	0.22
	METEOR	0.33	0.37	0.30	0.30
	BERTScore	0.88	0.89	0.87	0.87
Gemini Pro	SacreBLEU	7.02	8.15	4.71	4.54
	ROUGE-1	0.40	0.41	0.33	0.35
	ROUGE-2	0.10	0.13	0.09	0.09
	ROUGE-L	0.25	0.28	0.22	0.21
	METEOR	0.30	0.32	0.22	0.24
	BERTScore	0.88	0.88	0.87	0.87
Gemini Flash	SacreBLEU	8.29	7.09	5.15	4.93
	ROUGE-1	0.41	0.38	0.36	0.35
	ROUGE-2	0.13	0.10	0.10	0.09
	ROUGE-L	0.28	0.24	0.23	0.21
	METEOR	0.31	0.28	0.24	0.24
	BERTScore	0.89	0.88	0.87	0.87
Llama 70B	SacreBLEU	9.13	10.60	7.95	7.27
	ROUGE-1	0.42	0.43	0.41	0.39
	ROUGE-2	0.13	0.14	0.14	0.12
	ROUGE-L	0.25	0.27	0.26	0.23
	METEOR	0.36	0.37	0.31	0.30
	BERTScore	0.88	0.89	0.88	0.87
Llama 8B	SacreBLEU	8.41	9.76	7.16	5.83
	ROUGE-1	0.39	0.42	0.40	0.36
	ROUGE-2	0.12	0.14	0.12	0.10
	ROUGE-L	0.24	0.27	0.24	0.22
	METEOR	0.34	0.36	0.30	0.28
	BERTScore	0.88	0.89	0.87	0.87
R1 70B	SacreBLEU	7.73	8.47	6.27	5.89
	ROUGE-1	0.39	0.41	0.37	0.36
	ROUGE-2	0.10	0.12	0.11	0.10
	ROUGE-L	0.22	0.24	0.23	0.21
	METEOR	0.33	0.34	0.28	0.27
	BERTScore	0.88	0.88	0.87	0.87
R1 8B	SacreBLEU	5.25	6.48	4.85	4.12
	ROUGE-1	0.35	0.37	0.36	0.33
	ROUGE-2	0.08	0.10	0.09	0.08
	ROUGE-L	0.21	0.23	0.22	0.20
	METEOR	0.25	0.28	0.24	0.23
	BERTScore	0.87	0.88	0.87	0.86

Table 5.4: Automatic evaluation results by joke type. Human-authored, gold-standard explanations act as the reference.

Part V

Multimodal Understanding

Chapter 6

With Ears to See and Eyes to Hear: Sound Symbolism Experiments with Multimodal Large Language Models

Proceedings of EMNLP ([Loakman et al., 2024](#))

Tyler Loakman^a, Yucheng Li^b, and Chenghua Lin^c

^aDepartment of Computer Science, University of Sheffield, UK

^bDepartment of Computer Science, University of Surrey, UK

^cDepartment of Computer Science, The University of Manchester, UK

Contextual Statement

Transitioning away from the humour-based focus of previous chapters, this chapter addresses RQ5, investigating the extent to which large language models (LLMs) and vision language models (VLMs) are able to reflect human psycholinguistic associations through sound symbolism. Specifically, this chapter introduces tasks for assessing the presence of shape and magnitude symbolism, using VLMs to associate phonetically motivated pseudowords with images of entities with different physical characteristics, in addition to assessing the linguistic iconicity of words with LLMs. Rather than directly assessing phonetic or phonological knowledge, this work investigates the abstract non-arbitrary relationship between the phonetic realisations of words and the physical characteristics of what they describe, by assessing if such associations are learnt by VLMs despite not having access to audio. The findings of this chapter, therefore, have implications for in silico experimentation, using VLMs and LLMs in place of human participants for pilot studies regarding psycholinguistic phenomena. The transition towards assessing multimodal understanding using VLMs also reflected the growing prevalence of such models in NLP research at the time of publication, with OpenAI having recently released the first mainstream VLM, GPT-4 Vision.

Abstract

Large language models (LLMs) and vision language models (VLMs) have demonstrated aptitude as potential substitutes for human participants in experiments testing psycholinguistic phenomena. However, an understudied question is to what extent models that only have access to vision and text modalities are able to implicitly understand sound-based phenomena via abstract reasoning from orthography and imagery alone. We investigate the ability of LLMs and VLMs to demonstrate sound symbolism (i.e., recognise a non-arbitrary link between sounds and concepts) as well as their ability to “hear” via the interplay of the language and vision modules of open and closed-source multimodal models. We perform multiple experiments, including replicating the classic Kiki-Bouba and Mil-Mal shape and magnitude symbolism tests and comparing human judgements of linguistic iconicity with those of LLMs. Our results show that VLMs demonstrate varying levels of agreement with human labels, with more task information being required for *in silico* experimentation with VLMs than their human counterparts. We additionally see that magnitude symbolism is an easier pattern for VLMs to identify than shape symbolism, and that an understanding of linguistic iconicity is dependent on model size.

6.1 Introduction

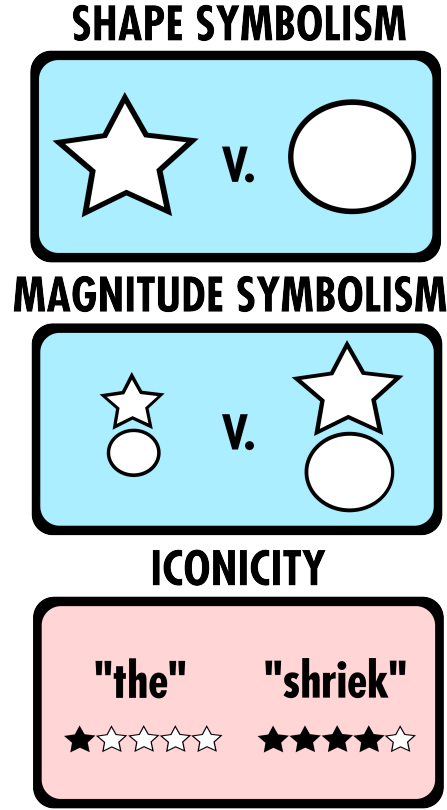


Figure 6.1: *Illustration of the three main experiments we perform: sound symbolism, magnitude symbolism, and linguistic iconicity.*

Sound symbolism refers to the perceived similarity between speech sounds and the conceptual meanings of the words they comprise. Evidence of this phenomenon can be found in linguistic devices such as onomatopoeia (e.g., “bang”, “shriek”, and “bellow”), where a word imitates the concept it describes via its phonetic form. The ability of large language models (LLMs) and vision language models (VLMs) to reflect a sense of sound symbolism would therefore suggest that these models are capable of acquiring phonetic knowledge indirectly through only the written orthographic form of a language through grapheme patterns (Loakman et al., 2025b, 2023b) and meta-level textual discussion of sound in training data, which has implications for the potential use of LLMa and

CHAPTER 6. WITH EARS TO SEE AND EYES TO HEAR

VLMs in perceptual studies usually reserved for humans (Jain et al., 2023; Dillion et al., 2023; Aher et al., 2023). In this work, we further explore the capability of such models to demonstrate human-like characteristics in a range of psycholinguistic perceptual tests investigating three main areas of sound symbolism: (i) *Shape symbolism* (i.e., the Kiki-Bouba effect, Ramachandran and Hubbard, 2001), where a forced choice must be made between two pseudowords as to which is the most appropriate to describe shapes and entities that are spiky or rounded; (ii) *magnitude symbolism* (i.e., the Mil-Mal Effect, Sapir, 1929), a similar test to (i), but where the entities are small or large rather than spiky or rounded; and (iii) *iconicity rating* (Winter et al., 2023), where LLMs are asked to rate a series of English words on their perceived “iconicity” (i.e., to what extent a word’s form is perceived to be analogous to the concept it describes). We illustrate these experiments in Figure 6.1.

In extending these experiments to LLMs and VLMs, our study aims to shed light on the processes underlying multimodal perception in language models.¹ Moreover, the presence of sound symbolism in such models could inform the development of more effective natural language processing algorithms, aiding tasks such as sentiment analysis, emotion recognition, and content generation that take into account more abstract layers of human reasoning and perception than semantics alone (Manzoor et al., 2023). An understanding of sound symbolism also has the potential to have a profound effect on creative generation, including language forms such as poetry and narratives, as well as their accompanying illustrations. Additionally, sound symbolism is a prevalent strategy used in marketing to create desirable associations in potential customers, allowing LLMs to be used in pilot testing (Ketron and Spears, 2021; Motoki et al., 2020; Spence, 2012).

¹The chapter title is inspired by the Sleeping With Sirens album of the same name: https://en.wikipedia.org/wiki/With_Ears_to_See_and_Eyes_to_Hear.

We summarise our main contributions as follows:

- We replicate the classic psycholinguistic Kiki-Bouba shape symbolism and Mil-Mal magnitude symbolism experiments with a range of open- and closed-source VLMs to investigate the understanding of the association between word forms and the characteristics of entities.
- We perform an in-depth analysis of the ability of a range of open- and closed-source LLMs to understand linguistic iconicity, comparing model judgements to an existing large-scale dataset of human ratings.
- We discuss the potential sources of sound symbolism abilities in LLMs and VLMs, as well as potential future approaches to bolstering these abilities, in addition to the implications of doing so.

6.2 Related Works

The work of early linguists, such as Saussure, addressed whether the relationship between the “sign” (i.e., a word) and the “signified” (i.e., the entity/concept to which the sign refers) is arbitrary, with there being nothing more “boat”-esque about the word “boat” than any other combination of phonotactically legal sounds (de Saussure, 1916). However, there are many types of language where this association is seen to be non-arbitrary, such as the onomatopoeia commonly used in literary works (e.g., “bang” for a loud noise or “shriek” for a high-pitched wail), where the phonetic realisation mirrors the concept it describes. These phenomena as a whole are known as sound symbolism, where there is thought to be a non-arbitrary link between the sign and the signified, in contrast with the popular stance of early linguistics.

CHAPTER 6. WITH EARS TO SEE AND EYES TO HEAR

Outside of onomatopoeia, sound symbolism is believed to have a range of effects on human perception, including applying to nonsense pseudowords. For example, even if a word was not created to explicitly denote a known concept or entity and therefore has no true denotation, it is nevertheless able to manifest a connotative meaning based on its phonological representation and phonetic realisation. The first identification of these patterns is frequently attributed to [Usnadze \(1924\)](#). In this work, 10 participants were given a series of pseudowords alongside drawings and asked to pair them. The results showed a higher-than-chance agreement between evaluators regarding which nonsense word best described each drawing. Perhaps the most famous example of sound symbolism is the Kiki-Bouba effect ([Sidhu et al., 2021](#); [Ramachandran and Hubbard, 2001](#); [Köhler, 1929](#)) which concerns the allocation of the name “Kiki” to sharp, hard-edged entities, and “Bouba” to softer, rounded entities. A similar relationship has been observed between the words “Mil” and “Mal”, where the changing vowel in the phonological minimal pairs² has a relationship to perceived size, in a phenomenon known as magnitude symbolism. This is where vowels with higher frequency content are associated with smaller entities due to the relationship between vocal tract length and vocal productions ([Sapir, 1929](#)). Extensive research has been performed in the area of sound symbolism, demonstrating that the phenomena are mostly language-agnostic ([Ćwiek et al., 2022](#)), weaker in neurodivergent individuals ([Ocelli et al., 2013](#)), and not developed until later childhood ([Sidhu et al., 2023](#)). Other research has also investigated the exact requirements and limits of the effects ([Sidhu and Vigliocco, 2023](#); [Passi and Arun, 2022](#); [Styles and Gawne, 2017](#); [Nielsen and Rendall, 2013](#)).

[Alper and Averbuch-Elor \(2023\)](#) investigated the ability of vision language models to exhibit the Kiki/Bouba effect using CLIP and Stable Diffusion to generate images from sound-symbolic prompts, providing evidence for this association. We build upon this

²A minimal pair refers to a pair of words that differ only in one phonological segment, such as “cat” /kat/ and “bat” /bat/.

work in §6.3, using a forced-naming task for “real” entities rather than abstract shapes, and additionally extending the investigation to magnitude symbolism with the Mil-Mal test in §6.4. We further differentiate our work by focusing on assigning pseudowords to provided visual stimuli, in contrast to [Alper and Averbuch-Elor \(2023\)](#) who investigate the effects of different pseudowords on the outputs of image generation models.

Additionally, large-scale efforts have been made to collect ratings of linguistic iconicity (i.e., the level of symbolism within a particular word), with [Winter et al. \(2023\)](#) collecting ratings of over 14k English words. Efforts have been made to analyse whether similar ratings would be assigned by an LLM, with [Trott \(2024\)](#) using GPT-4, reporting a moderate positive correlation between human and model ratings. We build upon this work in §6.5 by introducing a wider range of VLMs, including open-source alternatives. Numerous works have investigated other aspects of iconicity and sound symbolism, with [Abramova and Fernández \(2016\)](#) investigating the word embeddings of different aspects of morphology in relation to symbolism (see also [Yamshchikov et al., 2019](#); [Liu et al., 2018](#)). Additionally, [Sabbatino et al. \(2022\)](#) investigated the emotional intensity of nonsense words by analysing the phoneme combinations most responsible for emotional affect. Several works have also exemplified the ability of LLMs to demonstrate perceptual behaviour akin to humans and the potential for these models to replace participants in-silico, allowing evaluation to scale ([Jain et al., 2023](#); [Aher et al., 2023](#); [Dillion et al., 2023](#); [Ramezani and Xu, 2023](#); [Coda-Forno et al., 2023](#)).

6.3 Shape Symbolism



Figure 6.2: Examples of Kiki-style (*i.e.*, spiky) and Bouba-style (*i.e.*, rounded) generations from DALL-E 3. In total, 50 images were generated (25 per condition). Ground truth is taken as the human majority vote, rather than the generation intent.

In this section, we perform a replication of the classic Kiki-Bouba experiment ([Ramachandran and Hubbard, 2001](#)) using a range of vision language models. In the classic setup, human participants are presented with either a soft, rounded shape or a spiky, sharp-edged shape and asked to assign one of two pseudowords. Typically, words such as “Bouba” and “Maluma” are preferred for the rounded shapes, whilst “Kiki” and “Takete” are preferred for the spiked shapes ([Sidhu et al., 2021](#); [Ramachandran and Hubbard, 2001](#); [Köhler, 1929](#)). These findings demonstrate a non-arbitrary link between speech sounds and the physical characteristics of the entities to which they refer.

6.3.1 Methodology

Image Dataset. We use DALL-E 3 (Betker et al., 2023) to generate images of entities that are either “spiky” or “rounded”.³ Example generations are shown in Figure 6.2. We generate novel imagery rather than taking existing images, such as the classic 2D representation of Kiki/Bouba, in order to reduce the effects of memorisation and exert finer control of the physical characteristics of the entities. Furthermore, this allows further investigation into the extent to which human perception of this phenomenon extends from geometric shapes to entities in the “real” world, therefore increasing ecological validity and more closely representing how models may be tasked with demonstrating sound symbolism in production, such as naming products for marketing or new characters in a narrative. We use the prompt “Generate an image from the following description: [spiky/rounded + noun]”, where we generate 25 spiky examples, and 25 round examples, with each noun being used once per shape condition. The full list of entities is presented in Appendix 6.10.1.

Pseudowords. Akin to the need to generate novel imagery to overcome memorisation, it is also important to avoid bias from the eponymous pseudowords (i.e., “Kiki” or “Takete” for spiked entities, and “Bouba” and “Maluma” for rounded entities). As a result, we consult existing sound symbolism literature and select the following pseudowords that are legal in English phonotactics and imitate the same phonetic relationship that the original terms were designed to elicit. Borrowing from Ocelli et al. (2013), we adopt: Kalika/Mabobe, Zaki/Umbu, and Tiki/Giba. From Alper and Averbuch-Elor (2023) we additionally take Kitaki/Gugagu, Hatiha/Bodubo and Penape/Gunogu. Finally, we use the original Kiki/Bouba names as a point of reference

³Specifically, we access `dall-e-3` via the OpenAI API in November 2023, generating images with a resolution of 1024*1024 pixels.

for a best-case scenario where the link is explicitly learnt from mentions within the training data.

Task. We imitate the standard setup for the Kiki/Bouba experiment and present VLMs with the following zero-shot prompt:

“Look at the [ENTITY] in the provided image. Out of the following two options, which name would you most likely assign to the [ENTITY]: “[KIKI-WORD]” or “[BOUBA-WORD]”. Respond with only your decision”.

One candidate from either name category (i.e., Kiki or Bouba) is presented as outlined previously, and [ENTITY] refers to a noun that describes the entity, in order to direct the model’s attention to the correct element (i.e., the noun from the DALL-E 3 prompt). For this experiment, we set *max_tokens* to 10 for the VLM responses and leave all other hyperparameters at default. We additionally provide an extended *informed* prompt, which prepends:

“This task is related to the phenomenon of Sound Symbolism, which is a non-arbitrary relationship between the sound of a word and associations with its physical attributes”

to give the models additional task knowledge. This secondary prompt scenario is used to investigate whether human-like preferences can be encouraged from the model with additional awareness of which elements to focus on, without explicitly stating the relationship between particular sounds and characteristics. We present each prompt twice, placing each pseudoword in first or second position and averaging the results to mitigate positional bias.

Models. We use a selection of open- and closed-source VLMs, including Gemini Pro (referred to as Gemini) (Georgiev et al., 2024), GPT-4 Vision (referred to as GPT-4)

(OpenAI et al., 2023),⁴ and LLaVA (Liu et al., 2024a). To investigate the effects of model size, we include the 7B, 13B, and 34B parameter versions of LLaVA. Further details are presented in Appendix 6.10.1.

Human Labels. As the human point of comparison, we recruited ten evaluators with native-level English proficiency through internal communication (i.e., email lists at the primary author’s institution and word-of-mouth) who were presented with a task analogous to that which we present to the VLMs.⁵ Images are presented in a random order to participants for each pseudoword pair to avoid order effects.

6.3.2 Results

In Figure 6.3 we see mixed results as to which model performs best, with GPT-4 demonstrating the highest levels of agreement for the original Kiki/Bouba and the added Kalika/Mabobe, Gemini performing the best for Zaki/Umbu and Hatiha/Bodubo, and LLaVA performing the best across the remaining conditions. However, across all models, there is a general trend of low agreement with human ratings, with only a few condition-model combinations resulting in agreement above chance (50%). Regarding the introduction of the informed prompt containing additional task information, there is either no change or a moderate increase in agreement. This indicates that once models are aware of the characteristic of interest (i.e., shape), they are more likely to agree with human perception. However, there are select cases (e.g., Zaki/Umbu with LLaVA 7B and 13B, Kitaki/Gugagu with GPT-4, and Penape/Gugagu with Gemini, GPT-4, and LLaVA 13B) where performance decreases in the informed condition. However, these decreases are minor except where agreement was already very low. We observe

⁴This work was performed before the release of GPT-4o.

⁵Evaluators were recruited in different waves following revisions. All evaluators were paid above the 2024 UK Living Wage per hour.

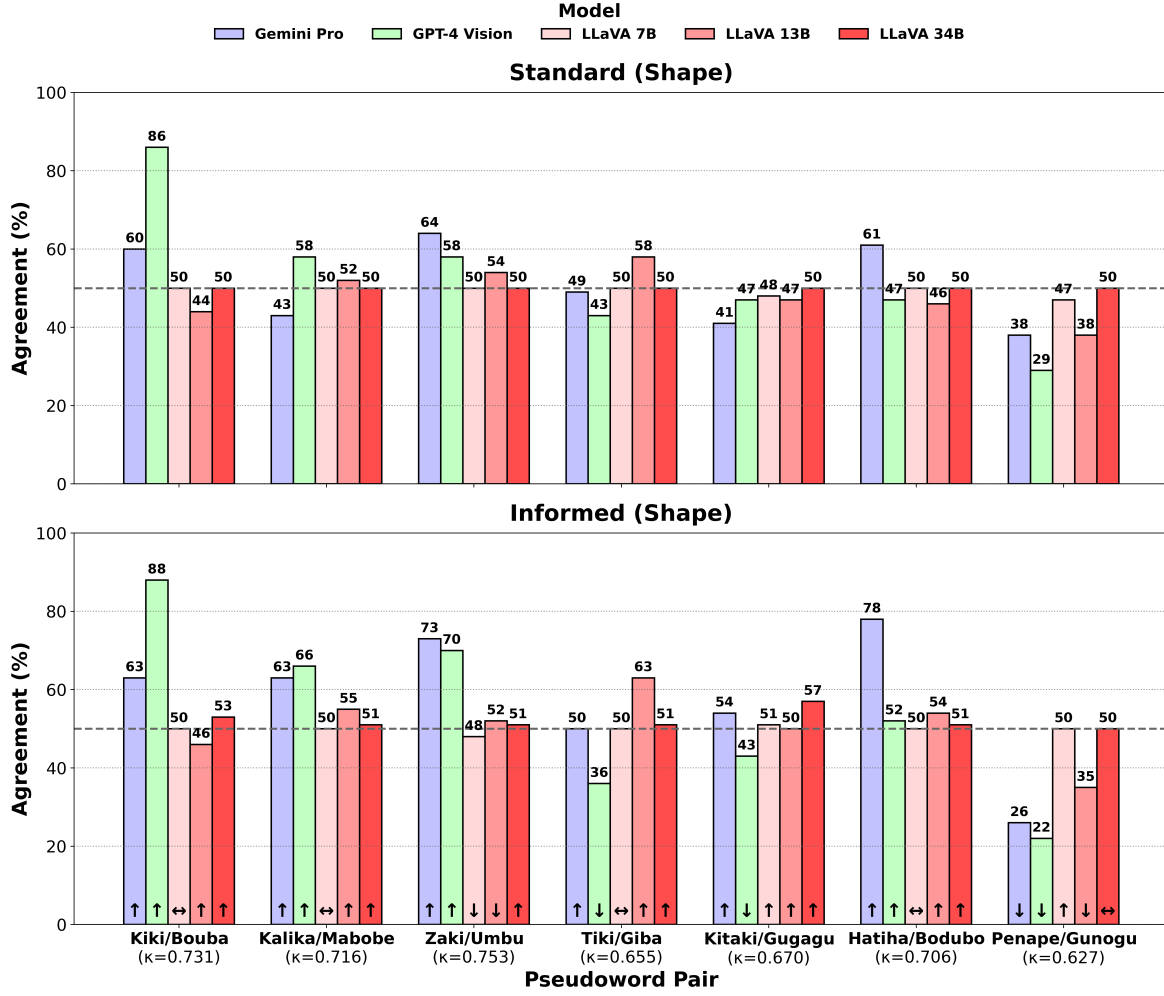


Figure 6.3: Results of the shape symbolism experiments per pseudoword pair. Fleiss’ κ (Fleiss, 1971) is presented next to each pseudoword pair for inter-annotator agreement between humans. Arrows indicate the direction of agreement change from the standard to the informed prompt. The dashed line represents chance-level agreement. Models are compared to the majority human vote.

varying performance across open and closed-source models. Whilst LLaVA outperforms GPT-4 and Gemini in many conditions, this is often close to chance-level agreement and may be the result of averaged label bias, in contrast to the closed models’ systematic disagreement.⁶ Finally, regarding the open-source LLaVA models, the 13B model outperforms the 7B and 34B models at times, most noticeably in Tiki/Giba. However,

⁶We see LLaVA variants demonstrate a clear preference for whatever pseudoword is presented first.

the largest 34B model performs best overall, in line with expectations (though only by a small margin). Additionally, the Penape/Gunogu pair presents difficulty for the tested models, with strong disagreement with humans demonstrated by Gemini and GPT-4, suggesting the potential influence of additional information contained within the models' training data as to the connotations of these pseudowords.

6.4 Magnitude Symbolism



Figure 6.4: *Examples of Mil-style (i.e., tiny) and Mal-style (i.e., huge) generations from DALL-E 3. In total, 50 images were generated (25 per condition). Ground truth is taken as the human majority vote, rather than the generation intent.*

Whilst the Kiki-Bouba effect demonstrates sound symbolism in relation to the perceived shape of an object, magnitude symbolism refers to the non-arbitrary relationship between a word's vowel quality and the perceived size of the entity to which it refers.

This is commonly demonstrated using the pseudowords “Mil” and “Mal”, where the high front vowel /ɪ/ in “Mil” is associated with small entities, and the low back vowel /a/ of “Mal” is associated with larger entities.

6.4.1 Methodology

Image Dataset. We follow the same process as §6.3.1, but use the characteristics of “tiny” and “huge” with the following prompt: “Generate an image of a/an [ENTITY] in isolation, with something else to help judge scale/perspective”. Example generations are presented in Figure 6.4. We verified that entity size in all images was perceived as intended (e.g., that the fish in the top left of Figure 6.4 is seen as small, rather than the coin seen as large).

Pseudowords. As in the shape symbolism experiment (§6.3), we wish to mitigate potential bias arising from memorisation. To this end, we use the “Mil” and “Mal” often associated with this test (Sapir, 1929) in addition to other non-eponymous phonetically similar pseudowords. Additionally, to avoid gross extraneous factors arising from using words that are meaningful in English (such as “mil” referring to millilitres, and “mal” being associated with badness, such as malpractice), we create the minimal pairs Dil/Dal, Zil/Zal, Geel/Gaal, Beel/Baal, Weel/Waal, and Leel/Laal. The former three exploit the contrast between /ɪ/ and /a/, whilst the latter exploit /i/ and /ɑ/, with a range of consonants for variation.

Task. We use the same setup as in §6.3.1 but present the models with one candidate from either magnitude-based name category (i.e., Mil-esque or Mal-esque pseudowords), rather than Kiki/Bouba-related pseudowords. We additionally provide an extended prompt we call *informed*, which prepends “*This task is related to the phenomenon of*

Magnitude Symbolism, which is a non-arbitrary relationship between the sound of a word and its association with size and scale” to give additional task knowledge. We use the same models and evaluation protocol as in §6.3.1.

6.4.2 Results

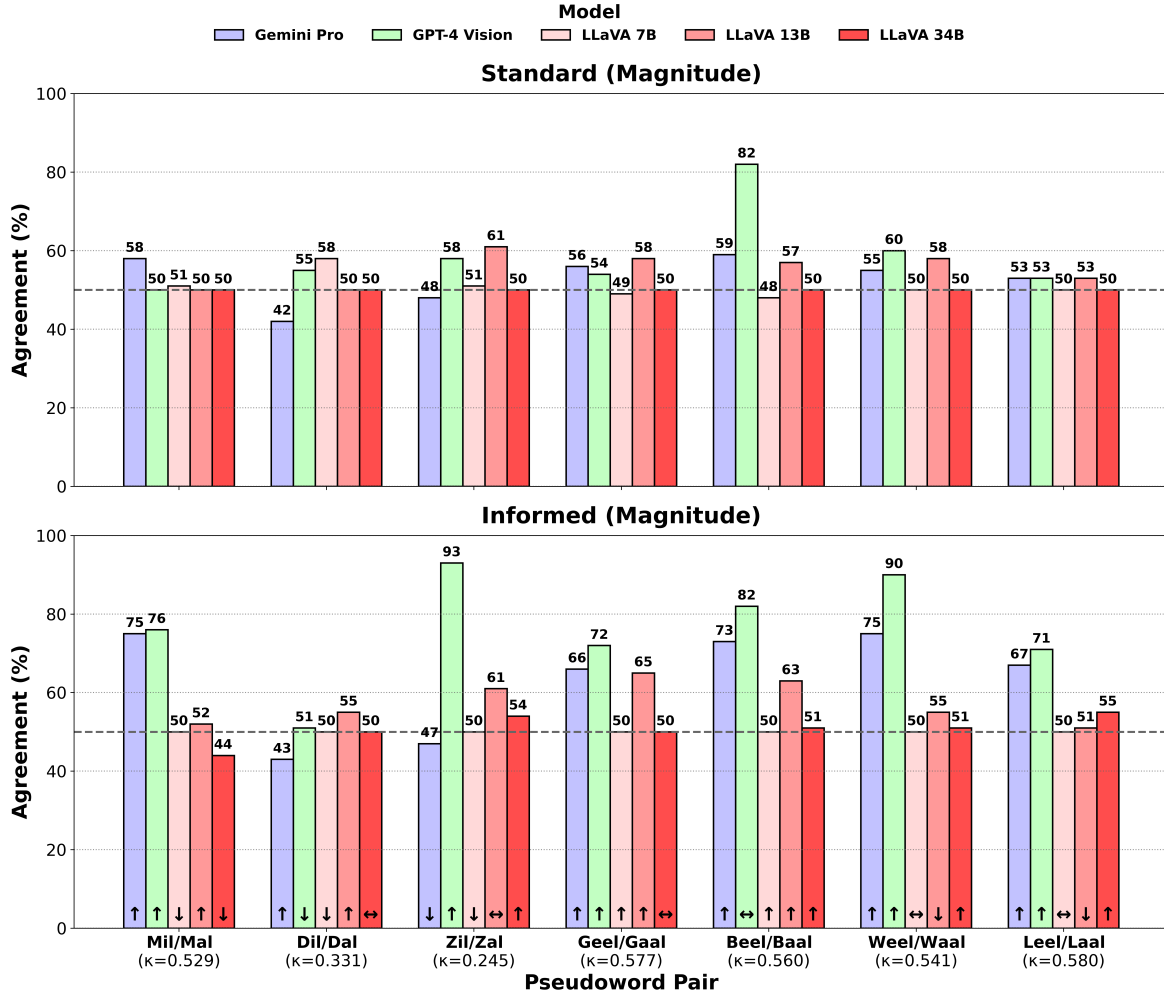


Figure 6.5: Results of the magnitude symbolism experiments per pseudoword pair. Fleiss’ κ (Fleiss, 1971) is presented beside each pseudoword pair, demonstrating inter-annotator agreement between humans. Arrows indicate the direction of agreement change from the standard to the informed prompt. The dashed line represents chance-level agreement. Models are compared to the majority human vote.

Overall, similar to §6.3, we see mixed results in Figure 6.5. However, GPT-4 demonstrates higher levels of agreement across most conditions in a clearer pattern than what was seen regarding shape symbolism. In several cases, model selections were aligned with human perception, with 90+% agreement (e.g., GPT-4 in the Zil/Zal and Weel/Waal conditions). When comparing the standard and informed prompts, we see a more substantial increase in performance, indicating that the VLMs fundamentally understand the relationship between sound and perceived size, but were focused on other aspects of the provided imagery when not explicitly directed towards size in the standard condition. Regarding the LLaVA models, we see the mid-sized 13B variant outperforming the 7B and 34B models in most conditions (rather than performance increasing alongside parameter count).

6.5 Iconicity Ratings

In this section, we investigate whether LLMs demonstrate human-like associations between word forms and the entities/concepts they symbolise.⁷ Winter et al. (2023) present a dataset of 14k+ human judgements of iconicity to which we compare our LLM ratings (where, on a 7-point scale, “how” scored 1.3, whilst “woof” scored 6.8 due to being onomatopoeic).

6.5.1 Methodology

Models. We use a range of modern LLMs for this task. Specifically, GPT-4 (OpenAI et al., 2023), GPT-3.5-Turbo (OpenAI et al., 2023), Llama 2 (7B, 13B, and 70B) (Touvron et al., 2023), Flan-T5 (250M and 3B) (Raffel et al., 2020; Chung et al., 2024a), and Mistral Instruct (7B) (Jiang et al., 2023). Implementation details are presented

⁷We treat this as VLMs imitating an understanding of sound symbolism, as they are trained on graphemes.

in Appendix 6.10.1. Existing work by [Trott \(2024\)](#) investigated whether GPT-4 is able to reflect human judgements in this task, reporting a Spearman correlation of 0.59. However, in trying to reproduce these findings with a newer version of GPT-4, we observe different performance. For this reason, we re-run this experiment on the first 7k entries of [Winter et al. \(2023\)](#) and compare GPT-4 ratings as of late November 2023 with the human and GPT-4 judgements in [Winter et al. \(2023\)](#).⁸ In all other cases, we use the entire 14,772-entry dataset, except for GPT-3.5-Turbo, where we remove 209 words that triggered OpenAI’s safeguarding filters.

Prompting. We adopt the same prompting strategy as [Trott \(2024\)](#), using a modified version of the instructions presented to human participants in [Winter et al. \(2023\)](#).⁹ In summary, we request ratings of iconicity on a 1-7 scale, where 1 is not at all iconic, and 7 is highly iconic. The full prompt presented to the models is available in Appendix 6.10.2.

6.5.2 Results

Correlations between models and humans are presented in Table 6.1. We observe that the level of agreement between models and humans on the linguistic iconicity of English words is dependent on model size. For instance, there is little correlation between human scores and responses from either variant of Flan-T5 or the smallest Llama 2 variant (7B). However, significant correlations are observed with another 7B parameter model, instruction-tuned Mistral. When investigating the larger models, the 13B variant of Llama 2 demonstrates Spearman/Pearson correlations of .379 and .381, respectively. Interestingly, however, the largest Llama 2 model, with 70B parameters, demonstrates

⁸We test only on the first 7k entries due to API expenses.

⁹This involved the removal of instructions that do not apply to model-based reproductions, such as elements of the specific setting and presentation method.

Model	Spearman		Pearson	
	ρ	p	r	p
Flan-T5 250M	-.035	<.001	-.037	<.001
Flan-T5 3B	.000	.991	-.002	.824
Mistral Instruct 7B	.382	<.001	.377	<.001
Llama 2 7B	-.003	.687	.005	.544
Llama 2 13B	.379	<.001	.381	<.001
Llama 2 70B	.304	<.001	.332	<.001
GPT-3.5-Turbo	.420	<.001	.439	<.001
GPT-4 (Trott, 2024)	.575	<.001	.615	<.001
GPT-4 (Ours, Nov. '23)	.537	<.001	.594	<.001

Table 6.1: *Correlations between human ratings from Winter et al. (2023) and tested models. Trott (2024) report a Spearman correlation of .590 across the entire dataset .*

lower agreement than the 13B variant. Regarding models from OpenAI, GPT-3.5-Turbo demonstrates moderate Spearman and Pearson correlations with humans at .420/.439, respectively. Finally, GPT-4 presents the strongest correlations. However, the version of GPT-4 used here (late November 2023) performs worse than the earlier version used in Trott (2024), demonstrating how continuously updated models also require continuous evaluation.

6.6 Discussion

6.6.1 Sound Symbolism Source

Across experiments, we observe that LLMs and VLMs are capable of making similar decisions to humans on sound symbolism tasks. This occurs whilst only having access to textual and visual modalities, whilst human decisions are primarily grounded in sound. We hypothesise several reasons for the emergence of sound symbolism in such models.

CHAPTER 6. WITH EARS TO SEE AND EYES TO HEAR

Firstly, due to most human languages exhibiting mostly regular, transparent orthography, auditory information in speech is reflected in orthography via grapheme sequences - a characteristic that grapheme-to-phoneme conversion models have long exploited. Through this, text-based models are able to learn associations between grapheme sequences and semantics, based on more abstract characteristics than morpheme combinations alone, such as phonaesthemes (Kaushal and Mahowald, 2022). Whilst such models have no embodied understanding of sound, such statistical patterns pose a viable signal for the implicit learning of sound-based phenomena.

Secondly, such associations between sounds (or grapheme combinations) and physical characteristics are naturally present in language, such as in poetry, narratives, or descriptions of entities that are cute, scary, small, or large. Such characteristics are therefore paired with relevant visual stimuli in image captions when training vision modules for multimodal systems. However, such associations are subtle and not entirely ubiquitous, which contributes to the inconsistent performance demonstrated by the tested models. For example, whilst the high front vowel /i/ is typically associated with small entities such as “mini”, the word “small” itself possesses a low back vowel /ɔ:/.

As a result, the relatively weak performance of our tested models could also be explained by the relative lack of sound-symbolism-heavy language in the models’ training data, which is overshadowed by prosaic language forms that do not exploit these phenomena as readily. This, in turn, would explain why the closed-source models we tested (i.e., GPT-4 and Gemini) outperform open-source models, due to significant differences in parameter count, allowing the larger closed models to retain more information within weights, in addition to being continuously updated with RLHF.

The results of the multimodal experiments demonstrate that under certain conditions, VLMs show systematic disagreement with human labels, indicating the influence of additional pre-training knowledge that affects the associations made between pseudowords

and images. However, in the presented experiments, we compare model selections against the majority vote or mean scores assigned by humans. Consequently, this results in a comparison to an “ideal” reference, overlooking individual differences in perception. It may be considered that for a decision to be “human-like”, it must match any human choice, rather than the majority vote.

6.6.2 Future Directions

As a result of the tested LLMs and VLMs not fully reflecting human preferences, it remains promising to explicitly pre-train language models on sound-symbolic datasets or explicitly include related tasks in pre-training. Additionally, investigating reasons behind model predictions is a promising direction, such as visual attention, addressing whether models are attending to characteristics closely associated with the concepts being tested (e.g., spikiness).

6.7 Conclusion

This work explored the processes underlying multimodal perception and understanding in language models. To do so, it presented a series of experiments regarding models’ understanding of sound symbolism. By comparing with human judgments, we see that VLMs can approximate human perception in sound symbolism tasks under certain conditions, such as when informed of the phenomenon’s concept, yet struggle overall. Additionally, we see that magnitude symbolism presents an easier pattern for VLMs to recognise than shape symbolism, with higher levels of agreement between models and humans. Furthermore, the ability of LLMs to emulate human judgements of iconicity is shown to scale with model size. These findings indicate room for future research on

explicitly encoding abstract perceptual properties into language models to facilitate better in silico experimentation and improve performance on downstream tasks.

6.8 Limitations

Whilst sound symbolism is believed to be largely language-agnostic, the presented experiments only used native English speakers and pseudowords that are phonotactically legal in English. Additionally, some of the chosen pseudowords were taken from existing literature. Whilst we investigated the prevalence of such words in the context of sound symbolism in internet resources, it remains possible that some level of data contamination may be present. Furthermore, we present only the orthographic forms of the pseudowords to human participants, resulting in potential variation regarding phonetic realisation. Statistical significance testing was omitted from cross-model evaluations due to the main contribution of this work being the identification of a general perceptual gap between language models and humans in sound symbolism associations. Whilst inter-model performance differences are discussed, the purpose of this work is primarily exploratory and to provide a novel dataset rather than to suggest a specific model for the processing of sound symbolic language or present a state-of-the-art architectural solution.

6.9 Ethics Statement

All image generations were provided via accessing the OpenAI API. In discovering imagery that triggered OpenAI’s built-in guardrails, we replaced these images with alternatives before finalising the test set. We present this work as a demonstration of interesting behaviours within LLMs/VLMs, but do not condone the wholesale

CHAPTER 6. WITH EARS TO SEE AND EYES TO HEAR

replacement of human participants in related psycholinguistic experimentation. Rather, we view in silico experimentation as a tool for prototyping, pilot studies, and additional insight.

6.10 Appendix

6.10.1 Implementation Details

Shape & Magnitude Symbolism. For the sound symbolism experiments, we access `gemini-pro-vision` via the Google Gemini API. For GPT-4, we use `gpt-4-vision-preview` via the OpenAI Chat Completions API. For the LLaVA models at various sizes, we specifically use `llava-hf/llava-v1.6-vicuna-7b-hf`, `llava-hf/LLaVA-v1.6-vicuna-13b-hf` and `llava-hf/LLaVA-v1.6-vicuna-34b-hf`. We use default hyperparameters for all models. Human participants were shown generated imagery with the binary pseudo-word options via Google Forms. The order of image presentation was randomised, whilst the order of the pseudowords was kept static. Separate forms were presented for each pseudoword pair. Participants completed the annotation at their own pace over approximately two weeks.

Iconicity Ratings. For the iconicity rating experiments, we use the following models from Hugging Face: `Flan-T5` (`google/flan-t5-base` and `google/flan-t5-xl`), `Mistral 7B Instruct` (`mistralai/Mistral-7B-Instruct-v0.2`), and `Llama 2` (`meta-llama/Llama-2-7b-hf`, `meta-llama/Llama-2-13b-hf`, and `meta-llama/Llama-2-70b-hf`). We access GPT-3.5-Turbo and GPT-4 via the OpenAI Chat Completions API. We keep all model hyperparameters at default settings. For the LLaVA models, we modify the prompt slightly by adding choice labels (A/B) rather than requesting the pseudoword itself to be returned in order to directly access single-token output probabilities.

Entities. We select the following list of entities for DALL-E-3 generations. including animate and inanimate entities: *alien, bed, bird, bottle, cat, chair, desk, dog, door, fish, flower, fruit, ghost, house, insect, lizard, machine, person, plane, planet, snake, toy, tree, vegetable, and vehicle.*

6.10.2 Iconicity Prompt

The full prompt provided to LLMs in the iconicity rating experiment is:

“Some English words sound like what they mean. These words are iconic. You might be able to guess the meaning of such a word even if you did not know English. Some words that people have rated high in iconicity are “screech,” “twirl,” and “ooze” because they sound very much like what they mean. Some words that people have rated moderate in iconicity are “porcupine,” “glowing,” and “steep,” because they sound somewhat like what they mean. Some words rated low in iconicity are “menu,” “amateur,” and “are,” because they do not sound at all like what they mean. In this task, you are going to rate words for how iconic they are. You will rate each word on a scale from 1 to 7. A rating of 1 indicates that the word is not at all iconic and does not at all sound like what it means. 7 indicates that the word is high in iconicity and sounds very much like what they mean. – It is important that you say the word out loud to yourself, and that you think about its meaning. If you are unsure of the meaning or the pronunciation of a word, you have the option of skipping it. – Try to focus on the word meaning of the whole word, rather than decomposing it into parts. For example, when rating ‘butterfly’ think of the insect rather than “butter” and “fly,” and rate how well the whole meaning relates to the sound of the whole word “butterfly.” – On a scale from 1 (not iconic at all) to 7 (very iconic), how iconic is the word ‘{word}’? Rating: ”

6.10.3 Full Results

Kiki/Bouba (κ .731)					
	Gemini Pro $\uparrow$	GPT-4 Vision $\uparrow$	LLaVA 7B	LLaVA 13B $\uparrow$	LLaVA 34B $\uparrow$
Standard	60% (κ .200)	86% (κ .720)	50% (κ .000)	44% (κ -.120)	50% (κ .000)
Informed	63% (κ .260)	88% (κ .760)	50% (κ .000)	46% (κ -.080)	53% (κ .060)
Kalika/Mabobe (κ .716)					
	Gemini Pro $\uparrow$	GPT-4 Vision $\uparrow$	LLaVA 7B	LLaVA 13B $\uparrow$	LLaVA 34B $\uparrow$
Standard	43% (κ -.140)	58% (κ .160)	50% (κ .000)	52% (κ .040)	50% (κ .000)
Informed	63% (κ .260)	66% (κ .320)	50% (κ .000)	55% (κ .100)	51% (κ .020)
Zaki/Umbu (κ .753)					
	Gemini Pro $\uparrow$	GPT-4 Vision $\uparrow$	LLaVA 7B $\downarrow$	LLaVA 13B $\downarrow$	LLaVA 34B $\uparrow$
Standard	64% (κ .277)	58% (κ .152)	50% (κ .002)	54% (κ .103)	50% (κ .000)
Informed	73% (κ .464)	70% (κ .402)	48% (κ .000)	52% (κ .065)	51% (κ .019)
Tiki/Giba (κ .655)					
	Gemini Pro $\downarrow$	GPT-4 Vision $\downarrow$	LLaVA 7B	LLaVA 13B $\uparrow$	LLaVA 34B $\uparrow$
Standard	49% (κ -.020)	43% (κ -.140)	50% (κ .000)	58% (κ .160)	50% (κ .000)
Informed	50% (κ .000)	36% (κ -.280)	50% (κ .000)	63% (κ .260)	51% (κ .020)
Kitaki/Gugagu (κ .670)					
	Gemini Pro $\uparrow$	GPT-4 Vision $\downarrow$	LLaVA 7B $\uparrow$	LLaVA 13B $\uparrow$	LLaVA 34B $\uparrow$
Standard	41% (κ -.180)	47% (κ -.060)	48% (κ -.040)	47% (κ -.060)	50% (κ .000)
Informed	54% (κ .080)	43% (κ -.140)	51% (κ .020)	50% (κ .000)	57% (κ .140)
Hatiha/Bodubo (κ .706)					
	Gemini Pro $\uparrow$	GPT-4 Vision $\uparrow$	LLaVA 7B	LLaVA 13B $\uparrow$	LLaVA 34B $\uparrow$
Standard	61% (κ .220)	47% (κ -.060)	50% (κ .000)	46% (κ -.080)	50% (κ .000)
Informed	78% (κ .560)	52% (κ .040)	50% (κ .000)	54% (κ .080)	51% (κ .020)
Penape/Gunogu (κ .627)					
	Gemini Pro $\downarrow$	GPT-4 Vision $\downarrow$	LLaVA 7B $\uparrow$	LLaVA 13B $\downarrow$	LLaVA 34B
Standard	38% (κ -.240)	29% (κ -.420)	47% (κ -.060)	38% (κ -.240)	50% (κ .000)
Informed	26% (κ -.480)	22% (κ -.560)	50% (κ .000)	35% (κ -.300)	50% (κ .000)
ALL (excl. Kiki/Bouba)					
	Gemini Pro $\uparrow$	GPT-4 Vision $\uparrow$	LLaVA 7B $\downarrow$	LLaVA 13B $\uparrow$	LLaVA 34B $\uparrow$
Standard	49.33% (κ -.014)	47.83% (κ -.045)	49.17% (κ -.016)	49.71% (κ -.013)	50.40% (κ .012)
Informed	57.33% (κ .147)	48.17% (κ -.036)	50.17% (κ .003)	51.50% (κ .034)	51.83% (κ .037)

Table 6.2: Results of the shape symbolism experiments per pseudoword pair. Fleiss’ κ (Fleiss, 1971) is presented next to each pseudoword pair for inter-annotator agreement between humans. For each VLM, we present Cohen’s κ (Cohen, 1960) for agreement between the models and the human majority vote. The model with the highest agreement per prompt is in **bold**. Arrows next to model names indicate the direction of agreement change from the standard to the informed prompt.

CHAPTER 6. WITH EARS TO SEE AND EYES TO HEAR

Mil/Mal (κ .529)					
	Gemini Pro $\uparrow$	GPT-4 Vision $\uparrow$	LLaVA 7B $\downarrow$	LLaVA 13B $\uparrow$	LLaVA 34B $\downarrow$
Standard	58% (κ .152)	50% (κ .031)	51% (κ .018)	50% (κ .000)	50% (κ .000)
Informed	75% (κ .512)	76% (κ .529)	50% (κ .000)	52% (κ .043)	44% (κ -.109)
Dil/Dal (κ .331)					
	Gemini Pro $\downarrow$	GPT-4 Vision $\downarrow$	LLaVA 7B $\downarrow$	LLaVA 13B $\uparrow$	LLaVA 34B
Standard	42% (κ -.168)	55% (κ .098)	58% (κ .149)	50% (κ .000)	50% (κ .000)
Informed	43% (κ -.128)	51% (κ .010)	50% (κ .000)	55% (κ .122)	50% (κ .000)
Zil/Zal (κ .245)					
	Gemini Pro $\downarrow$	GPT-4 Vision $\uparrow$	LLaVA 7B $\downarrow$	LLaVA 13B	LLaVA 34B $\uparrow$
Standard	48% (κ -.003)	58% (κ .149)	51% (κ .022)	61% (κ .222)	50% (κ .000)
Informed	47% (κ -.031)	93% (κ .860)	50% (κ .000)	61% (κ .207)	54% (κ .083)
Geel/Gaal (κ .577)					
	Gemini Pro $\uparrow$	GPT-4 Vision $\uparrow$	LLaVA 7B $\downarrow$	LLaVA 13B $\uparrow$	LLaVA 34B
Standard	56% (κ .120)	54% (κ .080)	49% (κ -.020)	58% (κ .160)	50% (κ .000)
Informed	66% (κ .320)	72% (κ .440)	50% (κ .000)	65% (κ .300)	50% (κ .000)
Beel/Baal (κ .560)					
	Gemini Pro $\uparrow$	GPT-4 Vision	LLaVA 7B	LLaVA 13B $\downarrow$	LLaVA 34B
Standard	59% (κ .180)	82% (κ .640)	48% (κ -.040)	57% (κ .140)	50% (κ .000)
Informed	73% (κ .460)	82% (κ .640)	50% (κ .000)	63% (κ .260)	51% (κ .020)
Weel/Waal (κ .541)					
	Gemini Pro $\uparrow$	GPT-4 Vision $\uparrow$	LLaVA 7B	LLaVA 13B $\downarrow$	LLaVA 34B $\uparrow$
Standard	55% (κ .100)	60% (κ .200)	50% (κ .000)	58% (κ .160)	50% (κ .000)
Informed	75% (κ .500)	90% (κ .800)	50% (κ .000)	55% (κ .100)	51% (κ .020)
Leel/Laal (κ .580)					
	Gemini Pro $\uparrow$	GPT-4 Vision $\uparrow$	LLaVA 7B	LLaVA 13B $\downarrow$	LLaVA 34B $\uparrow$
Standard	53% (κ .060)	53% (κ .060)	50% (κ .000)	53% (κ .060)	50% (κ .000)
Informed	67% (κ .340)	71% (κ .420)	50% (κ .000)	51% (κ .020)	55% (κ .100)
ALL (excl. Mil/Mal)					
	Gemini Pro $\uparrow$	GPT-4 Vision $\uparrow$	LLaVA 7B $\downarrow$	LLaVA 13B $\uparrow$	LLaVA 34B $\uparrow$
Standard	52.17% (κ .048)	58.50% (κ .169)	51.00% (κ .019)	56.17% (κ .127)	50.00% (κ .000)
Informed	61.83% (κ .243)	76.50% (κ .528)	50.00% (κ .000)	58.33% (κ .168)	51.83% (κ .037)

Table 6.3: Results of the magnitude symbolism experiments per pseudoword pair. Fleiss’ κ (Fleiss, 1971) is presented next to each pseudoword pair for inter-annotator agreement between humans. For each VLM, we present Cohen’s κ (Cohen, 1960) for agreement between the models and the human majority vote. The model with the highest agreement per prompt is in **bold**. Arrows next to model names indicate the direction of agreement change from the standard to the informed prompt.

Chapter 7

Seeing isn't Hearing: Benchmarking Vision Language Models at Interpreting Spectrograms

Proceedings of IJCNLP-AACL ([Loakman et al., 2025a](#))

Tyler Loakman^a, Joseph James^a, and Chenghua Lin^b

^aDepartment of Computer Science, University of Sheffield, UK

^bDepartment of Computer Science, The University of Manchester, UK

Contextual Statement

Building upon Chapter 6's introduction of multimodal understanding, this chapter addresses RQ6, investigating the extent to which vision language models (VLMs) are able to synthesise parametric knowledge of phonetics and phonology with visual inputs to aid in the interpretation of spectrograms and waveforms. Consequently, whilst the previous chapter explored the phenomena of sound symbolism, which arises indirectly in humans from correlations between speech sounds and physical characteristics, this chapter more explicitly assesses model knowledge of acoustic phonetics and English phonology. To achieve this, weighted phonemic edit distance (as seen in Chapter 4 for generating and assessing tongue twisters) is repurposed, this time for selecting minimally and maximally confusable distractors for the multiple choice spectrogram and waveform interpretation task. Furthermore, although transcribing speech from images is not a primary use of VLMs, the presented task assesses the general ability of VLMs to perform visual reasoning of complex data visualisations, whilst additionally reflecting the topic of esoteric knowledge retrieval presented in Chapter 5 regarding references to news and pop culture, but concerning knowledge of phonetics and phonology.

Abstract

With the rise of large language models (LLMs) and their vision-enabled counterparts (VLMs), numerous works have investigated their capabilities in tasks that fuse both vision and language modalities. In this work, we benchmark the extent to which VLMs are able to act as highly-trained phoneticians, interpreting spectrograms and waveforms of speech. To do this, we synthesise a novel dataset containing 4k+ English words spoken in isolation alongside stylistically consistent spectrogram and waveform figures. We test the ability of VLMs to understand these speech representations through a multiple-choice question task where models must select the correct phonemic or graphemic transcription of a spoken word when presented amongst three distractor transcriptions that have been selected based on their phonemic edit distance to the ground truth. We observe that both zero-shot and fine-tuned models rarely perform above chance-level. The results demonstrate the difficulty of this task and suggest models require esoteric parametric knowledge of how to interpret such figures, rather than paired samples alone.

7.1 Introduction

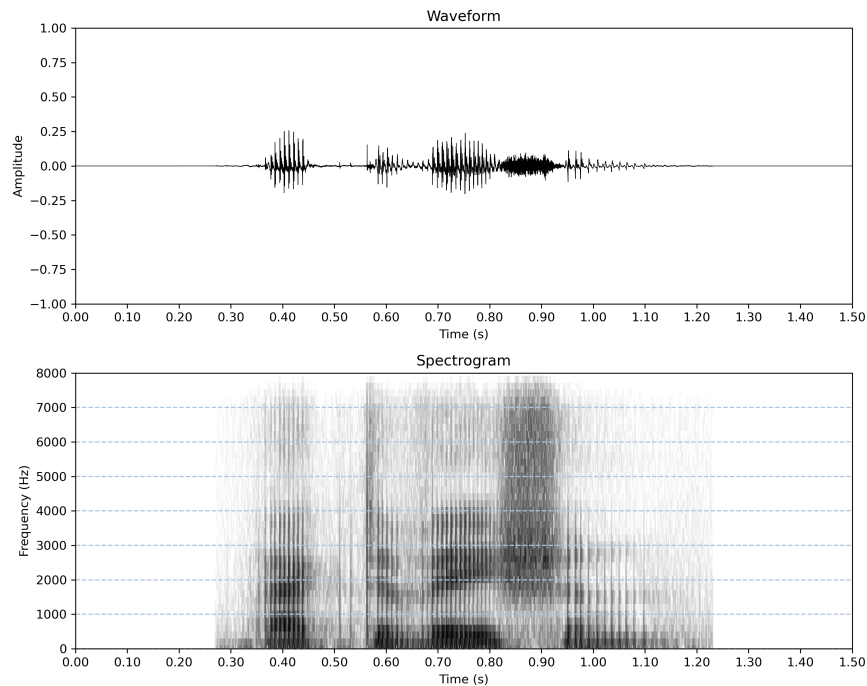


Figure 7.1: *An example waveform (top) and spectrogram (bottom) of “activation” spoken by a text-to-speech model. Horizontal frequency guidelines are presented in blue to not be mistaken for part of the speech signal.*

The ability of large language models (LLMs) and vision language models (VLMs) to reason about multimodal data has been studied extensively (Chia et al., 2024; Lin et al., 2023; Li and Zhang, 2023). One of the most productive domains for testing the unification between vision and language is that of the explanation and creation of data visualisations, such as graphs and figures (Zhang et al., 2024; Ge et al., 2024). For the most part, such works have tested models’ abilities to understand axes and labels, follow trend lines, and reformulate visual content as text (or vice versa). However, at the time of writing, little work had been performed on the ability of VLMs to understand more complex visualisations, where additional parametric knowledge must be combined with image inputs to facilitate visual reasoning.

To that end, we present the novel task of spectrogram and waveform interpretation, where models must identify words from visual representations of speech. As seen in Figure 7.1, the x-axis of a spectrogram represents time, whilst the y-axis represents sound frequency, and the heat-map intensity represents the density of energy within a particular frequency range at a given time. On the other hand, a waveform also represents time on the x-axis, but displays the amplitude of a signal on the y-axis. Unlike more common visualisations (e.g., bar charts, pie charts, and line graphs), to interpret such figures, models must reason about the visual content to establish what word is being spoken, utilising parametric knowledge regarding acoustic phonetics to aid in understanding the observed signal characteristics, such as vowel formants and fricative centre of gravity (Ladefoged, 1996).

The contributions of this work are as follows:

- We present a novel dataset of 4k+ words spoken in isolation using a range of synthetic voices, each with figures consisting of a spectrogram in isolation, or a spectrogram with a waveform.
- We perform what we believe to be the first analysis of whether VLMs can correctly interpret spectrograms and waveforms of speech in a phonologically-motivated multiple-choice paradigm.

7.2 Related Work

Data-to-text tasks have long been a staple of natural language processing (NLP) research (Castro Ferreira et al., 2019). Traditionally, such approaches have relied on access to largely text-based representations, such as converting tables to markdown or HTML (He et al., 2023; Liu et al., 2022b; Eisenschlos et al., 2020). However, the rise of VLMs has brought with it a range of benchmarks concerning image-based inputs (Zheng et al.,

2024) such as charts (Islam et al., 2024; Zhou et al., 2023a) and graphs (Ai et al., 2024), which have been fruitful in assessing data-to-text abilities (Zhu et al., 2025b; Islam et al., 2024; Ai et al., 2024). However, VLMs have been demonstrated to struggle with reasoning (Mukhopadhyay et al., 2024; Hou et al., 2025) and being easily subjected to training biases (Vo et al., 2025). Furthermore, works by Loakman et al. (2024) and Alper and Averbuch-Elor (2023) have investigated the meta-level abilities of models to reason about language, investigating whether or not they learn associations between vision and text that can be used to simulate an understanding of sound symbolism.

Building upon existing works in visual reasoning and the meta-level language understanding of VLMs, we present what we believe to be the first benchmark and analysis of the ability of VLMs to interpret spectrograms and waveforms of speech, combining visual acuity with phonetic knowledge.

7.3 Dataset

7.3.1 Word List

Due to the effects of coarticulation in fluent connected speech, we present VLMs with English words spoken in isolation. To reduce ambiguity, words are required to be representative of standard dictionary entries rather than the acronyms, initialisms, and titles often found in word frequency lists.¹ As a result, we use the Oxford 5000 word list, consisting of words from A1-C1 difficulty on the Common European Framework of Reference for languages (CEFR) (Council of Europe, 2001).² We remove words that have more than one possible pronunciation in the CMU Pronouncing Dictionary (CMUDict), such as homographs, removing the need to manually verify which version

¹This is to avoid ambiguity such as whether “Dr.” is pronounced as “doctor” [dɒktə] or “dr” [dɪ].

²<https://www.oxfordlearnersdictionaries.com/about/wordlists/oxford3000-5000>

of the word the speech synthesis model produces (owing to words being presented in isolation, void of contextual cues).³ This results in 4068 words after filtering.

7.3.2 Speech Synthesis

Owing to the high levels inter- and intra-speaker variation present in human speech and the range of possible acoustic environments for recording (Jessen, 2008), we create a novel dataset using Microsoft’s SpeechT5 trained for the task of speech synthesis (Ao et al., 2022a) to generate each word in isolation. We use synthetic voices with a General American (GenAm) (Wells, 1982a) accent due to this enabling the use of the CMUDict).⁴ A wider range of accents would require the manual creation of pronunciation dictionaries for each accent (intra-speaker variation notwithstanding), increasing the likelihood of phonetic realisations from the text-to-speech (TTS) model not accurately reflecting the transcriptions. We select 4 speaker embeddings from CMU ARCTIC X-Vectors⁵ (two male and two female perceived voices) for training, and two for testing (one male and one female). We additionally add the General American voice from Google Translate TTS to the test set to mitigate biases from the characteristics of a particular TTS system.⁶ Further details are presented in Appendix 7.8.1.

7.3.3 Spectrograms & Waveforms

As the image input to the VLMs, we provide a spectrogram of the target word produced by a TTS, either alone or with an accompanying waveform (akin to Figure 7.1). To ensure consistency across generated figures, audio files are padded or truncated to 1.5 seconds in length, which was validated to be longer than the speech portion of any

³This refers to words with multiple pronunciations in the General American accent covered by CMUDict. There is inter-speaker variation across English accents for all words.

⁴Accessed via the CMUDict library: <https://pypi.org/project/cmudict/>

⁵<https://huggingface.co/datasets/Matthijs/cmu-arctic-xvectors>

⁶Accessed via the gTTS library: <https://pypi.org/project/gTTS/>

generated audio files. All spectrograms are generated using Librosa (Brian McFee et al., 2015), with an n_fft of 128, a hop-length of 22, and a dynamic range of 70dB, using a Hann window. We find these settings to ensure an adequate frequency resolution without unnecessarily detailed harmonic information. An example of a generated figure is presented in Figure 7.1.

7.4 Methodology

7.4.1 Task

To test the ability of VLMs to interpret visual representations of speech, we use a multiple-choice question (MCQ) paradigm where the ground truth is paired with three distractors. Owing to the importance of high-quality distractors (Goel et al., 2025; Alhazmi et al., 2024; Ding et al., 2024), these are selected based on their weighted phonemic edit distance (PED) (Mortensen et al., 2016) from the ground truth, which is based on phonological distinctive features such as place and manner of articulation (Chomsky et al., 1991 - 1968). Consequently, distractors are selected based on whether they are phonemically similar or distant to the ground truth, allowing the identification of whether tested VLMs are making informed (yet incorrect) choices or selecting randomly. Specifically, the distractors consist of the words with the lowest, median, and highest PED from the ground truth that have less than 50% overlap in consecutive phonemes.⁷ We disallow distractors that start with the same character as the ground truth or other distractors to avoid overt biasing signals further.

⁷In preliminary testing, we observed that models were able to reach 50% accuracy without being presented with an image. This was due to the ground truth and most similar distractor frequently being minimal pairs (e.g., “sheer” and “cheer”).

CHAPTER 7. SEEING ISN'T HEARING

The filtered list of 4,068 words is broken down into an 80-10-10 split for training/validation/testing, whilst ensuring comparable distributions of word length and individual phonemes (see Appendix 7.8.1 for further details).

To mitigate ordering bias (Pezeshkpour and Hruschka, 2024; Wei et al., 2024), we expanded the training and validation sets by generating four permutations of each question. In these permutations, the target word appears once in each of the four possible positions, with the distractor positions randomised. For each permutation, one of the four synthetic voices was selected, ensuring an even distribution of voices across the training and validation splits. The test set consists of only one permutation per word, ensuring that the target word appears equally in all positions, with one of the three held-out voices randomly selected. Again, we ensure that all three voices appear equally across the test set. Consequently, the final augmented splits consist of 13,016/1,628/407 questions for training/validation/testing, respectively.

Conditions. In total, experiments are performed under four conditions. Answer options were presented in either graphemic (i.e., orthographic words) or phonemic (i.e., International Phonetic Alphabet) forms. Additionally, image inputs consisted of either a spectrogram or a spectrogram with a waveform.

7.4.2 Models

We test a range of open-source VLMs from different model families in this task.⁸ For zero-shot performance, we investigate the results from both small and large model variants, whilst for fine-tuning, we focus only on the small variants, owing to resource constraints. Specifically: Qwen 2.5 VL (7B and 32B) (Bai et al., 2025b), LLaVA 1.6 (7B and 34B) (Liu et al., 2024b), and InternVL 3 (8B and 38B) (Zhu et al., 2025a).

⁸Closed-source proprietary models were not tested due to API expenses.

Additional details, including fine-tuning hyperparameters, specific model variants, and prompts, are provided in Appendix 7.8.2

7.5 Results

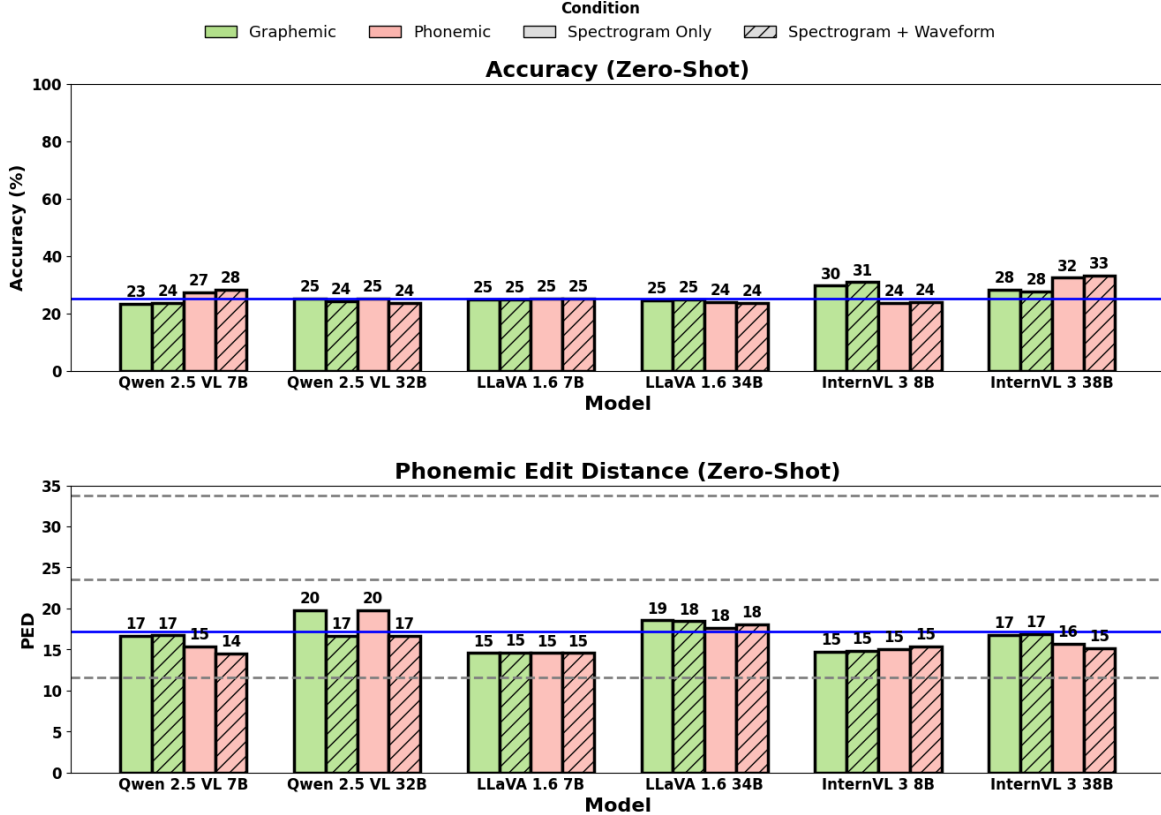


Figure 7.2: Zero-shot results in the multiple choice spectrogram interpretation task. The solid horizontal blue lines represent expected values from random selection. The dashed lines in the phonemic edit distance plot relate to the expected mean distance for consistently selecting the 2nd, 3rd or 4th most similar option.

Zero-shot performance results are presented in Figure 7.2. Regarding accuracy, all models consistently perform around chance-level (25%) in all conditions. No demonstrable benefit is observed from the additional information presented in the waveform, which would usually be considered useful for identifying fricatives and plosives from aperiodic

CHAPTER 7. SEEING ISN'T HEARING

waves, or vowels from periodic waves. Performance is even shown to decrease for Qwen 2.5 VL when waveforms are included in the image input. Furthermore, regarding the format of options, performance is generally higher for Qwen 2.5 VL and InternVL 3 38B when presented as phonemes, but this does not hold for other models.

One reason for this is that the phoneme condition may be more likely to activate knowledge directly related to acoustic phonetics, which is relevant to the task of spectrogram and waveform interpretation. On the other hand, the average phonemic edit distance of the selected options suggests that low-distance distractors are frequently selected. Whilst a random selection is expected to have a phonemic edit distance of 17.2, most models perform slightly better than this, suggesting that models may be making informed decisions, even if not selecting the correct answer.

The results of fine-tuning the smaller model variants are presented in Figure 7.3. Higher levels of performance are observed in the fine-tuned setting than in the zero-shot condition, particularly for InternVL 3. However, no statistically significant difference is observed between the performance of the fine-tuned models when tested with and without image inputs, suggesting that this performance increase may be the result of learning the remaining biases in the dataset.

To evaluate the impact of each condition on VLM performance, paired-sample t-tests were performed on model summary statistics, aggregating results across model architectures to investigate the effects of option type (grapheme or phoneme) and input type (spectrogram or spectrogram with waveform). Regarding accuracy, no significant effect was found for either variable at $\alpha = 0.05$, either for option type ($t(8) = -0.716$, $p = .494$) or input type ($t(8) = -0.327$, $p = .752$). There was also no significant effect on phoneme edit distance, for option type ($t(8) = 2.028$, $p = .077$) or input type ($t(8) = -0.345$, $p = .739$).

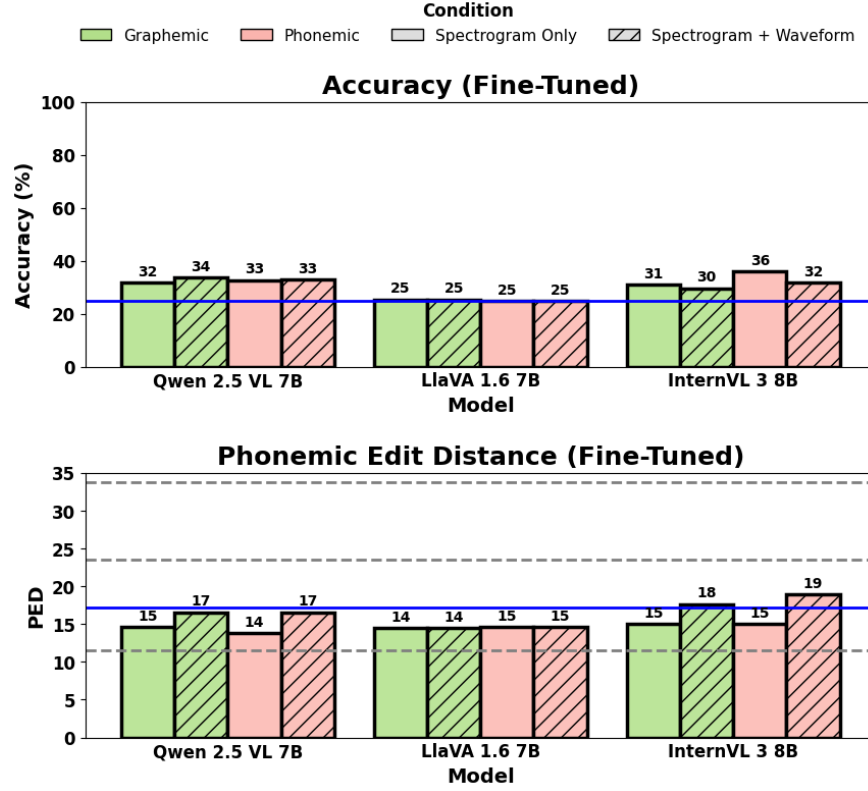


Figure 7.3: Fine-tuned results in the MCQ spectrogram and waveform interpretation task. The solid horizontal blue lines represent expected values from random selection. The dashed lines in the phonemic edit distance plot relate to the expected mean distance for consistently selecting the 2nd, 3rd or 4th most similar option.

7.5.1 Automatic Speech Recognition & Phonetician Baselines

As a traditional baseline for speech understanding, the results of Nvidia’s NeMo Canary Qwen 2.5B, an automatic speech recognition (ASR) system, are presented on the test set’s corresponding audio (see Appendix 7.8.2 for more details). When accounting for case folding and punctuation removal to determine exact matches, the ASR model achieves an accuracy of 87.56%, demonstrating that the underlying task is easily solvable when the acoustic information is given as input to a purpose-built ASR system. This contrasts with VLM performance, where models struggled when the same information was presented visually as spectrograms and waveforms. Finally, human evaluation

was conducted on a subset of 100 random examples from the test set consisting of a spectrogram and waveform. This allowed the establishment of an upper bound for the task of spectrogram and waveform interpretation, using a trained phonetician, who is the primary author of this work. Importantly, the second author was responsible for generating and presenting the random subset to avoid bias from pre-exposure. We observe an accuracy of 75% and an average phonemic edit distance of 5.17, highlighting the considerable gap between human and model performance, whilst demonstrating that there is sufficient information within the visualisations to achieve high accuracy.⁹

7.6 Conclusion

This work presented the first analysis of the ability of VLMs to interpret speech from spectrograms and waveforms, using a phonologically motivated multiple-choice question approach. The results demonstrated that both zero-shot and fine-tuned models struggle to identify the correct answer, demonstrating the difficulty of the task and its potential as a benchmark assessment of the ability of VLMs to combine esoteric parametric knowledge with vision and text inputs.

7.7 Limitations

Owing to computational limitations, we did not fine-tune the larger model variants or benchmark large proprietary vision-enabled models such as OpenAI’s o1 and o3, which have associated API expenses. Furthermore, our dataset consists of synthetic speech to reduce the effects of co-articulation and inconsistent acoustic environments. However, synthetic speech is less variable than natural human speech and therefore presents an easier form of the task. Additionally, whilst longer speech sequences present a more

⁹We expect layperson accuracy on this task to be no better than chance.

practical test, such sequences increase the complexity of generating equal distributions across splits, whilst impacting the resolution of phonetic characteristics, with more speech signal being included in each image patch. Furthermore, we analyse models using zero-shot and fine-tuned approaches with a simple multiple-choice paradigm. However, we do not include any explicit knowledge within the prompt itself. Future work may be best served by training models to recognise the characteristics of individual phonemes before learning entire words, as well as exploring the benefits of pedagogical prompts that explicitly contain phonetic knowledge. However, pedagogical prompting was explored in a small-scale pilot study (without prompt engineering), with no significant performance benefit observed.

Additional Acknowledgments

Joseph James is supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [grant number EP/S023062/1].

7.8 Appendix

7.8.1 Dataset Characteristics

Figure 7.4 presents the distribution of words by length (in phonemes) across the training, development, and test sets, in addition to the distribution of specific phonemes.

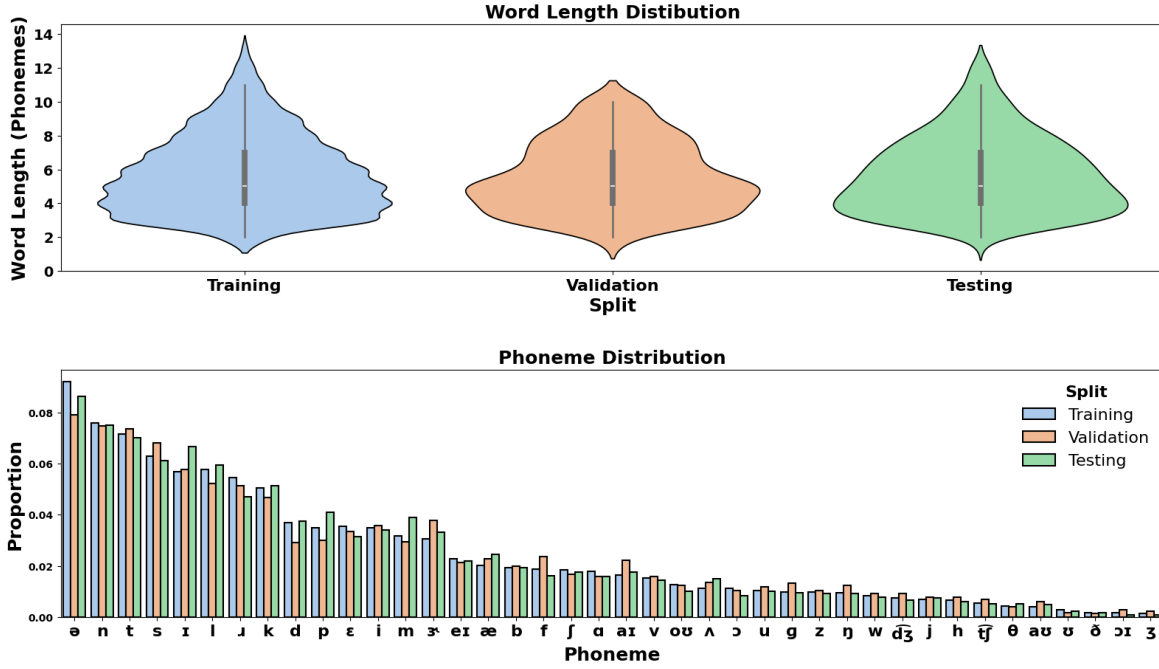


Figure 7.4: *Distribution of word lengths in phonemes (top) and individual phoneme counts (bottom) across the training, validation and test sets.*

7.8.2 Implementation Details

Text-to-Speech Voices. From the CMU ARCTIC X-Vectors, we use speaker embeddings 69 (M), 750 (M), 2500 (F), and 3333 (F) for training, alongside 4444 (M) and 7500 (F) for testing. We additionally use Google Translate’s General American TTS voice (which is only available in a female-presenting form) for testing.

CHAPTER 7. SEEING ISN'T HEARING

Models. All selected models were taken from Hugging Face. Specifically, Qwen 2.5 VL (Qwen/Qwen2.5-VL-7B-Instruct and Qwen/Qwen2.5-VL-32B-Instruct), LLaVA 1.6 (i.e., LLaVA NeXT) (llava-hf/llava-v1.6-vicuna-7b-hf and llava-hf/llava-v1.6-34b-hf), and InternVL 3 (OpenGVLab/InternVL3-8B-hf and OpenGVLab/InternVL3-38B-hf). For the ASR baseline, we used NeMo Canary Qwen (nvidia/canary-qwen-2.5b). At the time of publication, ranked among the top-performing models on the Open ASR leaderboard in terms of word error rate.¹⁰

Prompt. “You are a Vision-Language Model specialized in the phonetic interpretation of speech spectrograms. Your primary role is to act as a highly trained phonetician: given visual representations of spoken English words in the General American accent (spectrograms and/or waveforms), you must determine which graphemic or phonetic transcription correctly matches the spoken word.\n For each example, you will be shown:\n A spectrogram (and optionally a waveform) of a single English word spoken in isolation.\n Four candidate transcriptions labelled 0 to 3 (one correct, three phonetic distractors).\n Provide the final label as <label>label</label>, where label is the number (0-3) of the correct transcription.\n Just output the label and nothing else in the format: <label>label</label>”

Training Details. Fine-tuning and inference for all models were performed on a single Nvidia A100 GPU. Hyperparameters used during fine-tuning are presented in Table 7.1.

¹⁰https://huggingface.co/spaces/hf-audio/open_asr_leaderboard

Hyperparameter	Setting
Epochs	5
Batch Size	1
Gradient Accumulation Steps	8
Optimiser	AdamW
Learning Rate (LR)	2e-4
Weight Decay	0.01
Warmup Steps	150
Evaluation Steps	150
Evaluation Strategy	steps
Early Stopping Callback	3
Hardware	Nvidia A100

Table 7.1: *Fine-tuning hyperparameters.*

Part VI

Perspectives on AI Development

Chapter 8

Do as I Say, Not as You Do:

Assessing Public Versus Practitioner Perspectives on the Development of Artificial Intelligence and Language Technologies

Under Review

Tyler Loakman^a and Chenghua Lin^b

^aDepartment of Computer Science, University of Sheffield, UK

^bDepartment of Computer Science, The University of Manchester, UK

Contextual Statement

In the final publication-oriented chapter of this thesis, attention shifts from computational experiments regarding phonetic and phonological understanding and humour, towards investigations concerning AI ethics and impact. This chapter addresses RQ7, investigating the extent to which AI practitioners (comprising academics and industry professionals) differ in their perspectives towards AI development and usage versus the general public. Specifically, this chapter presents the results of a questionnaire on topics including AI art, creativity, and consciousness, as well as the use of copyrighted material for model training and the overall societal impact of AI. The parameter counts of large language models (LLMs) had been rapidly accelerating during the earlier stages of work towards this thesis, achieving widespread public awareness with the release of ChatGPT. As a result, the capabilities of such models and the subsequent societal impact were yet to be understood. As the body of work that constituted this thesis progressed, the research focus transitioned from the direct generation of creative language seen in Chapter 3 and Chapter 4. This shift reflects a deepening critical engagement with the field and growing concerns about the automation of creative labour. Consequently, this chapter serves to quantify the observed concerns in the wider population (particularly regarding the impact of AI on creative domains), and additionally present a forward-looking assessment of how AI development and research should continue to progress to best serve society as a whole.

Abstract

With the public awareness of artificial intelligence (AI) applications such as large language models (LLMs) and text-to-video models reaching unprecedented levels, there is an increasing need to reconcile the perspectives held by informed practitioners, who research and develop these technologies, with the interests and opinions of the general public, who constitute the majority of people impacted by technological developments. In this work, we survey practitioners (academics and industry professionals) alongside members of the general public on perspectives relating to pertinent areas of artificial intelligence, including perceived benefit and risk, the ability for models to achieve consciousness and output art, and the use of copyright-protected material for model training. Alongside comparisons of diverse perspectives on these topics, we additionally contextualise the findings with a discussion of relevant literature that supports and refutes the myriad observed stances.

8.1 Introduction

The advent of technologies such as large language models (LLMs) like ChatGPT (OpenAI et al., 2023) and other artificial intelligence (AI) applications has driven the unprecedented integration of AI into all aspects of life, including healthcare (Luo et al., 2025; Mondal et al., 2022), education (Kumar and Lata, 2025; Han et al., 2024b; Selwyn, 2022), entertainment (Hemraj, 2025), and everyday consumer goods such as fridges and e-cigarettes. The available investment capital for AI technology has additionally resulted in “AI washing”, where products and services exaggerate the extent and effectiveness of AI integration (Ozturkcan and Bozdağ, 2025; Elhajjar and Itani, 2025; Wu and Chen, 2022). As a result, the integration of AI into everyday life causes substantial economic gains for some parties, whilst presenting an economic risk to others (Restrepo, 2025; Yuan et al., 2024a).

There has been a growing body of work in natural language processing (NLP) on the post-training stage of model development, particularly regarding the alignment of model behaviours to human preferences (Pan et al., 2025; Shu et al., 2025). However, a substantial quantity of human preference data is typically required. One issue that commonly arises during this stage is that such approaches do not adequately account for the plurality of perspectives held by individuals across different cultures and social groups, leaving some demographics under-represented (Zhi-Xuan et al., 2025; Barretto et al., 2021). Similarly, the domain of model debiasing and detoxification relies on prior expectations about what is considered harmful and offensive (Zhi-Xuan et al., 2025). However, research shows that over-reliance on such approaches results in the degradation of overall model performance and a preference for abstaining from responding to topics that may be construed as sensitive or contentious, effectively neutering model usability (Kaneko et al., 2023; Meade et al., 2022).

CHAPTER 8. DO AS I SAY, NOT AS YOU DO

Alongside the rapid deployment of AI technology and rising challenges in achieving representative alignment for models, there remains a substantial gap in understanding regarding the resulting societal impact. Crucially, the priorities and policies governing those developing AI (frequently driven by chasing leaps in technical capabilities and corporate objectives) may not align with the concerns of the public consumers whose lives these systems fundamentally affect ([Horwitz, 2025](#); [Amazon Employees for Climate Justice, 2025](#); [Right to Warn Coalition, 2024](#)).

To address this critical gap, we conduct a survey to compare the perspectives of AI practitioners (consisting of academics and industry professionals) with those of the general public. In doing so, we ascertain whether the creators and consumers of such technologies are aligned, thereby gaining a better understanding of the benefits and risks of continued AI development.

We summarise the primary contributions of this work as follows:

- We survey the three primary stakeholders in the development of artificial intelligence: academics, industry professionals, and the general public (dichotomised as practitioners and the public) on topics relating to both practical and philosophical domains (e.g., economic and legal concerns, art, and consciousness).
- Responses are analysed to explore whether practitioners and the public are aligned in their perspectives on AI, assessing whether the development of such technologies is to the benefit of society as a whole.
- We provide an overview of the literature surrounding AI development, including areas of legal, economic, and social concern.

8.2 Related Works

Public Perception of AI. Existing works have explored the perspectives of different stakeholders in AI development. For instance, [Kendall Roundtree \(2024\)](#) presents a meta-analysis of literature surrounding public perceptions of AI development alongside a sentiment analysis of social media communication on Quora and Reddit. Their findings demonstrate concerns regarding the impact of AI development on employment and the risks of AI-driven healthcare. Furthermore, [Brauner et al. \(2023\)](#) surveyed 122 members of the German public on opinions regarding personal, economic, industrial, social, cultural, and health-related factors relating to AI development. They identified cybersecurity risks as being the most critical point of public concern when accounting for the interaction between risk and likelihood, whilst threats to careers and personal lives were seen as unlikely, and the position of AI performing unpleasant and undesirable human tasks was viewed as both likely and desirable. Furthermore, [Selwyn and Gallo Cordoba \(2022\)](#) analysed the understanding of the Australian public on AI in 2020, prior to the widespread adoption of performant LLMs. Their results showed that most participants had a sufficient working understanding of the technology, even if they self-identified as lacking knowledge.

Comparative Studies. Comparing the general public and practitioners, [Pew Research Center \(2025\)](#) directly surveyed 5k+ members of their American Trends Panel and 1k+ authors of AI publications in 2024 (all of which reside in the United States).¹ Results showed that AI experts were significantly more optimistic than the public regarding the potential impact of AI, particularly towards medical care, the economy, and education. We compare the findings of the present work with [Pew Research Center](#)

¹As a result, this work was both American-centric and also surveyed members of a questionnaire panel, resulting in a substantial response bias. In contrast, this work is international (though predominantly European), and anonymous.

(2025) in Appendix 8.11.1. Whilst these findings establish a trend of practitioner optimism and public pessimism, existing comparative works have focused on broad societal applications (e.g., the economy and healthcare), whilst the present work additionally explores philosophically-laden topics such as art, creativity, and consciousness.

Present Work. Building upon these findings, we address gaps in the literature by surveying public and practitioner views on AI with a predominant (though non-exclusive) focus on language technologies. In addition to topics pertaining to practical areas such as the economy and healthcare, we further investigate perspectives surrounding philosophical concerns, such as the ability for AI to generate art and demonstrate consciousness. Such exploration enables a deeper understanding of people’s fundamental perspectives regarding subjective topics where opinions are more resistant to change, in comparison to concerns that can be empirically verified (e.g., the impact of AI on economic growth and employment).

8.3 Methodology

Recruitment of Participants. Recruitment was primarily performed via word-of-mouth to maintain anonymity, bootstrapped by addressing email lists at the lead author’s institution, posting on social media, and approaching industry contacts. A wider range of general public respondents was sought through posts on Reddit and the placement of the survey on SurveySwap. All participants were encouraged to share the survey to further eligible participants.

Response Filtering. All participants were asked to respond to questions outlining their understanding and familiarity with AI technologies in addition to their understanding of the technical underpinnings of machine learning (e.g., architectures and

CHAPTER 8. DO AS I SAY, NOT AS YOU DO

training methodologies). Participants additionally self-identified their position within AI development as being in academia (e.g., researchers and lecturers in natural language processing, computer vision, and general machine learning), industry (e.g., private researchers and applied AI software developers) or the general public. Using responses to a five-point Likert scale, we removed self-identified in-domain academics and industry practitioners who did not identify their own general and technical AI knowledge as above or equal to 3. As a result, we filter out mistaken self-identifications (as practitioners necessarily must have sufficient domain-knowledge), whilst not being overly restrictive (e.g., requiring technical knowledge ratings of 4-5), therefore allowing multidisciplinary practitioners who may regularly use AI systems as the focus of their work, but for whom a deep technical knowledge may not be mandatory. Finally, response timestamps were checked for suspicious activity, such as spam.

Demographic Information. In total, the survey received 336 responses: 69 academics, 33 industry professionals, and 234 members of public. Following filtering, this resulted in 76 practitioners across academia and industry, and 212 members of the general public. Over half of respondents were Millennials (born 1981-1996), with similar numbers of Gen X (1965-1980) and Gen Z (1997-2006), at $\approx 20\%$ each).² A further small number of respondents were from the Baby Boomer generation (1946-1964). In terms of geographical spread, the majority of respondents were from Europe ($\approx 60\text{-}70\%$), with a considerable number of North American responses ($\approx 17\text{-}30\%$). A further nominal number of responses were received from Oceania, East Asia, Southeast Asia, and the Middle East/North Africa (MENA). In total, age and geographical demographics were comparable across practitioners and the public, providing a sample that is not explicitly restricted to a specific locale or other demographic variable.

²Respondents were required to be at least 18 years of age.

Rating Scales. All non-categorical statements were rated on a 5-point Likert (i.e., “strong disagree” to “strong agree”, or the semantic equivalent for a given question phrasing).³ Only broad demographic information was collected to encourage honesty, as some respondents may have opinions that challenge the public stance of the organisations they represent. Question phrasing has been condensed for inclusion in visualisations, whilst the full original questions are presented in Table 8.1 within Appendix 8.11.2, in addition to significance testing for ordinal responses.⁴

8.4 Artificial Intelligence Usage

With the wide range of forms that artificial intelligence and machine learning models take, from simple grammar correction and translation tools to in-line code suggestions and autonomous driving assistance, in Figure 8.1 we first outline the different technologies being used both at home and for work.

Similar popularity rankings can be seen for different technologies between practitioners and the public, but with different levels of general use. For instance, the most popular technologies for personal use across both groups are translation tools and open-domain chatbots, with more than half of respondents from both groups using such technologies, whilst a considerable portion of people additionally interact with image generation, virtual assistants, and grammar correction AI ($\approx 20\%$). Across both groups, the rarest technologies used by participants are autonomous driving, AI-driven telemedicine, and music generation models - likely owing to their niche application, relatively rudimentary abilities, or high financial barrier to entry. Importantly, over half of public respondents used open-domain chatbots (i.e., LLMs). Such results are unsurprising, with OpenAI revealing that by 2025, approximately 10% of the world’s

³Participants were only presented with the linguistic labels for the two extremes (i.e., 1 and 5).

⁴Differences between practitioner and public responses are significant at $\alpha = 0.01$ for all statements.

CHAPTER 8. DO AS I SAY, NOT AS YOU DO

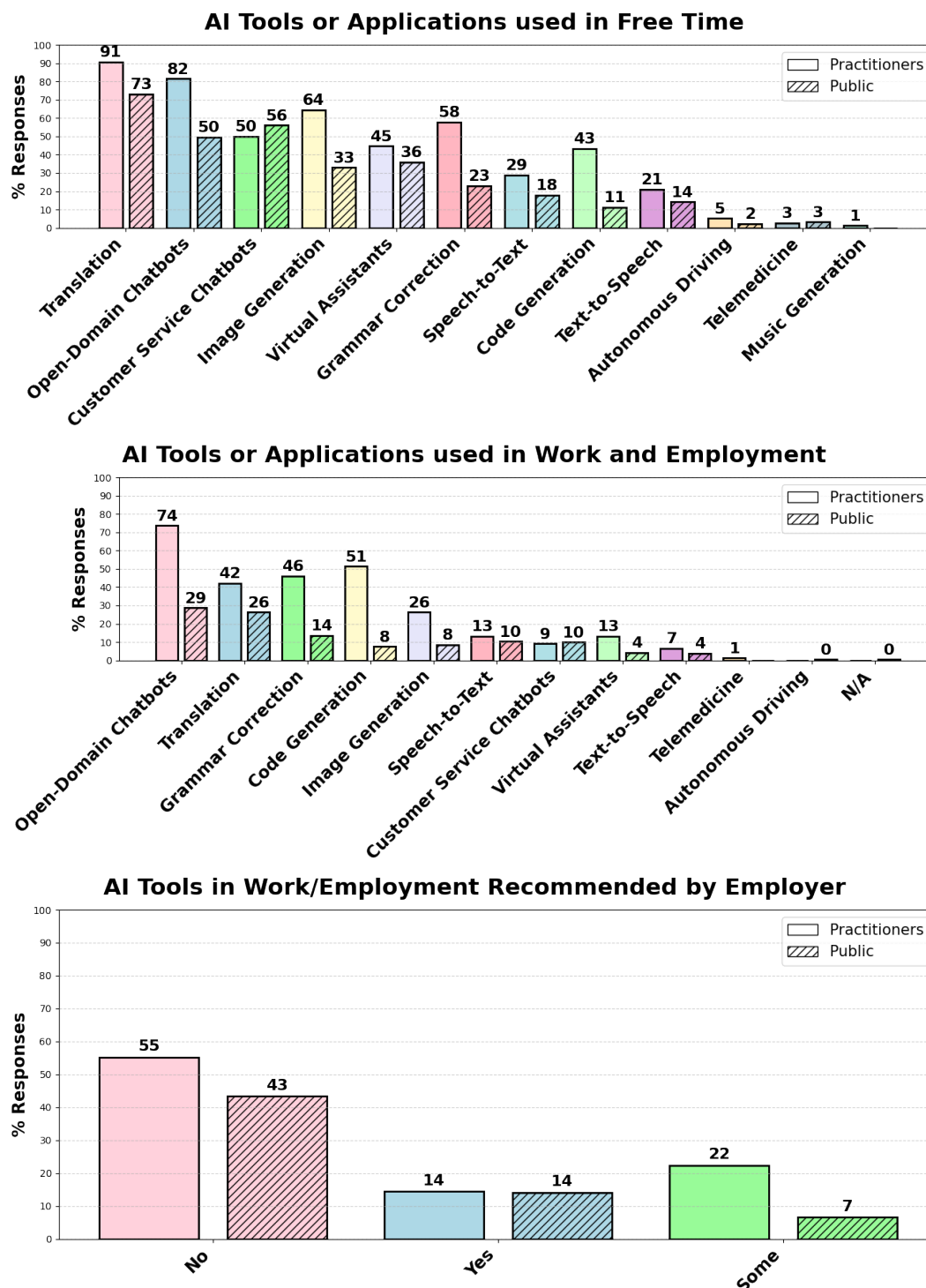


Figure 8.1: *Practitioner and public use of AI technologies in free time and employment.*

adult population had interacted with ChatGPT alone (Hitzig et al., 2025). Regarding AI-usage at work (or in study, for full-time students), less than 30% of public respondents interact with any, whilst the majority of practitioners make use of open-domain chatbots and specialised code models. Of the respondents that reported using any AI applications in their employment, almost half of the public (43%) and more than half of practitioners (55%), did so without endorsement or recommendation by their employer, presenting data protection concerns. Furthermore, 22% of practitioners and 7% of the public reported that only a subset of the AI technologies used in employment were endorsed by their employer, whilst still making use of technologies outside of this scope.

8.5 Training Data Collection

The following section focuses on concerns surrounding the use of public material for model training. Owing to the need for large volumes of data to imbue models with sufficient world-knowledge and enable reasoning abilities, much of the available training data is sourced via scraping large volumes (i.e., the majority) of available web content. As a result, popular pre-training corpora include collections such as CommonCrawl (Common Crawl Foundation, 2025), RedPajama (Weber et al., 2024), and The Pile (Gao et al., 2020), which consist largely of publicly available data.

However, whilst models exist that have been trained on highly curated and ethically collected data, such as CommonCorpus (Langlais et al., 2025) and Olmo (Ettinger et al., 2025; Walsh et al., 2025; Groeneveld et al., 2024), the majority of large-scale pre-training still concerns mass collected data where little regard is given to the licenses of the original works, the rights of copyright holders, and the general desires of the material creators.

CHAPTER 8. DO AS I SAY, NOT AS YOU DO

Whilst at the time of publication, the regulation of AI remains inconsistent and unclear across jurisdictions (Sloane and Wüllhorst, 2025; Kaliisa et al., 2025), there has been a growing volume of legal challenges on the topic of model training, including for language generation (Advance Local Media LLC v. Cohere, 2025), code generation (Doe v. Github, Inc., 2022), and image generation (Andersen v. Stability AI, 2024). In part, such cases concern the disregard of traditionally upheld copyright protections and the debated matter of model training constituting fair use. This has resulted in loopholes being exploited by some companies, such as Anthropic legally obtaining non-pirated physical copies of print publications and destructively scanning them (Teilmann-Lock and Savin, 2025; Bartz v. Anthropic PBC., 2025), following fair-use precedent set by Google Books (Authors Guild Inc. v. Google Inc., 2015).

Additional complications arise from unclear agreements between platforms and users, where user consent is purportedly given for data to be used for AI purposes, whilst the layperson user is not fully aware of the ramifications or details of said agreements. Such patterns have been previously observed with the use of end-user license agreements (EULAs), where details are obfuscated behind legalese or buried with extensive terms and conditions that users frequently accept without critique or further scrutiny (Corbett, 2019; Hutton and Henderson, 2017).

Figure 8.2 presents survey results regarding whether companies should be permitted to train on public data without the explicit consent of material creators. Overall consensus of both groups is against the motion that such material should be used without explicit consent (64% for practitioners and 90% for the public). However, a clear discrepancy is seen between the groups, with 31% of practitioners being in favour of being able to make use of copyright-protected material without consent, versus only 4% of the public.

CHAPTER 8. DO AS I SAY, NOT AS YOU DO

One potential explanation is that practitioners view the often lengthy and expensive process of obtaining license information, contacting rights-holders, and obtaining explicit informed consent to be burdensome, whilst members of the public may not believe that the benefit they receive from more performant AI products outweighs the implications of private organisations profiting from models trained on the work of uncompensated individuals. In particular, this topic raises a potential conflict of interest in practitioners, with many positions being highly lucrative compared to most employment performed by the public (Bone et al., 2025; PwC, 2025). As a result, being able to more quickly perform research and develop successful products may lead to increased compensation or career progression for practitioners, whilst from the public perspective many jobs are placed under growing threat of replacement (Restrepo, 2025). On the other hand, mandating that companies have the requisite permissions for all training data can be viewed as a competitive disadvantage in the AI race, allowing companies in jurisdictions with more relaxed intellectual property laws to have a significant advantage by having access to substantially larger volumes of data (Rosati, 2025, 2019). We speculate that open-access models with lower barriers to entry (i.e., no fees) are more likely to receive licensing permissions owing to the wider democratisation of their capabilities.

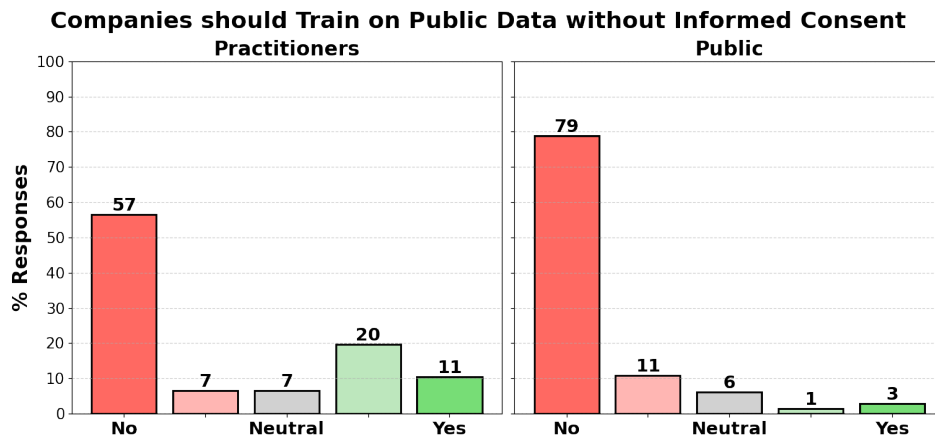


Figure 8.2: *Practitioner and public beliefs concerning AI models being trained on public data without explicit informed consent.*

8.6 Art & Creativity

Whilst AI applications have previously been developed predominantly for more mundane or supra-human tasks such as text summarisation and cancer screening, models have now become capable of generating articulate, engaging text and life-like images and video, resulting in such systems increasingly being used in tasks that have historically relied on human creativity, such as writing narratives (Teleki et al., 2025) and poetry (Gonçalo Oliveira, 2017), or designing infographics (Ghosh et al., 2025; Zhu et al., 2020) and movie trailers (Yao et al., 2025; Argaw et al., 2024).

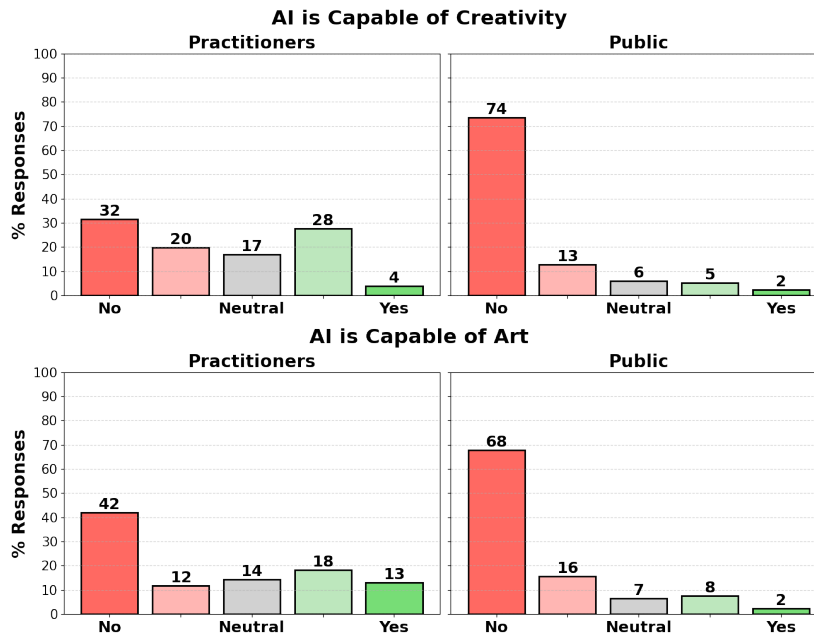


Figure 8.3: *Practitioner and public beliefs concerning the ability of AI models to be creative and generate art.*

Consequently, respondents were surveyed on their perspectives regarding the ability for AI models to demonstrate creativity and generate art - two philosophically-loaded concepts that are frequently related to intention and embodied experience, which AI may be considered to lack (Leong, 2025). Results are presented in Figure 8.3. Across both practitioners and the public, consensus is against AI models being capable of

CHAPTER 8. DO AS I SAY, NOT AS YOU DO

either of these feats. However, responses from practitioners on the topic of AI creativity are less unanimous, with only a small majority (32%) being strongly of the belief that such systems are incapable of creativity, compared to 74% of public respondents. An intra-group evaluation demonstrates that the discrepancy between the concepts of art and creativity is stronger in the general public, evaluating the artistic capabilities of AI significantly differently than its creative capabilities (Wilcoxon $W = 357.5, p = .038$), with 78.3% of respondents giving the same rating on both questions, whilst 9.0% rated creativity higher and 12.7% rated art higher. In contrast, practitioners make no such distinction ($W = 236.0, p = .803$), with 59.2% giving the same rating, 23.7% rating creativity higher and 17.1% rating art higher.

Whilst art is often seen as a highly human-centric concept that requires intention and meaning, creativity may also be defined as the ability to produce unique ideas (Dow, 2017). Consequently, technologies such as large language models are capable of rearranging sequences observed from training data to output novel linguistic strings (and therefore ideas). Practitioners may be more likely to consider such outputs as being creative, even if the ideas are amalgamations and abstractions of existing human thoughts contained in training corpora. Such comparatively shallow abilities have been referred to as pseudo-creativity (Runco, 2023). On the other hand, AI may be capable of core levels of creativity, yet popular architectures are fundamentally bounded in their creative limits (Cropley, 2025). Even within humans, there is debate as to whether creativity is truly novel, or the result of exposure to embodied experiences that constitute our training data (Lynn and Bassett, 2020).

Extensive work has demonstrated that people are more likely to rate content negatively when it is suggested that it has been created by AI rather than another human, regardless of tangible characteristics Bellaiche et al. (2023). This preference for human-created content over machine-created content is a simple extension of longstanding

patterns observed alongside industrialisation, such as the value placed behind handmade items versus the perceived coldness of machine-produced items (Degirmenci et al., 2025; Fan et al., 2024; Fuchs et al., 2015; Kruger et al., 2004).

8.7 Thought & Consciousness

Whilst once firmly in the remit of science fiction (Nader et al., 2024; Hermann, 2023; Hudson et al., 2023), the concept of artificial consciousness and the ability for machines to emulate internal cognitive processes such as thought is progressively being viewed as more viable. A former Google engineer made news by claiming that the LaMDA (Thoppilan et al., 2022) language model was sentient (Tiku, 2022). Whilst dismissed by the community at large, this presented one of the first notable moments where an in-domain practitioner with expertise made positive claims regarding the sentience of computational models. Previously, such claims were largely limited to non-experts whose beliefs about sentient machines may have been dismissed as a lack of understanding (Selwyn and Gallo Cordoba, 2022).

Implications for the consciousness of AI include needing to ensure that such models are not subject to undue suffering and other protections analogous to human-rights Metzinger (2021); Schwitzgebel and Garza (2020); Laukyte (2019). Furthermore, the consciousness of such models has further potential legal implications pertaining to the allocation of blame and responsibility for the actions made by such technologies (Veluwenkamp, 2025; Noh, 2025).

Survey respondents were queried on whether AI models are, or will become, capable of thought and consciousness, and the consequent real-world implications regarding legal protections for AI analogous to human rights. The results in Figure 8.4 demonstrate that by and large, both practitioners and the public dismiss the potential sentience of

CHAPTER 8. DO AS I SAY, NOT AS YOU DO

such models, with approximately half of all respondents strongly rejecting both claims. However, only 42% of practitioners strongly rejected the claim that AI models will ever require human-rights style protections, whilst 17% and 21% gave a weaker rejection or were neutral on the topic, respectively. Furthermore, for both questions and for both practitioners and the public, we see at least some respondents giving positive responses, demonstrating that the belief in artificial consciousness and the subsequent need for legal protections is not an entirely fringe belief.

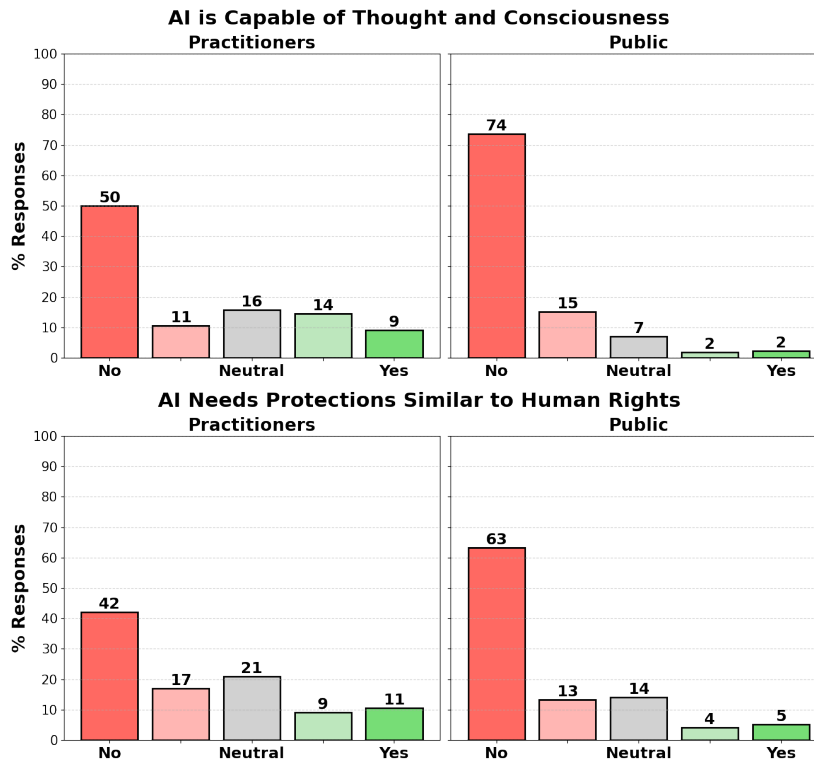


Figure 8.4: *Practitioner and public beliefs concerning the ability for AI systems to be capable of thought and consciousness, and the likelihood of requiring legal protections.*

8.8 Societal Impact

In this final section, we investigate the aspects of society that people believe are impacted by AI in a net positive or negative way. Figure 8.5 presents public and practitioner perspectives on the overall impact of AI systems on society at large, both now and in the future. The results show that the public have a much more negative view of the societal impact of AI, with a strong majority of respondents having negative views on the impact of AI at the time of response (83%), and longer term forecasts (77%), versus less than half of practitioners (46% and 51%, for the present and future, respectively).

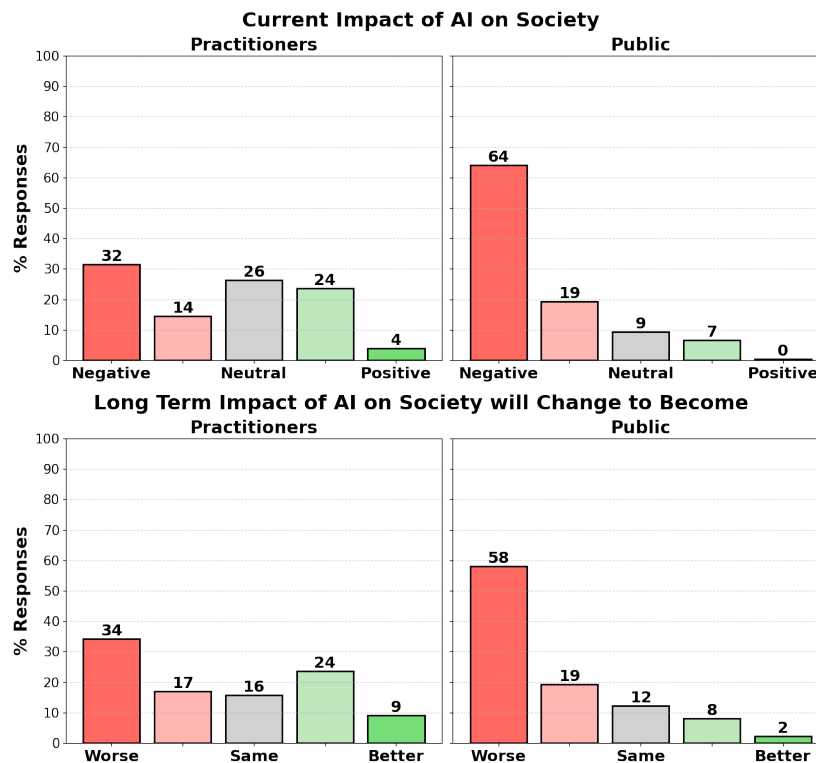


Figure 8.5: *Practitioner and public beliefs concerning the present and future societal impact of AI.*

Figure 8.6 presents a breakdown of the specific domains and industries where AI is considered to be a net positive or negative. Most strikingly, 56% of public respondents and 21% of practitioners believe that AI provides no net benefit to any of the presented

CHAPTER 8. DO AS I SAY, NOT AS YOU DO

areas. In contrast, 3% of public respondents take the optimistic stance that none of the proposed areas are impacted by AI in a net negative way. Whist raw percentages differ, the results for both net positive and net negative areas are comparable between both practitioners and the general public, demonstrating that specific areas of concern are shared across the surveyed population.

Of the most positively viewed domains, disagreement between the public and practitioners is highest for research, healthcare, education, entertainment, and creativity, with an approximately 2-4x higher proportion of practitioners presenting positive views of the potential impact of AI in these domains than public respondents.

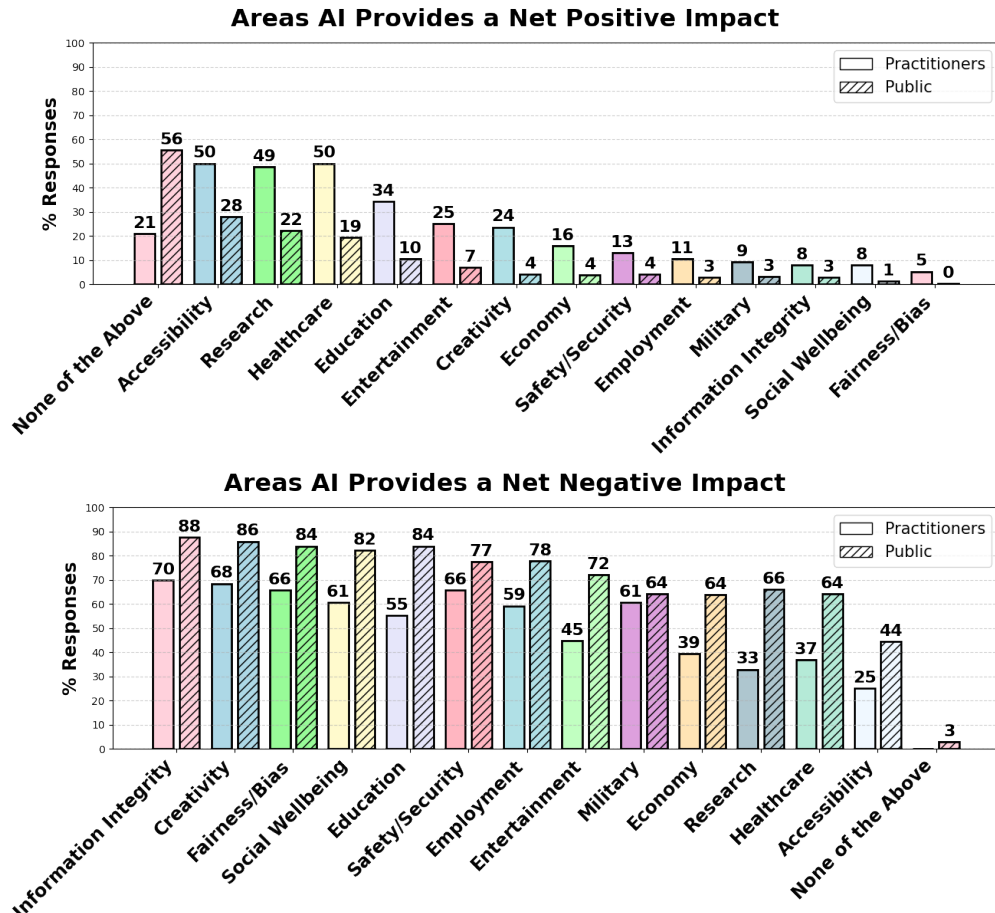


Figure 8.6: Practitioner and public beliefs concerning the subdisciplines of AI that present overall positive or negative impacts on society.

Overall, the finding that half of the surveyed practitioners view AI's societal impact to be net negative is concerning. One interpretation of this finding is that practitioners with this view may wish to remain involved in the field in order to help direct AI development towards positive societal utility, even if they view the domain as a whole as trending negatively. In the following section, we elaborate on some of the key areas of societal impact:

8.8.1 Accessibility

AI technologies can significantly lower barriers to partaking in activities that have traditionally required specialised skill sets and knowledge (Costa et al., 2024; Dessimoz and Thomas, 2024). Whilst the dawn of the internet presented a positive step in this direction, much specialised knowledge remained complex and difficult to understand. Consequently, the use of technologies such as large language models to summarise and explain concepts in a more accessible, simplified manner further reduces barriers. Barriers to other skills have likewise progressively been reduced, such as being able to create music or animations. However, reducing the skill level required to partake in activities has been shown to reduce a skill's perceived value, therefore garnering negativity from those who learnt with less assistance (Kruger et al., 2004). Additionally, the over-reliance on AI aids may generally reduce the ability levels of a population over time (Gerlich, 2025; Macnamara et al., 2024). Alongside lowering the overall barrier to entry for the general public, AI technologies also provide assistance to those with disabilities, for whom access to a given activity may have otherwise been impossible, rather than just requiring more effort (Brotosaputro et al., 2024).

8.8.2 Research

The use of AI in academic research has enabled the significant scaling of experiments, resulting in a faster rate of progress (Nolte and Tomforde, 2025; Gottweis et al., 2025; Lu et al., 2024). However, in specific research domains, the use of AI has increasingly presented challenges. For instance, the pre-print platform arXiv decided to halt the acceptance of survey papers and opinion-pieces due to an influx of low-quality AI-generated works (Boboris, 2025), whilst many academic venues report the use of AI by reviewers, usually without disclosure. This results in a pipeline where research works have their aims and hypotheses generated by AI, the methodology implemented with AI, and the results written into papers by AI, to be later reviewed by AI (Gottweis et al., 2025; Lu et al., 2024). This is compounded by the volume of research works published in fields such as machine learning and natural sciences (van der Zwaard et al., 2020; Pontis et al., 2017; Bornmann and Mutz, 2015), meaning that a significant number of these works go untouched by human eyes throughout their entire life cycle.

8.8.3 Healthcare

The survey responses show that the impact of AI on healthcare is considered positive by many people. For areas such as genome sequencing, early screening of life-threatening conditions, drug discovery, and other big-data-reliant healthcare fields, AI and machine learning are integral to meaningful, timely progress (Bajwa et al., 2021; Zhavoronkov et al., 2019). Additionally, the availability of AI doctors may help to alleviate strains on medical systems, including in less economically developed nations where medical expertise may not be readily available at the scale required for the population (Yousef and Schmollgruber, 2024). However, negative sentiment arises when considering the use of AI for final diagnoses and decision-making roles, such as determining the likelihood

of survival for someone on life support, without human intervention being required for the final decision.

8.8.4 Wellbeing

In terms of mental health and social wellbeing, there has been an increasing number of individuals turning to AI assistance for therapy and counselling services, owing to the relative affordability and the perceived anonymity of conversing with a language model (Cox et al., 2025; Lee et al., 2020). However, most implementations of AI for this task rely on third-party hosted instances of proprietary models, where sensitive data is shared with international corporations, rather than self-hosted open-source implementations that require substantial financial and educational investments to use (Maslej et al., 2025). Additionally, model debiasing and detoxification are not solved issues, resulting in instances where models have prompted vulnerable users to either make unsupported medical decisions, or take their own lives whilst in depressive episodes (Schoene and Canca, 2025; Judd et al., 2025; Arnaiz-Rodriguez et al., 2025).

8.8.5 Information Integrity

As the most negatively viewed impact of AI on society in our survey, information integrity refers to the reliability and trustworthiness of outputs presented by such technologies. Technologies such as large language models are able to facilitate bad actors in the creation and propagation of fake or misleading information at an unprecedented scale, largely going undetected by users (Williams et al., 2025; Radivojevic et al., 2024a; Chen and Shu, 2024). Such materials exist both as web pages and as bots acting as real social media users to promote anger and civil unrest (Radivojevic et al., 2024b). Further adding to the concern, whilst image and video evidence of an event was once considered

almost infallible, developments in the automatic generation of such mediums has slowly eroded the utility of such evidence (Iannuzzi et al., 2025; Pataranutaporn et al., 2025; Maras and Alexandrou, 2019). Combined with the growing distrust of governments and news media, false information is able to propagate faster than ever (Garry et al., 2024; Prats et al., 2024; Mittelstadt et al., 2023; Ognyanova et al., 2020). Furthermore, even within the scope of expected use, hallucination remains a prevalent concern, casting doubt on the factual accuracy of model outputs and requiring additional verification that many users skip (Lee et al., 2025; Gerlich, 2025). As such problems worsen and more content becomes model-generated, progress towards realising the dead internet theory increases (Walter, 2025; Muzumdar et al., 2025).

8.9 Conclusion

This work contributed further insights into the differences of opinion held by the individuals responsible for the development of artificial intelligence technologies (i.e., practitioners) and the members of public that constitute the majority of people impacted by such developments. Through a survey, we highlight key differences between both parties, particularly concerning opinions on copyright responsibilities, societal impact, and the philosophical concerns of art, creativity, and consciousness. Our results support the findings of prior work, empirically demonstrating that practitioners have more optimistic views toward AI technologies than the general public, yet both parties remain pessimistic overall. We argue that such differences in opinion between practitioners and the public are not irreconcilable, but highlight the need for the public to become more well-informed about the nature of AI technologies, as well as the need for practitioners to engage more actively with stakeholders outside of the technology industry and corporate interests to better understand the perspectives of society at large.

8.10 Limitations

We performed a survey of opinions on the development of artificial intelligence from the perspectives of practitioners and the general public. However, we were only able to capture a small subset of opinions from each of these parties. Owing to logistics and ethics approval procedures, we were unable to collect responses from individuals in rural communities across the globe with limited access to technology, in addition to the young (who were unable to participate in this work due to the 18+ age restriction), and the elderly. Due to the survey being presented only in English, we were additionally not able to receive responses from individuals who are not fluent in English. Whilst we did not enforce a geographical boundary on responses to the survey, the primary demographic represented in this work is that of Millennial Europeans, rather than a stratified sample of the world's true population for representativeness. We highly encourage research groups globally to perform further work in this vain (as well as longitudinal analyses of how perspectives evolve over time as the technology advances). Furthermore, the largest field from which public respondents were recruited from is technology (excluding ML/AI), constituting $\approx 30\%$ of respondents. We split technology workers ($N=64$) from the remainder of the public ($N=148$) and performed Mann-Whitney U tests, finding no significant difference between this subset and the public as a whole.

8.11 Appendix

8.11.1 Comparison to Existing Work

In the following section, we compare our findings to those reported in [Pew Research Center \(2025\)](#), where comparable lines of questioning are available. Regarding long term impact of AI (Figure 8.5), we observe less optimistic results, with only 33% of

practitioners and 10% of the public having favourable views, versus 17% of the public and 56% of the experts surveyed by the Pew Research Center. Furthermore, regarding specific areas of benefit (Figure 8.6), we observe no specific domain in which the majority of practitioners or the public believe AI will present an overall benefit. In contrast, [Pew Research Center \(2025\)](#) observe a majority of positive perspectives towards AI's effect on the economy and medical care by practitioners, whilst the public remains mostly pessimistic. However, our results echo the overall finding that practitioners are almost universally more optimistic than the public. However, regarding AI concerns, both the present work and [Pew Research Center \(2025\)](#) demonstrate that inaccurate information (i.e., information integrity) is one of the most pressing concerns.

8.11.2 Original Question Phrasing & Statistical Testing

The exact question phrasing as presented to participants is presented in Table 8.1, alongside the results of statistical significance tests regarding differences between practitioner and public responses.

8.11.3 Additional Demographic Information

Additional information regarding the specific fields that public respondents came from is presented in Figure 8.7.

CHAPTER 8. DO AS I SAY, NOT AS YOU DO

Figure	Original Question	Practitioner	Public	U	p_{BH}	r
-	How familiar are you with the general idea of AI and ML and their applications?	-	-	-	-	-
-	How familiar are you with technical aspects of AI/ML (e.g., neural networks, deep learning)?	-	-	-	-	-
8.1	Which AI-powered tools or applications have you used in your own free time?	-	-	-	-	-
8.1	Which AI-powered tools or applications have you used to aid in your work/employment?	-	-	-	-	-
8.1	Was using the AI-powered tools selected above in your work/employment mandated/recommended by your employer?	-	-	-	-	-
8.2	Do you think that AI/ML companies should be able to train models on publicly available data (e.g., social media posts, PDFs, YouTube videos, etc.) without explicit informed consent?	2.21	1.39	10,205.0	< 0.001	0.267
8.3	Do you believe that AI is capable of generating art?	2.49	1.61	10,653.5	< 0.001	0.322
8.3	Do you think that AI is capable of thinking and being conscious?	2.22	1.44	10,407.5	< 0.001	0.292
8.4	Do you believe that AI/ML systems are capable of being creative?	2.53	1.50	11,738.0	< 0.001	0.457
8.4	Do you think that an AI system will ever need to be granted protections similar to human-rights?	2.29	1.75	9,927.5	< 0.001	0.232
8.5	In its current state, how do you perceive the impact of AI on society?	2.54	1.60	11,400.0	< 0.001	0.415
8.5	How do you perceive the impact of AI on society to change long-term (e.g., the next 10 years)?	2.57	1.77	10,628.0	< 0.001	0.319
8.6	In which areas do you believe AI provides an overall positive impact?	-	-	-	-	-
8.6	In which areas do you believe AI provides an overall negative impact?	-	-	-	-	-

Table 8.1: *Original questions as posed to survey respondents. Two-sided Mann-Whitney U Test results comparing practitioner and public responses are provided, alongside mean Likert scores and absolute rank-biserial correlation (r) for effect size. p -values are adjusted using the Benjamini-Hochberg procedure across all 12 evaluation metrics. Practitioner $N = 76$, Public $N = 212$.*

CHAPTER 8. DO AS I SAY, NOT AS YOU DO

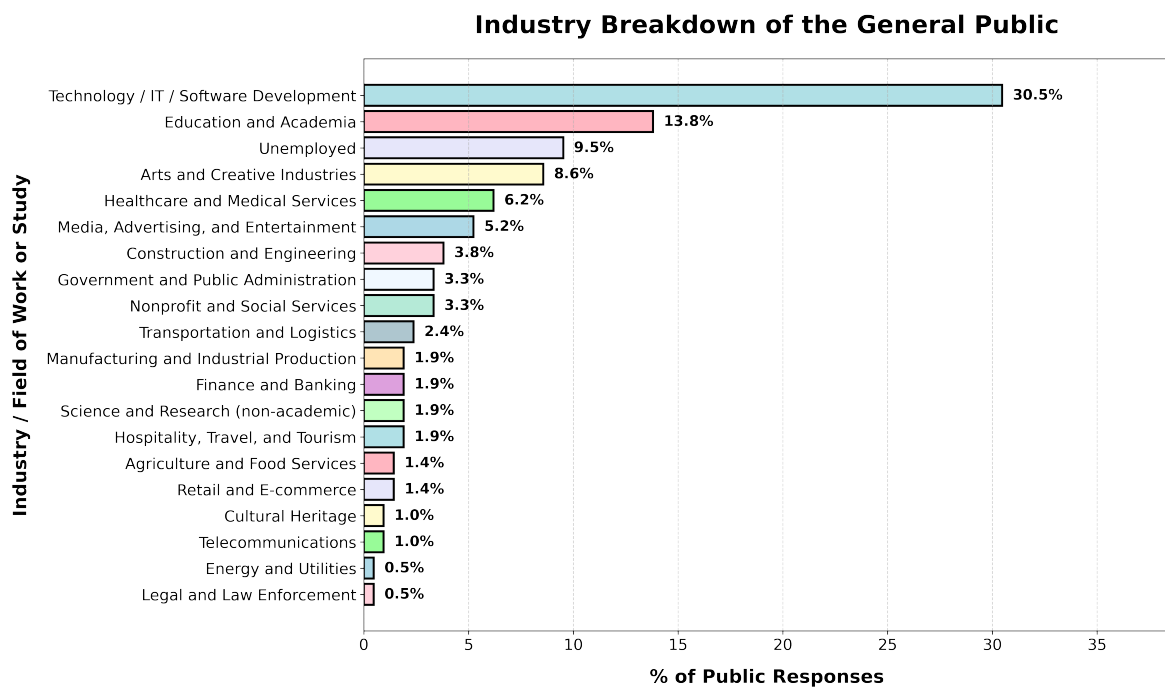


Figure 8.7: *Breakdown of industries within which public respondents operate.*

Part VII

Conclusion

Chapter 9

Thesis Conclusions

In the following sections, the overall findings and implications of this thesis are outlined and discussed, including potential impacts and promising directions for future research in the area of phonetically and phonologically motivated language understanding. First, each research question posed in Chapter 1 is presented alongside a recap of the relevant chapter, elucidating how each research question has been addressed.

9.1 Research Questions Discussion

RQ0. *To what extent can large language models and vision-language models bridge the inherent information loss between orthographic text and spoken language to generate, reason about, and visually interpret phonetic, phonological, and acoustic phenomena?*

Regarding the overarching primary research question, this thesis has demonstrated mixed abilities within language models to process such phenomena. It was observed in Chapter 3 and Chapter 4 that generative models can learn foundational characteristics of tongue twisters (e.g., alliteration and general sound repetition), yet applying knowl-

CHAPTER 9. THESIS CONCLUSIONS

edge of phonetics to contextual tasks such as humour explanation in Chapter 5 was demonstrably prone to hallucinations, particularly where different grapheme sequences are used to represent the same sound. Similarly, more abstract tasks using VLMs such as the assessment of sound symbolism in Chapter 6 and spectrogram reading in Chapter 7 were shown to be even more difficult for models to understand. In the remainder of the section, detailed discussions of each publication-specific research question is provided.

RQ1. To what extent do demographic factors affect the interpretation of the intensity of humour, irony, and sarcasm; how transparently are human evaluation procedures reported in existing works; and how can the reporting of human evaluation be standardised to account for this subjectivity?

Chapter 2 outlined the demographic factors that impact the interpretation of humour, irony, and sarcasm. These variables included the skill-based factors of language proficiency and education, in addition to complex sociocultural factors such as age and political leaning. Whilst some instances of subjective interpretation lead to innocuous miscommunications, other factors are demonstrated to be the difference between humorous interaction and potential offence, a distinction also explored in Survey II regarding demographic-aware models. Regarding the transparency of reporting human evaluation, Chapter 2 included a critical survey of reporting practices regarding evaluator panel demographics and logistical factors such as recruitment methods and compensation. Overall, Chapter 2 revealed that the reporting transparency for demographic variables was poor, with fewer than half of the surveyed papers reporting any evaluator characteristics. Regarding logistical factors, the majority of works reported the recruitment strategies used to gain evaluators, yet only 30% of works disclosed how evaluators were compensated. However, the work contained in Chapter 2 was performed at the start of this research project, before changes such as the Responsible NLP

CHAPTER 9. THESIS CONCLUSIONS

Checklist being a mandatory inclusion for submissions to Automatic Rolling Review (ARR). Consequently, modern reporting practices may have improved. The observed impact of demographic variables and the lack of reporting transparency resulted in Chapter 2 presenting criteria for reporting human evaluation. Whilst the reporting transparency of more recent works has not been verified, this set of standardised criteria informed the reporting of human evaluation procedures in subsequent thesis chapters.

RQ2. *To what extent can grapheme-based, pre-trained language models learn phonological constraints from human-authored samples to generate tongue twisters?*

Chapter 3 presented the task of topic-to-twister generation, using only orthographic text to fine-tune a range of small, pre-trained language models. To achieve this, a human-authored dataset of tongue twisters was collected (TwistList 1.0), presenting the largest known single collection at the time of publication. Whilst some of the fine-tuned models were able to learn general characteristics of tongue twisters, producing moderately coherent outputs, the overall findings revealed that these small models, trained on a relatively small-scale dataset (2.1k entries), largely failed to produce high-quality outputs. One observed trend was a tendency to rely exclusively on single-word repetition, suggesting that the lexical overlap present in tongue twisters overshadowed grammatical and semantic cohesion. Furthermore, to assess the quality of generations, Chapter 3 introduced two phoneme-based metrics centred on overlap. Whilst indicative of the observed pure sound repetition, the scores of such metrics were observed to be too coarse in their categorisation of different phonemes, treating them all as equally dissimilar. This resulted in overlooking desirable traits such as the alternation between similar sounds. On the other hand, an early version of ChatGPT (GPT-3.5-Turbo) was shown to produce coherent tongue twisters, but relied heavily on performing minor

modifications to well-known formats.

RQ3. *To what extent does incorporating synthetic fine-tuning data and constrained decoding improve the ability of grapheme-based language models to generate tongue twisters, and what insights are provided by phonologically informed metrics?*

Following the limited success observed with using TwistList 1.0 to fine-tune pre-trained language models for the task of tongue twister generation, Chapter 4 aimed to take advantage of observations from Chapter 3, namely that larger models (that have both higher parameter counts and more extensive training data) could be used to expand the existing tongue twister dataset. To do so, Chapter 4 presented TwisterLister, a phonologically informed pipeline using a large language model (LLM) for generating synthetic tongue twister examples for topic-to-twister and style-transfer tasks, resulting in TwistList 2.0 (a dataset 8x larger than TwistList 1.0). This enabled RQ3 to be addressed by fine-tuning pre-trained language models on various volumes of in-domain data. Whilst tongue twister quality remained inconsistent, training volume was demonstrated to have an overall positive impact on generation quality, though with diminishing returns. Furthermore, a phoneme-aware constrained decoding module was introduced (PACD), which applied a cascade of constraints to the output of an autoregressive language model. Though this, it was observed that successful tongue twister generation could be achieved without in-domain training data, by constraining the output of a general-purpose model. Despite this, all fine-tuned and constrained models were outperformed by both LLMs (i.e., ChatGPT) and humans. All of these findings were aided by the inclusion of new metrics based on phonemic edit distance. These metrics moved beyond the overlap-based focus of those presented in Chapter 3, accounting for phoneme similarity via phonological feature vectors. These metrics provided deeper

CHAPTER 9. THESIS CONCLUSIONS

insight, no longer unfairly rewarding degenerate patterns such as single-word repetition, by more accurately rewarding outputs that alternated between similar phonemes.

RQ4. *To what extent does joke typology affect the ability of large language models to explain humour, and how does the transition from sound-based wordplay to topical humour expose reasoning limits?*

Chapter 5 marked the transition from generating phonetically and phonologically motivated language to assessing its understanding by LLMs. Specifically, this chapter investigated the ability of models to accurately and comprehensively explain a range of joke types. To facilitate this, a novel dataset was introduced, combining jokes from existing sources and authoring gold-standard natural language explanations against which to compare model outputs. In relation to RQ4, joke type was seen to significantly impact the ability of models to provide accurate and complete explanations. In line with the focus of this thesis, sound-based heterographic puns were shown to be more difficult to explain than homographic puns. This was due to phonetic hallucination, where models identify that a piece of text is a pun, yet hallucinate the pair of words that have phonetic similarity. On the other hand, homographic puns rely on polysemy, which is better captured by word embeddings. Overall, however, topical jokes were shown to present the most difficulty. Such jokes routinely rely on abstract and ambiguous references to pop culture phenomena and real-world entities, resulting in hallucination from models generating explanations based on imperfect information. This effect was exacerbated by reasoning-focused models, which continued down a reasoning “rabbit hole”, with explanations becoming progressively less accurate over time.

RQ5. To what extent do large language models and vision language models reflect human psycholinguistic associations regarding sound symbolism?

Chapter 6 continued the investigation of LLMs’ understanding of phonetic and phonological patterns by replicating a range of psycholinguistic studies concerning sound symbolism with LLMs and VLMs, introducing the first use of multimodality within this thesis. Specifically, to address RQ5, Chapter 6 investigated shape and magnitude symbolism via the Kiki-Bouba and Mil-Mal tests, in addition to evaluating linguistic iconicity. To achieve this, the chapter introduced novel datasets for assessing shape and magnitude symbolism, combining pseudowords with unseen generated imagery and asking a panel of human annotators to determine ground truth associations. The results demonstrated that VLMs performed at approximately chance-level for assigning the same pseudoword to a visual entity as humans, regarding both shape and magnitude symbolism. However, once the models were informed of the nature of the task (e.g., that shape symbolism concerns a relationship between the sound of a word and the physical characteristics of the entity it describes), agreement levels were shown to increase for most models, particularly regarding magnitude symbolism. On the other hand, model size was shown to have an even clearer relationship to human agreement in the iconicity rating task, with large-scale proprietary models showing the highest alignment. Consequently, this chapter demonstrated the potential for language models to be used for in silico sound symbolism experiments.

CHAPTER 9. THESIS CONCLUSIONS

RQ6. *To what extent can vision language models synthesise parametric knowledge of phonetics and phonology to correctly interpret visual representations of speech?*

Chapter 7 furthered the use of multimodal inputs to test the abilities of VLMs, tasking models with interpreting spoken words from spectrogram and waveform inputs. To do so, this chapter introduced a novel dataset for the task of spectrogram and waveform interpretation that included standardised visualisations of a range of synthetic voices producing common English words in isolation. From this, a multiple-choice question dataset was produced, with distractors selected based on their phonemic edit distance from the ground truth label. Regarding RQ6, both fine-tuned and zero-shot models were demonstrated to struggle immensely on this task, achieving chance-level accuracy. However, as in Chapter 4, phoneme edit distance provided a more detailed understanding, suggesting that models had learnt some phonetic knowledge, selecting options that were closer to the ground truth. Overall, this chapter presents a phonetically and phonologically motivated task for challenging both the linguistic knowledge of VLMs, as well as their ability to interpret esoteric data visualisations.

RQ7. *To what extent do the perspectives of practitioners and the general public differ towards the development and use of artificial intelligence technology, encompassing factors such as creativity, consent, and wellbeing?*

Chapter 8 presented a shift in direction for the thesis, focusing on the sentiment and perspectives of practitioners and the general public in the broader field of artificial intelligence (AI). Motivated by the transition from creative generation (Chapter 3 and Chapter 4) to the analysis of model understanding of phonetically and phonologically motivated language (Chapter 5), this chapter investigated whether the development

CHAPTER 9. THESIS CONCLUSIONS

of such technologies is in society’s best interest. Consequently, Chapter 8 addressed RQ7 by presenting a retrospective analysis of the desirability of the technologies investigated in this thesis, from creative generation to in silico experimentation. The results showed that both practitioners and the general public have pessimistic views regarding the present and future societal impact of AI. Information integrity and the reliability of model outputs were of particular concern, in addition to the detrimental impact on human creativity and employment. In terms of positive impacts, almost half of the practitioners surveyed believe that AI provides a net positive to accessibility, research, and healthcare, whilst over half of the public respondents believe that AI provides no net benefit to any of the presented areas. Overall, the general public was demonstrated to be more pessimistic towards AI development than practitioners, who themselves were largely negative. However, broader areas of general concern were shared.

9.2 Impact

This thesis has motivated the investigation of phonetic and phonologically motivated language forms, demonstrating that such knowledge is relevant to a broad range of NLP domains including creative language generation (Chapter 3 and Chapter 4) and explanation (Chapter 5), the psycholinguistic alignment of models for in silico psycholinguistic experimentation (Chapter 6), and esoteric knowledge retrieval (Chapter 5 and Chapter 7), presenting novel datasets for all of the presented tasks. Furthermore, Chapter 2 brought attention to the lack of reporting transparency present in humour, sarcasm, and irony research regarding human evaluation of model outputs. It is hoped that future work in similar domains will take into account best practices for transparent, reproducible reporting, following the evaluation reporting standards presented. Finally,

CHAPTER 9. THESIS CONCLUSIONS

Chapter 8 explored public and industry concerns over the development of AI technology. It is hoped that this work will generate impact by encouraging academics, developers and the general public to better engage with one another, ensuring AI development becomes a net benefit for society. It is further hoped that current and future researchers consider the unintended negative impacts of their research, even in tasks that appear innocuous in isolation.

Materially, the humour explanation work presented in Chapter 5 contributed insights that directed the performance of [Wang et al. \(2025b\)](#) in terms of rating systems and taxonomy creation, whilst the model inference infrastructure used to generate joke explanations (designed by a co-author) was demonstrated to be of further use in my co-author’s works. Additionally, the design-laden posters presented at multiple conferences (see Figure 2) garnered significant attention from conference attendees, frequently including members of the community that primarily worked in different domains, yet who were attracted to discuss the work owing to the graphic design and unique research topics, demonstrating that the marketing of a work is often almost as important as the work itself. An individual also informed me that my writing style (particularly in the work presented within Chapter 4) and background as a non-computer scientist inspired them to seek a PhD position in NLP themselves.

At the time of writing, the works authored within this PhD have gained 259 citations, demonstrating sizeable readership of the published works.

9.3 Future Work

This thesis has presented a range of novel datasets for myriad tasks concerning phonetic and phonological understanding in language models, including TwistList and TwistList 2.0 for tongue twister generation (Chapter 3 and Chapter 4), a dataset of jokes with gold standard natural language explanations (Chapter 5), a dataset for performing psycholinguistic experiments concerning shape and magnitude symbolism with VLMs (Chapter 6), and a dataset for tasking VLMs with interpreting spectrograms and waveforms (Chapter 7). In several instances, this thesis investigated the zero-shot abilities of models to perform novel tasks, providing insights into their out-of-the-box capabilities with canonical prompts. As a result, future work is able to explicitly focus on improving the standard performance of models in these tasks, including through training models on phonetically transcribed text, integrating retrieval augmented generation to improve performance on tasks that require esoteric knowledge, and general prompt engineering. Furthermore, large-scale reasoning models, which are becoming progressively more accessible, may outperform many of the models studied within this thesis. Similarly, true speech-to-speech models that understand the characteristics of speech without relying on lossy textual representations are starting to emerge in the research landscape, building upon existing multimodal models that use text as an intermediary between speech recognisers and speech synthesis engines. Finally, Chapter 8 demonstrated that practitioners and the public are misaligned in many areas concerning AI use and development, whilst both being pessimistic towards the present and future impact of such technologies on society. It is for this reason that future work in the domain should focus on the understanding of subjective and creative language forms, rather than the direct generation of such language, either through human-in-the-loop paradigms or by focusing on practical tasks such as explanation.

In the following subsection, more specific future directions are presented for the domain of phonetically motivated language understanding, followed by promising directions for future work in the secondary research area explored by this thesis of computational humour.

9.3.1 Phonetically Motivated Language Understanding

Linguistic Bias. As a research domain, a pressing concern remains the inadvertent promotion of specific linguistic characteristics over others (Fleisig et al., 2024b; Ryan et al., 2024; Joshi et al., 2020). Whilst accent robustness is a common concern in areas such as automatic speech recognition (Yong et al., 2025; Markl and McNulty, 2022), much research in phonetically and phonologically grounded language generation fails to address these concerns. For instance, most G2P systems are trained to represent high-resource, idealised language varieties (e.g., Southern Standard British English or General American), and gold-standard resources such as the CMU Pronouncing Dictionary remain static. As a result, the understanding of phonetically and phonologically informed language overlooks differences in phoneme inventories and articulatory tendencies of non-standard, non-prestige varieties, resulting in language technologies that do not adequately serve such communities. Outside of accent variation, similar representational limitations are observed across different languages. Whilst English is known for having opaque orthography, the sheer volume of available data (including data relating to phonetic and phonological characteristics) facilitates implicit learning to a much greater extent than would be expected from a language that is both low-resource and orthographically opaque, in contrast to languages with transparent orthography and lower levels of inter-speaker variation (Marjou, 2021; Borleffs et al., 2017). To this end, research must continue in the direction of the wider NLP trend towards customisable,

CHAPTER 9. THESIS CONCLUSIONS

personalised models that work for a language community or individual, rather than against (Guan et al., 2025).

Tokenisation Approaches. Whilst existing research has demonstrated that phonetic and phonological knowledge can be implicitly gained by LLMs with significant scale increases in computational power and data, lower resource application settings require more deft approaches. Of these, the limits of subword tokenisation for many of the tasks presented in this thesis are evident, with the broadness of tokenisation resulting in a loss of understanding concerning the relationship between orthography and pronunciation. Several works have found success in applying alternative tokenisation approaches, such as character-level tokenisers as used in ByT5 (Loakman et al., 2025b; Rahmati and Sameti, 2024; Zhu et al., 2022).

Knowledge Integration. Current foundation models demonstrate remarkable understanding of a wide range of complex topics. However, despite information on phonetics, phonology, and orthography being widely available for high-resource languages, many of the processing difficulties observed in LLMs indicate an inability to integrate this knowledge in practice (Shojaee et al., 2025; Loakman et al., 2025a; Jiang et al., 2025). Consequently, research toward active reasoning using latent knowledge remains fruitful.

Audio-Enabled Models. Finally, a pivotal future direction concerns LLMs explicitly trained on speech representations, such as audio-enabled LLMs or speech-to-speech understanding models (Cui et al., 2025; Wang et al., 2025a; Yang et al., 2024c; Shan et al., 2025). Whilst such models remain in their relative infancy, these approaches alleviate the need for labour-intensive handcrafted phonetic and phonological knowledge, such as gold-standard transcriptions, whilst additionally avoiding the lack of representational fidelity that is inherent in learning sound-based characteristics from

broad phonemic transcriptions or directly from orthography. However, accent diversity concerns remain with such models, with many of the most popular speech datasets being relatively homogenous in language variety (Markl and McNulty, 2022). Multimodal prompting approaches demonstrate promise in facilitating phonetic and phonological understanding, lending support to future advancements as multimodal fusion improves and the handling of speech by LLMs transitions from speech recognition front-ends for text-based understanding models to speech-to-speech models.

9.3.2 Computational Humour Processing

Generative tasks in the area of computational humour present a range of challenges owing to aspects such as the subjective and potentially offensive nature of humour and the ethics of generating any creative language form.

9.3.3 Demographic-Aware Humour Generation

In following the trends established by works such as Sun et al. (2022b) and Garimella et al. (2020), an essential primary focus of future research should be audience-tailored, demographic-aware, and contextually nuanced humour generation that is able to better address the preferences of particular end-users. Owing to the wide gamut of topics that humour may be found in, such approaches would aid in decreasing the risk of generating material that is considered offensive to a given end-user. Consequently, a promising direction is to explicitly incorporate human preference alignment techniques such as RLHF (Ouyang et al., 2022; Stiennon et al., 2020; Christiano et al., 2017), DPO (Rafailov et al., 2023), and PPO (Schulman et al., 2017) to the task of humour generation.

CHAPTER 9. THESIS CONCLUSIONS

Additionally, the developing area of perspectivist approaches in NLP (Fleisig et al., 2024a; Valette, 2024; Abercrombie et al., 2022) provide a route through which to consider individual preferences for subjective domains such as humour. Progress to this end is demonstrated by Casola et al. (2024) and Frenda et al. (2023), who present perspective-aware datasets for irony processing (a phenomenon highly related to humour). The types of humour and jokes that are most often required to be explained to someone are those that are ambiguous, nuanced, and frequently focused on sensitive topics. Whilst the potential sensitivity of a joke topic can have an intensifying impact on the level of humour perceived by the *intended* audience, it acts as part of a risk-to-reward trade-off, likewise increasing the risk of offence if presented to the wrong audience (McGraw and Warren, 2010).

9.3.4 Human-in-the-Loop Humour Generation

Humour is a quintessential demonstration of human intelligence and creativity. As such, the development of computational models for the generation of such content poses the risk of increasing the anthropomorphism of machine learning models - something that is at times desirable, but also disproportionately impacts vulnerable users of these technologies (Shevlin, 2024). In addition to this, the creation of (intentional) humour is a creative endeavour. The generation of creative language forms with computational models has the potential to have severe negative impacts on the arts and creative industries as a whole, and reduce the outlets available to people to realise financial gains from their own creativity. Whilst systems such as Witscript (Toplyn, 2023) were developed by real-world comedy writers, such examples are the exception to the rule. As a result, future work should focus increasingly on human-in-the-loop approaches (Wu et al., 2022) to humour generation, requiring meaningful contributions from the end-user, or combining elements of humour generation and explanation to develop systems for

joke workshoping, offering valuable feedback and direction on human-authored humour, rather than generating jokes wholesale.

9.3.5 Complex Humour Explanation

Humour explanation is an important task for assessing the verbal and common-sense reasoning abilities of models (particularly LLMs), yet is a scarcely researched domain. If the end goal is to imbue systems with human-level reasoning abilities, being able to explain humour is fundamental. Particular focus should be given to context-sensitive, esoteric humour such as topical jokes. Such jokes present a rich area for aiding in the development of advanced information retrieval systems that are able to work with ambiguous topics such as the references to world events and pop culture phenomena found in complex jokes, furthering the work presented in Chapter 5.

9.3.6 Intent-Aware Humour Explanation

Whilst humour explanation is a challenging task with a real-world benefit of reducing communication barriers between people from different backgrounds, it is also an area with significant ethical considerations. Another potential future direction for research in the area of humour explanation is the development of intent-aware models (Ma et al., 2025; Park et al., 2024). Such approaches would attempt to model the characteristics and intent of the author based on available information, such as prior language use and known/inferred demographic variables, before attempting to explain potentially humorous language use. Not only could this information aid in inferring references to ambiguous content, improving reasoning and relevant document retrieval, it would also aid in overcoming the undesirable effect of legitimising the instances where “*it’s just a joke!*” is used as a guise for spreading hatred through offensive material (Ford

CHAPTER 9. THESIS CONCLUSIONS

et al., 2015; Brommage, 2015; Hodson et al., 2010). Such approaches have extensive applications in online communication venues such as social media and in handling interpersonal conflict. Whilst individual preferences and beliefs as to what topics are possible to derive humour from remain highly subjective (Smuts, 2010; Gaut, 1998), improved modelling of the underlying intent would help identify instances where offence was likely caused accidentally.¹

¹We of course do not advocate for the extreme version of such systems in practice, whereby individuals are effectively accused of thought-crime due to an intent assigned to an author via a computational model.

Bibliography

Almas Abdibayev, Allen Riddell, and Daniel Rockmore. 2021. BPoMP: The benchmark of poetic minimal pairs – limericks, rhyme, and narrative coherence. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1–9, Held Online. INCOMA Ltd.

Gavin Abercrombie, Valerio Basile, Sara Tonelli, Verena Rieser, and Alexandra Uma, editors. 2022. *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*. European Language Resources Association, Marseille, France.

Gavin Abercrombie, Amanda Cercas Curry, Mugdha Pandya, and Verena Rieser. 2021. Alexa, Google, Siri: What are your pronouns? gender and anthropomorphism in the design and perception of conversational assistants. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 24–33, Online. Association for Computational Linguistics.

Ekaterina Abramova and Raquel Fernández. 2016. Questioning arbitrariness in language: a data-driven study of conventional iconicity. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 343–352, San Diego, California. Association for Computational Linguistics.

BIBLIOGRAPHY

- Ibrahim Abu Farha, Silviu Vlad Oprea, Steven Wilson, and Walid Magdy. 2022. SemEval-2022 task 6: iSarcasmEval, intended sarcasm detection in English and Arabic. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 802–814, Seattle, United States. Association for Computational Linguistics.
- Ibrahim Abu Farha, Wajdi Zaghouani, and Walid Magdy. 2021. Overview of the WANLP 2021 shared task on sarcasm and sentiment detection in Arabic. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 296–305, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.
- Advance Local Media LLC v. Cohere. 2025. *Advance Local Media LLC et al. v. Cohere Inc.*, no. 1:25-cv-01305, (s.d.n.y. feb. 13, 2025). United States District Court for the Southern District of New York (S.D.N.Y.).
- Rajat Agarwal and Katharina Kann. 2020. Acrostic poem generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1230–1240, Online. Association for Computational Linguistics.
- Siddhant Agarwal, Shivam Sharma, Preslav Nakov, and Tanmoy Chakraborty. 2024. MemeMQA: Multimodal question answering for memes via rationale-based inferencing. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5042–5078, Bangkok, Thailand. Association for Computational Linguistics.
- Gati Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning, ICML’23*. JMLR.org.
- Qihang Ai, Jiafan Li, Jincheng Dai, Jianwu Zhou, Lemao Liu, Haiyun Jiang, and Shuming Shi. 2024. Advancement in graph understanding: A multimodal benchmark

BIBLIOGRAPHY

- and fine-tuning of vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7485–7501, Bangkok, Thailand. Association for Computational Linguistics.
- D Akila and C Jayakumar. 2014. Semantic similarity- a review of approaches and metrics. *International Journal of Applied Engineering Research*, 9:27581–27600.
- Elaf Alhazmi, Quan Z. Sheng, Wei Emma Zhang, Munazza Zaib, and Ahoud Alhazmi. 2024. Distractor generation in multiple-choice tasks: A survey of methods, datasets, and evaluation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14437–14458, Miami, Florida, USA. Association for Computational Linguistics.
- Morris Alper and Hadar Averbuch-Elor. 2023. Kiki or bouba? sound symbolism in vision-and-language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 78347–78359. Curran Associates, Inc.
- Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. The (un)suitability of automatic evaluation metrics for text simplification. *Computational Linguistics*, 47(4):861–889.
- Amazon Employees for Climate Justice. 2025. AI Open Letter: We’re the Workers Who Develop, Train, and Use AI, So We Have a Responsibility to Intervene.
- Jacopo Amidei, Paul Piwek, and Alistair Willis. 2019. The use of rating and Likert scales in natural language generation human evaluation tasks: A review and some recommendations. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 397–402, Tokyo, Japan. Association for Computational Linguistics.

BIBLIOGRAPHY

- Miriam Amin and Manuel Burghardt. 2020. A survey on approaches to computational humor generation. In *Proceedings of the 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 29–41, Online. International Committee on Computational Linguistics.
- Andersen v. Stability AI. 2024. *Andersen et al. v. Stability AI Ltd. et al.*, no. 3:23-cv-00201 (n.d. cal. jan. 13, 2023). United States District Court for the Northern District of California.
- Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, et al. 2023. Gemini: A family of highly capable multimodal models.
- Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, Zhihua Wei, Yao Qian, Jinyu Li, and Furu Wei. 2022a. SpeechT5: Unified-modal encoder-decoder pre-training for spoken language processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5723–5738, Dublin, Ireland. Association for Computational Linguistics.
- Xiao Ao, Danae Sanchez Villegas, Daniel Preotiuc-Pietro, and Nikolaos Aletras. 2022b. Combining humor and sarcasm for improving political parody detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1800–1807, Seattle, United States. Association for Computational Linguistics.
- Angelina Aquino, Joshua Lijandro Tsang, Crisron Rudolf Lucas, and Franz de Leon. 2019. G2p and asr techniques for low-resource phonetic transcription of tagalog, cebuano, and hiligaynon. In *2019 International Symposium on Multimedia and Communication Technology (ISMAC)*, pages 1–5.

BIBLIOGRAPHY

- Dawit Mureja Argaw, Mattia Soldan, Alejandro Pardo, Chen Zhao, Fabian Caba Heilbron, Joon Son Chung, and Bernard Ghanem. 2024. Towards automated movie trailer generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7445–7454.
- Aristotle. 1902. *Poetics*. Macmillan and Co., London. First composed c. 335 BCE.
- Adrian Arnaiz-Rodriguez, Miguel Baidal, Erik Derner, Jenn Layton Annable, Mark Ball, Mark Ince, Elvira Perez Vallejos, and Nuria Oliver. 2025. Between help and harm: An evaluation of mental health crisis handling by llms.
- Arian Askari, Mohammad Aliannejadi, Chuan Meng, Evangelos Kanoulas, and Suzan Verberne. 2023. Expand, highlight, generate: RL-driven document generation for passage reranking. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10087–10099.
- Salvatore Attardo. 2008. *A primer for the linguistics of humor*, pages 101–156. De Gruyter Mouton, Berlin, New York.
- Authors Guild Inc. v. Google Inc. 2015. Authors guild, inc. v. google inc., no. 13-4829-cv, (2d cir. oct. 16, 2015). United States Court of Appeals for the Second Circuit. Decided Oct. 16, 2015.
- Peng Bai, Meizhen Zheng, and Xiaodong Shi. 2025a. A survey of singing voice synthesis. In *Proceedings of the 10th Conference on Sound and Music Technology*, pages 19–30, Singapore. Springer Nature Singapore.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo

BIBLIOGRAPHY

- Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025b. Qwen2.5-vl technical report.
- Junaid Bajwa, Usman Munir, Aditya Nori, and Bryan Williams. 2021. Artificial intelligence in healthcare: transforming the practice of medicine. *Future Healthcare Journal*, 8(2):e188–e194.
- Ashwin Baluja. 2025. Text is not all you need: Multimodal prompting helps LLMs understand humor. In *Proceedings of the 1st Workshop on Computational Humor (CHum)*, pages 9–17, Online. Association for Computational Linguistics.
- Scott Barnett, Stefanus Kurniawan, Srikanth Thudumu, Zach Brannelly, and Mohamed Abdelrazek. 2024. Seven failure points when engineering a retrieval augmented generation system. In *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering - Software Engineering for AI*, CAIN '24, page 194–199, New York, NY, USA. Association for Computing Machinery.
- Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. 2025. τ^2 -bench: Evaluating conversational agents in a dual-control environment.
- Daphne Barretto, Julianne LaChance, Emanuelle Burton, and Soohyun Nam Liao. 2021. Exploring why underrepresented students are less likely to study machine learning and artificial intelligence. In *Proceedings of the 26th ACM Conference on Innovation and Technology in Computer Science Education V. 1*, ITiCSE '21, page 457–463, New York, NY, USA. Association for Computing Machinery.
- Antonia Bartulović and Dóra Paula Váradi. 2024. University of split and university of malta (team ab&dpv) at the clef 2024 joker track: From 'lol' to 'mdr' using artificial intelligence models to retrieve and translate puns. In *Working Notes of the Conference*

BIBLIOGRAPHY

- and Labs of the Evaluation Forum (CLEF 2024)*, volume 3740 of *CEUR Workshop Proceedings*. CEUR-WS.org.
- Bartz v. Anthropic PBC. 2025. Bartz v. anthropic pbc., no. 3:24-cv-05417-wha, (n.d.c. jun. 6, 2025). United States District Court for the Northern District of California (N.D.C.).
- Nancy Bell and Salvatore Attardo. 2010. Failed humor: Issues in non-native speakers’ appreciation and understanding of humor. *Intercultural Pragmatics*, 7(3):423–447.
- Lucas Bellaiche, Rohin Shahi, Martin Harry Turpin, Anya Ragnhildstveit, Shawn Sprockett, Nathaniel Barr, Alexander Christensen, and Paul Seli. 2023. Humans versus ai: whether and why we prefer human-created compared to ai-created artwork. *Cognitive Research: Principles and Implications*, 8(1):42.
- Jonas Belouadi and Steffen Eger. 2023. ByGPT5: End-to-end style-conditioned poetry generation with token-free language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7364–7381, Toronto, Canada. Association for Computational Linguistics.
- Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.
- Abbas Bertina, Shahab Beirami, Hossein Biniazian, Elham Esmaeilnia, Soheil Shahi, and Mahdi Pirnia. 2025. Bridging the gap: An intermediate language for enhanced and cost-effective grapheme-to-phoneme conversion with homographs with multiple pronunciations disambiguation. *arXiv preprint arXiv:2505.06599*.
- James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey

BIBLIOGRAPHY

- Chu, Yunxin Jiao, and Aditya Ramesh. 2023. Improving image generation with better captions.
- Kim Binsted. 1996. *Machine humour: an implemented model of puns*. Ph.D. thesis, University of Edinburgh.
- Kim Binsted and Graeme Ritchie. 1994. An implemented model of punning riddles. In *Proceedings of the Twelfth AAAI National Conference on Artificial Intelligence*, AAAI’94, page 633–638. AAAI Press.
- Alan W Black, Kevin Lenzo, and Vincent Pagel. 1998. Issues in building general letter to sound rules. In *3rd ESCA/COCOSDA Workshop on Speech Synthesis (SSW 3)*, pages 77–80.
- Dawn G. Blasko, Victoria A. Kazmerski, and Shariffah Sheik Dawood. 2021. Saying what you don’t mean: A cross-cultural study of perceptions of sarcasm. *Canadian Journal of Experimental Psychology / Revue canadienne de psychologie expérimentale*, 75(2):114–119.
- Leonard Bloomfield. 1933. *Language*. Holt, New York.
- Kat Boboris. 2025. Attention authors: Updated practice for review articles and position papers in arxiv cs category.
- Nikolay Bogoychev and Pinzhen Chen. 2023. Terminology-aware translation with constrained decoding and large language model prompting. In *Proceedings of the Eighth Conference on Machine Translation*, pages 890–896, Singapore. Association for Computational Linguistics.

BIBLIOGRAPHY

- Matthew Bone, Eugenia González Ehlinger, and Fabian Stephany. 2025. Skills or degree? the rise of skill-based hiring for ai and green jobs. *Technological Forecasting and Social Change*, 214:124042.
- Elisabeth Borleffs, Ben A M Maassen, Heikki Lyytinen, and Frans Zwarts. 2017. Measuring orthographic transparency and morphological-syllabic complexity in alphabetic orthographies: a narrative review. *Reading and Writing*, 30(8):1617–1638.
- Lutz Bornmann and Rüdiger Mutz. 2015. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, 66(11):2215–2222.
- Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.
- Philipp Brauner, Alexander Hick, Ralf Philipsen, and Martina Zieffle. 2023. What does the public think about artificial intelligence?—a criticality map to understand bias in the public perception of ai. *Frontiers in Computer Science*, Volume 5 - 2023.
- Aurélien Breidenbach, Caroline Mahlow, and Andreas Schreiber. 2021. Implicit gender bias in computer science – a qualitative study.
- Brian McFee, Colin Raffel, Dawen Liang, Daniel P.W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and Music Signal Analysis in Python. In *Proceedings of the 14th Python in Science Conference*, pages 18–24.
- Thomas Brommage. 2015. Just kidding, folks!: An expressivist analysis of offensive humor. *Florida Philosophical Review*, 15(1).

BIBLIOGRAPHY

- Goenawan Brotosaputro, Agung Supriyadi, and Michael Jones. 2024. Ai-powered assistive technologies for improved accessibility. *International Transactions on Artificial Intelligence (ITALIC)*, 3(1):76–84.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.
- Peter Brugger. 2001. From haunted brain to haunted science: A cognitive neuroscience view of paranormal and pseudoscientific thought. In James Hournan and RenseEditors Lange, editors, *Hauntings and Poltergeists: Multidisciplinary Perspectives*, page 195–213. McFarland.
- Gregory A Bryant and Jean E Fox Tree. 2002. Recognizing verbal irony in spontaneous speech. *Metaphor and symbol*, 17(2):99–119.
- Cristian Buciluă, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model compression. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, page 535–541, New York, NY, USA. Association for Computing Machinery.
- Bastian Bunzeck, Daniel Duran, Leonie Schade, and Sina Zarriß. 2024. Graphemes vs. phonemes: battling it out in character-based language models. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 54–64, Miami, FL, USA. Association for Computational Linguistics.

BIBLIOGRAPHY

- C. Burgers. 2010. *Verbal irony: Use and effects in written discourse*. Ph.D. thesis, UB Nijmegen.
- Chris Callison-Burch. 2009. Fast, cheap, and creative: Evaluating translation quality using Amazon’s Mechanical Turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 286–295, Singapore. Association for Computational Linguistics.
- E. Camp. 2012. Sarcasm, pretense, and the semantics/pragmatics distinction. *Nous*, 46:587–634.
- Silvia Casola, Simona Frenda, Soda Maren Lo, Erhan Sezerer, Antonio Uva, Valerio Basile, Cristina Bosco, Alessandro Pedrani, Chiara Rubagotti, Viviana Patti, and Davide Bernardi. 2024. MultiPICO: Multilingual perspectivist irony corpus. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16008–16021, Bangkok, Thailand. Association for Computational Linguistics.
- Thiago Castro Ferreira, Chris van der Lee, Emiel van Miltenburg, and Emiel Krahmer. 2019. Neural data-to-text generation: A comparison between pipeline and end-to-end architectures. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 552–562, Hong Kong, China. Association for Computational Linguistics.
- Tuhin Chakrabarty, Debanjan Ghosh, Smaranda Muresan, and Nanyun Peng. 2020. R³: Reverse, retrieve, and rank for sarcasm generation with commonsense knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7976–7986, Online. Association for Computational Linguistics.

BIBLIOGRAPHY

- Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2024. Art or artifice? large language models and the false promise of creativity. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.
- Tuhin Chakrabarty, Arkadiy Saakyan, and Smaranda Muresan. 2021. Don't go far off: An empirical study on neural poetry translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7253–7265, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Jeanne Sternlicht Chall and Edgar Dale. 1995. *Readability revisited: The new Dale-Chall readability formula*. Brookline Books.
- Arjun Chandrasekaran, Devi Parikh, and Mohit Bansal. 2018. Punny captions: Witty wordplay in image descriptions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 770–775, New Orleans, Louisiana. Association for Computational Linguistics.
- Tyler A. Chang and Benjamin K. Bergen. 2024. Language model behavior: A comprehensive survey. *Computational Linguistics*, 50(1):293–350.
- Yongzhu Chang, Rongsheng Zhang, Lin Jiang, Qihang Chen, Le Zhang, and Jiashu Pu. 2023. Sudowoodo: A Chinese lyric imitation system with source lyrics. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 99–105.
- Canyu Chen and Kai Shu. 2024. Can LLM-generated misinformation be detected? In *The Twelfth International Conference on Learning Representations*.

BIBLIOGRAPHY

- Chen Chen, Xinlong Hao, Weiwen Liu, Xu Huang, Xingshan Zeng, Shuai Yu, Dexun Li, Yuefeng Huang, Xiangcheng Liu, Wang Xinzhi, and Wu Liu. 2025. ACEBench: A comprehensive evaluation of LLM tool usage. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 12970–12998, Suzhou, China. Association for Computational Linguistics.
- Hong Chen, Raphael Shu, Hiroya Takamura, and Hideki Nakayama. 2021. Graph-Plan: Story generation by planning with event graph. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 377–386.
- Wenhu Chen, Yu Su, Xifeng Yan, and William Yang Wang. 2020. KGPT: Knowledge-grounded pre-training for data-to-text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8635–8648, Online. Association for Computational Linguistics.
- Yang Chen, Chong Yang, Tu Hu, Xinhao Chen, Man Lan, Li Cai, Xinlin Zhuang, Xuan Lin, Xin Lu, and Aimin Zhou. 2024. Are U a joke master? pun generation via multi-stage curriculum learning towards a humor LLM. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 878–890, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.
- Yihao Chen and Alexander Lerch. 2020. Melody-Conditioned Lyrics Generation with SeqGANs . In *2020 IEEE International Symposium on Multimedia (ISM)*, pages 189–196, Los Alamitos, CA, USA. IEEE Computer Society.
- Shiyang Cheng, Pengcheng Zhu, Jueting Liu, and Zehua Wang. 2024. A survey of grapheme-to-phoneme conversion methods. *Applied Sciences*, 14(24):11790.
- Yew Ken Chia, Vernon Toh, Deepanway Ghosal, Lidong Bing, and Soujanya Poria. 2024. PuzzleVQA: Diagnosing multimodal reasoning challenges of language models with

BIBLIOGRAPHY

- abstract visual patterns. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16259–16273, Bangkok, Thailand. Association for Computational Linguistics.
- Cheng-Han Chiang and Hung-yi Lee. 2023. Are synonym substitution attacks really synonym substitution attacks? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1853–1878.
- Won Ik Cho, Seok Min Kim, and Nam Soo Kim. 2020. Towards an efficient code-mixed grapheme-to-phoneme conversion in an agglutinative language: A case study on to-Korean transliteration. In *Proceedings of the 4th Workshop on Computational Approaches to Code Switching*, pages 65–70, Marseille, France. European Language Resources Association.
- Yin-Ping Cho, Fu-Rong Yang, Yung-Chuan Chang, Ching-Ting Cheng, Xiao-Han Wang, and Yi-Wen Liu. 2021. A Survey on Recent Deep Learning-driven Singing Voice Synthesis Systems . In *2021 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, pages 319–323, Los Alamitos, CA, USA. IEEE Computer Society.
- Francois Chollet, Mike Knoop, Gregory Kamradt, Bryan Landers, and Henry Pinkard. 2025. Arc-agi-2: A new challenge for frontier ai reasoning systems.
- Noam Chomsky. 1957. *Syntactic structures*. Syntactic structures. Mouton, Oxford, England.
- Noam. Chomsky, Morris. Halle, and American Council of Learned Societies. 1991 - 1968. *The sound pattern of English*, 1st mit press pbk. ed. edition. ACLS Humanities E-Book. MIT Press, Cambridge, Mass.

BIBLIOGRAPHY

- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayanan Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.
- Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS’17, page 4302–4310, Red Hook, NY, USA. Curran Associates Inc.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tai, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav

BIBLIOGRAPHY

- Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2024a. Scaling instruction-finetuned language models. 25(1).
- Jiwan Chung, Seungwon Lim, Jaehyun Jeon, Seungbeen Lee, and Youngjae Yu. 2024b. Can visual language models resolve textual ambiguity with visual cues? let visual puns tell you! In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2452–2469, Miami, Florida, USA. Association for Computational Linguistics.
- Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that’s ‘human’ is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296, Online. Association for Computational Linguistics.
- G. Nick Clements and Rachid Ridouane. 2011. *Where Do Phonological Features Come From? : Cognitive, Physical and Developmental Bases of Distinctive Speech Categories.*, 1st ed. edition. Language Faculty and Beyond Series. John Benjamins Publishing Company, Amsterdam/Philadelphia.
- Jennifer Coates. 2016. *Women, men and language: A sociolinguistic account of gender differences in language.* Routledge.
- Julian Coda-Forno, Kristin Witte, Akshay K. Jagadish, Marcel Binz, Zeynep Akata, and Eric Schulz. 2023. Inducing anxiety in large language models increases exploration and bias. *ArXiv Pre-print*.
- Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

BIBLIOGRAPHY

- Common Crawl Foundation. 2025. Common crawl dataset. <https://commoncrawl.org/>.
- Susan Corbett. 2019. Computer game licences: The eula and its discontents. *Computer Law & Security Review*, 35(4):453–461.
- Carlos J. Costa, Manuela Aparicio, Sofia Aparicio, and Joao Tiago Aparicio. 2024. The democratization of artificial intelligence: Theoretical framework. *Applied Sciences*, 14(18).
- Council of Europe. 2001. Common european framework of reference for languages: Learning, teaching, assessment.
- Samuel Rhys Cox, Rune Møberg Jacobsen, and Niels van Berkel. 2025. The impact of a chatbot’s ephemerality-framing on self-disclosure perceptions. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces*, CUI ’25, New York, NY, USA. Association for Computing Machinery.
- David H. Cropley. 2025. “The Cat Sat on the ...?” Why Generative AI Has Limited Creativity. *Journal of Creative Behavior*, (4):1–9.
- Wenqian Cui, Dianzhi Yu, Xiaoqi Jiao, Ziqiao Meng, Guangyan Zhang, Qichao Wang, Steven Y. Guo, and Irwin King. 2025. Recent advances in speech language models: A survey. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13943–13970, Vienna, Austria. Association for Computational Linguistics.
- Christine Cuskley, Rebecca Woods, and Molly Flaherty. 2024. The limitations of large language models for understanding human language and cognition. *Open Mind*, 8:1058–1083.

BIBLIOGRAPHY

- Aleksandra Ćwiek, Susanne Fuchs, Christoph Draxler, Eva Liina Asu, Dan Dediu, Katri Hiovain, Shigeto Kawahara, Sofia Koutalidis, Manfred Krifka, Pärtel Lippus, Gary Lupyan, Grace E. Oh, Jing Paul, Caterina Petrone, Rachid Ridouane, Sabine Reiter, Nathalie Schümchen, Ádám Szalontai, Özlem Ünal-Logacev, Jochen Zeller, Marcus Perlman, and Bodo Winter. 2022. The bouba/kiki effect is robust across cultures and writing systems. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 377(1841). Publisher Copyright: © 2021 The Authors.
- Iria de Dios-Flores, Carmen Magariños, Adina Ioana Vladu, John E. Ortega, José Ramon Pichel, Marcos García, Pablo Gamallo, Elisa Fernández Rei, Alberto Bugarín-Diz, Manuel González González, Senén Barro, and Xosé Luis Regueira. 2022. The nós project: Opening routes for the Galician language in the field of language technologies. In *Proceedings of the Workshop Towards Digital Language Equality within the 13th Language Resources and Evaluation Conference*, pages 52–61, Marseille, France. European Language Resources Association.
- Paul V. De Lacy. 2007. *The Cambridge handbook of phonology*. Cambridge University Press, Cambridge.
- Ferdinand de Saussure. 1916. *Course in General Linguistics*. Columbia University Press.
- Tuba Degirmenci, Frank Mathmann, and Gary Mortimer. 2025. Made with love: Examining consumer engagement with handmade versus machine-made production cues. *Journal of Consumer Behaviour*, 24(2):982–1002.
- Mathieu Dehouck and Marine Delaborde. 2025. Rule-based approaches to the automatic generation of puns based on given names in French. In *Proceedings of the 1st*

BIBLIOGRAPHY

- Workshop on Computational Humor (CHum)*, pages 18–22, Online. Association for Computational Linguistics.
- Aliya Deri and Kevin Knight. 2016. Grapheme-to-phoneme models for (almost) any language. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 399–408, Berlin, Germany. Association for Computational Linguistics.
- Christophe Dessimoz and Paul D. Thomas. 2024. Ai and the democratization of knowledge. *Scientific Data*, 11(1):268.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can ai language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600.
- Wenjian Ding, Yao Zhang, Jun Wang, Adam Jatowt, and Zhenglu Yang. 2024. Can we learn question, answer, and distractors all from an image? a new task for multiple-choice visual question answering. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2852–2863, Torino, Italia. ELRA and ICCL.
- Doe v. Github, Inc. 2022. *Doe et al. v. Github, Inc. et al.*, no. 4:22-cv-06823-jst (n.d. cal. jan. 13, 2023). United States District Court for the Northern District of California.

BIBLIOGRAPHY

- Gayle T. Dow. 2017. *Defining creativity.*, Creativity and innovation: Theory, research, and practice., pages 5–21. Prufrock Press Inc., Waco, TX, US.
- Wei Duan, Yi Yu, Xulong Zhang, Suhua Tang, Wei Li, and Keizo Oyama. 2023. Melody generation from lyrics with local interpretability. *ACM Trans. Multimedia Comput. Commun. Appl.*, 19(3).
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, et al. 2024. The llama 3 herd of models.
- Marta Dynel. 2009. Beyond a joke: Types of conversational humour. *Language and Linguistics Compass*, 3(5):1284–1299.
- Julian Eisenschlos, Syrine Krichene, and Thomas Müller. 2020. Understanding tables with intermediate pre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 281–296, Online. Association for Computational Linguistics.
- Samer Elhajjar and Omar S. Itani. 2025. Ai washing: A conceptual exploration. *AMS Review*.
- Liana Ermakova, Ricardo Campos, Anne-Gwenn Bosser, and Tristan Miller. 2026. Overview of the clef 2025 joker lab: Humour in machine. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 315–337, Cham. Springer Nature Switzerland.
- Allyson Ettinger, David Heineman, Amanda Bertsch, Bailey Kuehl, Dirk Groeneveld, Hamish Ivison, Jacob Morrison, Matt Jordan, Mayee Chen, Pete Walsh, Pradeep Dasigi, Scott Geng, Shane Arora, Tyler Murray, Tyler Romero, Faeze Brahman, Jake Poznanski, David Graham, Finbarr Timbers, Kyle Lo, Luca Soldaini, Michael

BIBLIOGRAPHY

- Noukhovitch, Nathan Lambert, Robert Berry, Shashank Gupta, Victoria Graf, Saumya Malik, Saurabh Shah, Taira Anderson, Teng Xiao, Akari Asai, Akshita Bhagia, Alex Wettig, Alisa Liu, Chloe Anastasiades, Costa Huang, Dustin Schwenk, Aman Rangapur, Harsh Trivedi, Ian Magnusson, Jaron Lochner, Jiacheng Liu, Lj Miranda, Maarten Sap, Malia Morgan, Michael Schmitz, Michal Guerquin, Michael Wilson, Regan Huff, Ronan Le Bras, Rui Xin, Rulin Shao, Sam Skjonsberg, Shannon Zejiang, Tucker Shen, Valentina Pyatkin, Will Merrill, Yapei Chang, Yuling Gu, Ashish Sabharwal, Luke Zettlemoyer, Pang Wei Koh, Ali Farhadi, Noah A. Smith, Hannaneh Hajishirzi, Shuyue Stella Li, Zhiyuan Zeng, and the Olmo Team. 2025. Olmo 3. Technical report, Allen Institute for AI (AI2).
- European Commission and Directorate-General for Research and Innovation. 2019. She figures 2018. Technical report, European Commission.
- Xiaoming Fan, Anqi Lai, and Hean Tat Keh. 2024. Handmade vs. machine-made: the effects of handmade gifts on social relationships. *Marketing Letters*, 35(3):1–14.
- Dalya Faraj and Malak Abdullah. 2021. SarcasmDet at SemEval-2021 task 7: Detect humor and offensive based on demographic factors using RoBERTa pre-trained model. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 527–533, Online. Association for Computational Linguistics.
- Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press.
- Elena Filatova. 2012. Irony and sarcasm: Corpus generation and analysis using crowdsourcing. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*, pages 392–398, Istanbul, Turkey. European Language Resources Association (ELRA).

BIBLIOGRAPHY

- Ruth Filik, Alexandra Turcan, Christina Ralph-Nearman, and Alain Pitiot. 2019. What is the difference between irony and sarcasm? an fmri study. *Cortex*, 115:112–122.
- Eve Fleisig, Su Lin Blodgett, Dan Klein, and Zeerak Talat. 2024a. The perspectivist paradigm shift: Assumptions and challenges of capturing human labels. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2279–2292, Mexico City, Mexico. Association for Computational Linguistics.
- Eve Fleisig, Genevieve Smith, Madeline Bossi, Ishita Rustagi, Xavier Yin, and Dan Klein. 2024b. Linguistic bias in ChatGPT: Language models reinforce dialect discrimination. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13541–13564, Miami, Florida, USA. Association for Computational Linguistics.
- JL Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378—382.
- R.R Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32:2211–2223.
- Thomas E. Ford, Kyle Richardson, and Whitney E. Petit. 2015. Disparagement humor and prejudice: Contemporary theory and research. *HUMOR*, 28(2):171–186.
- Mary Ellen Foster and Michael White. 2007. Avoiding repetition in generated text. In *Proceedings of the Eleventh European Workshop on Natural Language Generation (ENLG 07)*, pages 33–40.
- Giorgio Franceschelli and Mirco Musolesi. 2024. On the creativity of large language models. *AI Soc.*, 40(5):3785–3795.

BIBLIOGRAPHY

- W Nelson Francis and Henry Kucera. 1979. The brown corpus. *Department of Linguistics, Brown University*.
- Simona Frenda, Alessandro Pedrani, Valerio Basile, Soda Marem Lo, Alessandra Teresa Cignarella, Raffaella Panizzon, Cristina Marco, Bianca Scarlini, Viviana Patti, Cristina Bosco, and Davide Bernardi. 2023. Epic: Multi-perspective annotation of a corpus of irony. In *ACL 2023*, pages 13844–13857, Toronto, Canada. Association for Computational Linguistics.
- Sigmund Freud. 1963. *Jokes and their relation to the unconscious*. Jokes and their relation to the unconscious. W. W. Norton, Oxford, England.
- Christoph Fuchs, Martin Schreier, and Stijn M.J. Van Osselaer. 2015. The handmade effect: What’s love got to do with it? *Journal of Marketing*, 79(2):98–110.
- Robert C. Gale, Alexandra C. Salem, Gerasimos Fergadiotis, and Steven Bedrick. 2023. Mixed orthographic/phonemic language modeling: Beyond orthographically restricted transformers (BORT). In *Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023)*, pages 212–225, Toronto, Canada. Association for Computational Linguistics.
- Vaishali Ganganwar, Manvinder, Mohit Singh, Priyank Patil, and Saurabh Joshi. 2024. Sarcasm and humor detection in code-mixed hindi data: A survey. In *Computing and Machine Learning*, pages 453–469, Singapore. Springer Nature Singapore.
- Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The pile: An 800gb dataset of diverse text for language modeling.
- Yingming Gao, Baorian Nuchged, Ya Li, and Linkai Peng. 2024. An investigation of applying large language models to spoken language learning. *Applied Sciences*, 14(1).

BIBLIOGRAPHY

- Judit García-González, Patricia Forcén, and Maria Jimenez-Sanchez. 2019. Men and women differ in their perception of gender bias in research institutions. *PloS one*, 14(12):e0225763–e0225763.
- Aparna Garimella, Carmen Banea, Nabil Hossain, and Rada Mihalcea. 2020. “judge me by my size (noun), do you?” YodaLib: A demographic-aware humor generation framework. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2814–2825, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Maryanne Garry, Way Ming Chan, Jeffrey Foster, and Linda A Henkel. 2024. Large language models (LLMs) and the institutionalization of misinformation. *Trends in Cognitive Science*, 28(12):1078–1088.
- Berys Nigel Gaut. 1998. Just joking: The ethics and aesthetics of humor. *Philosophy and Literature*, 22(1):51–68.
- Yan Ge, Victor Junqiu Wei, Yuanfeng Song, Jason Chen Zhang, and Raymond Chi-Wing Wong. 2024. Automatic data visualization generation from Chinese natural language questions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1889–1898, Torino, Italia. ELRA and ICCL.
- Jonas Geiping, Sean McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian R. Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. 2025. Scaling up Test-Time Compute with Latent Reasoning: A Recurrent Depth Approach.
- Theodore Seuss Geisel. 1965. *Fox in socks: Dr. Seuss’s book of tongue tanglers*. Random House.

BIBLIOGRAPHY

- Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context.
- Tatjana Georgievska-Jancheska. 2019. Lambdacism, rhotacism and sigmatism in preschool children: Frequency and distribution. *Open Access Macedonian Journal of Medical Science*, 7(3):336–340.
- Michael Gerlich. 2025. Ai tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies*, 15(1).
- Ricky Gervais. 2011. Ricky gervais: The difference between american and british humour.
- Marjan Ghazvininejad, Xing Shi, Yejin Choi, and Kevin Knight. 2016. Generating topical poetry. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1191, Austin, Texas. Association for Computational Linguistics.
- Akash Ghosh, Aparna Garimella, Pritika Ramu, Sambaran Bandyopadhyay, and Sriparna Saha. 2025. Infogen: Generating complex statistical infographics from documents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20552–20570, Vienna, Austria. Association for Computational Linguistics.
- Debanjan Ghosh, Avijit Vajpayee, and Smaranda Muresan. 2020. A report on the 2020 sarcasm detection shared task. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 1–11, Online. Association for Computational Linguistics.
- Raymond W Gibbs. 1986. On the psycholinguistics of sarcasm. *Journal of experimental psychology: general*, 115(1):3.

BIBLIOGRAPHY

- Bryan Gick, Ian Wilson, and Donald Derrick. 2013. *Articulatory phonetics*. John Wiley & Sons.
- Mayank Goel, Parameswari Krishnamurthy, and Radhika Mamidi. 2024. Automating humor: A novel approach to joke generation using template extraction and infilling. In *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*, pages 442–448, AU-KBC Research Centre, Chennai, India. NLP Association of India (NLP AI).
- Rashi Goel, Harsh Nandwani, Eshaan Shah, Ashalatha Nayak, and Archana Praveen Kumar. 2025. Enhancing visual question answering for multiple choice questions. *IEEE Access*.
- Seraphina Goldfarb-Tarrant, Tuhin Chakrabarty, Ralph Weischedel, and Nanyun Peng. 2020. Content planning for neural story generation with aristotelian rescoring. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4319–4338, Online. Association for Computational Linguistics.
- Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. 2022. Making science simple: Corpora for the lay summarisation of scientific literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10589–10604, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. 2023. Domain-driven and discourse-guided scientific summarisation. In *European Conference on Information Retrieval*, pages 361–376.
- Carlos Gómez-Rodríguez and Paul Williams. 2023. A confederacy of models: a comprehensive evaluation of LLMs on creative writing. In *Findings of the Association for*

BIBLIOGRAPHY

- Computational Linguistics: EMNLP 2023*, pages 14504–14528, Singapore. Association for Computational Linguistics.
- Hugo Gonalo Oliveira. 2017. A survey on intelligent poetry generation: Languages, features, techniques, reutilisation and evaluation. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 11–20, Santiago de Compostela, Spain. Association for Computational Linguistics.
- Hugo Gonalo Oliveira. 2015. Tra-la-lyrics 2.0: Automatic generation of song lyrics on a semantic domain. *Journal of Artificial General Intelligence*, 6(1):87–110.
- Hugo Gonalo Oliveira, F Cardoso, Amialcar Cardoso, and Francisco C. Pereira. 2007. Tra-la-lyrics: An approach to generate text based on rhythm. *Proceedings of the 4th International Joint Workshop on Computational Creativity*, pages 47–55.
- Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2, NIPS’14*, page 2672–2680, Cambridge, MA, USA. MIT Press.
- Z bulon Goriely, Richard Diehl Martinez, Andrew Caines, Paula Buttery, and Lisa Beinborn. 2024. From babble to words: Pre-training language models on continuous streams of phonemes. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 37–53, Miami, FL, USA. Association for Computational Linguistics.
- Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, Khaled Saab, Dan Popovici, Jacob Blum, Fan Zhang, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat, Pushmeet Kohli, Yossi Matias, Andrew Carroll, Kavita

BIBLIOGRAPHY

- Kulkarni, Nenad Tomasev, Yuan Guan, Vikram Dhillon, Eeshit Dhaval Vaishnav, Byron Lee, Tiago R D Costa, José R Penadés, Gary Peltz, Yunhan Xu, Annalisa Pawlosky, Alan Karthikesalingam, and Vivek Natarajan. 2025. Towards an ai co-scientist.
- P.H. Grice. 1978. *Further notes on logic and conversation*, volume 9, pages 113–127. P. Cole, Syntax and Semantics, New York: Academic Press.
- Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, William Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah Smith, and Hannaneh Hajishirzi. 2024. OLMo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15789–15809, Bangkok, Thailand. Association for Computational Linguistics.
- David Gros, Yu Li, and Zhou Yu. 2022. Robots-dont-cry: Understanding falsely anthropomorphic utterances in dialog systems. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3266–3284, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Jian Guan, Junfei Wu, Jia-Nan Li, Chuanqi Cheng, and Wei Wu. 2025. A survey on personalized Alignment—The missing piece for large language models in real-world

BIBLIOGRAPHY

- applications. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5313–5333, Vienna, Austria. Association for Computational Linguistics.
- Marco Guerini, Gözde Özbal, and Carlo Strapparava. 2015. Echoes of persuasion: The effect of euphony in persuasive communication. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1483–1493, Denver, Colorado. Association for Computational Linguistics.
- R. Gunning. 1971. *The Technique of Clear Writing*. McGraw-Hill.
- Manish Gupta and Puneet Agrawal. 2022. Compression of deep learning models for text: A survey. *ACM Trans. Knowl. Discov. Data*, 16(4).
- Prakhar Gupta, Shikib Mehri, Tiancheng Zhao, Amy Pavel, Maxine Eskenazi, and Jeffrey Bigham. 2019. Investigating evaluation of open-domain dialogue systems with human generated multiple references. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 379–391.
- Lorenz Gutscher and Michael Pucher. 2024. Exploring phonetic features in language embeddings for unseen language varieties of Austrian German. In *Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024)*, pages 317–325, Vienna, Austria. Association for Computational Linguistics.
- Michael Hammond. 2023. Low-resource grapheme-to-phoneme mapping with phonetically-conditioned transfer. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 245–248.
- Dongrui Han, Mingyu Cui, Jiawen Kang, Xixin Wu, Xunying Liu, and Helen Meng. 2024a. Improving grapheme-to-phoneme conversion through in-context knowledge

BIBLIOGRAPHY

- retrieval with large language models. In *2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 631–635. IEEE.
- Jieun Han, Haneul Yoo, Junho Myung, Minsun Kim, Hyunseung Lim, Yoonsu Kim, Tak Yeon Lee, Hwajung Hong, Juho Kim, So-Yeon Ahn, and Alice Oh. 2024b. LLM-as-a-tutor in EFL writing education: Focusing on evaluation of student-LLM interaction. In *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*, pages 284–293, Miami, Florida, USA. Association for Computational Linguistics.
- Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque. 2019. UR-FUNNY: A multimodal language dataset for understanding humor. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2046–2056, Hong Kong, China. Association for Computational Linguistics.
- He He, Nanyun Peng, and Percy Liang. 2019. Pun generation with surprise. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1734–1744, Minneapolis, Minnesota. Association for Computational Linguistics.
- Xinyi He, Mengyu Zhou, Mingjie Zhou, Jialiang Xu, Xiao Lv, Tianle Li, Yijia Shao, Shi Han, Zejian Yuan, and Dongmei Zhang. 2023. AnaMeta: A table understanding dataset of field metadata knowledge shared by multi-dimensional data analysis tasks. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9471–9492, Toronto, Canada. Association for Computational Linguistics.

BIBLIOGRAPHY

- Shaif Hemraj. 2025. Ai and the future of creative development: redefining digital media production. *AI and Ethics*, 5(5):5105–5119.
- Joseph Henrich, Steven J. Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2-3):61–83.
- Isabella Hermann. 2023. Artificial intelligence in fiction: between narratives and metaphors. *AI & SOCIETY*, 38(1):319–329.
- Jack Hessel, Ana Marasovic, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2023. Do androids laugh at electric sheep? humor “understanding” benchmarks from the new yorker caption contest. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 688–714, Toronto, Canada. Association for Computational Linguistics.
- Tim Highfield. 2015. Tweeted joke lifespans and appropriated punch lines: Practices around topical humor on social media. *International Journal of Communication*, 9(0).
- Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.
- Zoë Hitzig, Aaron Chatterji, Tom Cunningham, David Deming, Christopher Ong, Carl Shan, and Kevin Wadman. 2025. How people use chatgpt. *OpenAI (affiliated with Duke University and Harvard University)*. Dated September 15, 2025.
- Thomas Hobbes. 1660. *The Leviathan*.

BIBLIOGRAPHY

- Gordon Hodson, Jonathan Rush, and Cara C. Macinnis. 2010. A joke is just a joke (except when it isn't): cavalier humor beliefs facilitate the expression of group dominance motives. *Journal of Personality and Social Psychology*, 99(4):660–682.
- Thomas R. Hofmann. 1993. *Realms of meaning : an introduction to semantics*. Learning about language. Longman, London.
- Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1546.
- Anita Holdcroft. 2007. Gender bias in research: how does it affect evidence based medicine? *Journal of the Royal Society of Medicine*, 100(1):2–3. PMID: 17197669.
- Bryan Anthony Hong and Ethel Ong. 2009. Automatically extracting word relationships as templates for pun generation. In *Proceedings of the Workshop on Computational Approaches to Linguistic Creativity*, pages 24–31, Boulder, Colorado. Association for Computational Linguistics.
- Xudong Hong, Asad Sayeed, Khushboo Mehra, Vera Demberg, and Bernt Schiele. 2023. Visual writing prompts: Character-grounded story generation with curated image sequences. *Transactions of the Association for Computational Linguistics*, 11:565–581.
- Jack Hopkins and Douwe Kiela. 2017. Automatically generating rhythmic verse with neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 168–178, Vancouver, Canada. Association for Computational Linguistics.
- Zachary Horvitz, Nam Do, and Michael L. Littman. 2020. Context-driven satirical news generation. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 40–50, Online. Association for Computational Linguistics.

BIBLIOGRAPHY

- Jeff Horwitz. 2025. Meta’s AI rules have let bots hold ’sensual’ chats with kids, offer false medical info. *Reuters Special Report*.
- Nabil Hossain, John Krumm, Michael Gamon, and Henry Kautz. 2020. SemEval-2020 task 7: Assessing humor in edited news headlines. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 746–758, Barcelona (online). International Committee for Computational Linguistics.
- Yifan Hou, Buse Giledereli, Yilei Tu, and Mrinmaya Sachan. 2025. Do vision-language models really understand visual language?
- David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuezhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.
- Henglin Huang, Chen Tang, Tyler Loakman, Frank Guerin, and Chenghua Lin. 2022. Improving Chinese story generation via awareness of syntactic dependencies and semantics. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*.
- Andrew Dana Hudson, Ed Finn, and Ruth Wylie. 2023. What can science fiction tell us about the future of artificial intelligence policy? *AI & SOCIETY*, 38(1):197–211.

BIBLIOGRAPHY

- Lluís-F. Hurtado, Encarna Segarra, Ferran Pla, Pascual Carrasco, and José-Ángel González. 2017. ELiRF-UPV at SemEval-2017 task 7: Pun detection and interpretation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 440–443, Vancouver, Canada. Association for Computational Linguistics.
- Luke Hutton and Tristan Henderson. 2017. *Beyond the EULA: Improving Consent for Data Mining*, pages 147–167. Springer International Publishing, Cham.
- Jessica Huynh, Jeffrey Bigham, and Maxine Eskenazi. 2021. A survey of nlp-related crowdsourcing hits: what works and what does not.
- EunJeong Hwang and Vered Shwartz. 2023. MemeCap: A dataset for captioning and interpreting memes. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1433–1445, Singapore. Association for Computational Linguistics.
- Micol Iannuzzi, Glynis Bogaard, and Jennifer M. Schell-Leugers. 2025. Ai-generated media and the erosion of trust in legal settings: a review on impact and distortion of belief. *AI & SOCIETY*.
- Oana Ignat, Zhijing Jin, Artem Abzaliev, Laura Biester, Santiago Castro, Naihao Deng, Xinyi Gao, Aylin Ece Gunal, Jacky He, Ashkan Kazemi, Muhammad Khalifa, Namho Koh, Andrew Lee, Siyang Liu, Do June Min, Shinka Mori, Joan C. Nwatu, Veronica Perez-Rosas, Siqi Shen, Zekun Wang, Winston Wu, and Rada Mihalcea. 2024. Has it all been solved? open NLP research questions not solved by large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8050–8094, Torino, Italia. ELRA and ICCL.

BIBLIOGRAPHY

- Vijayasaradhi Indurthi and Subba Reddy Oota. 2017. Fermi at SemEval-2017 task 7: Detection and interpretation of homographic puns in English language. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 457–460, Vancouver, Canada. Association for Computational Linguistics.
- Neslihan Iskender, Tim Polzehl, and Sebastian Möller. 2021. Reliability of human evaluation for text summarization: Lessons learned and challenges ahead. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 86–96.
- Mohammed Saidul Islam, Raian Rahman, Ahmed Masry, Md Tahmid Rahman Laskar, Mir Tafseer Nayeem, and Enamul Hoque. 2024. Are large vision language models up to the challenge of chart comprehension and reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3334–3368, Miami, Florida, USA. Association for Computational Linguistics.
- Hayate Iso. 2022. Autotemplate: A simple recipe for lexically constrained text generation. In *Proceedings of the 17th International Natural Language Generation Conference*.
- Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2024. Livecodebench: Holistic and contamination free evaluation of large language models for code.
- Shailee Jain, Vy A. Vo, Leila Wehbe, and Alexander G. Huth. 2023. Computational Language Modeling and the Promise of in Silico Experimentation. *Neurobiology of Language*, pages 1–27.
- Sophie Jentzsch and Kristian Kersting. 2023. ChatGPT is fun, but it is not funny! humor is still challenging large language models. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 325–340, Toronto, Canada. Association for Computational Linguistics.

BIBLIOGRAPHY

- Michael Jessen. 2008. Forensic phonetics. *Language and linguistics compass*, 2(4):671–711.
- Prince Jha, Krishanu Maity, Raghav Jain, Apoorv Verma, Sriparna Saha, and Pushpak Bhattacharyya. 2024. Meme-ingful analysis: Enhanced understanding of cyberbullying in memes through multimodal explanations. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 930–943, St. Julian’s, Malta. Association for Computational Linguistics.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, et al. 2023. Mistral 7b. *ArXiv Pre-Print*.
- Ruili Jiang, Kehai Chen, Xuefeng Bai, Zhixuan He, Juntao Li, Muyun Yang, Tiejun Zhao, Liqiang Nie, and Min Zhang. 2025. A survey on human preference learning for aligning large language models. *ACM Comput. Surv.*, 58(6).
- Tonglin Jiang, Hao Li, and Yubo Hou. 2019. Cultural differences in humor perception, usage, and implications. *Frontiers in Psychology*, 10.
- Valentin Jijkoun and Katja Hofmann. 2009. Generating a non-english subjectivity lexicon: Relations that matter. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 398–405.
- Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan. 2024. SWE-bench: Can language models resolve real-world github issues? In *The Twelfth International Conference on Learning Representations*.
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world.

BIBLIOGRAPHY

- In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.
- Nick Judd, Alexandre Vaz, Kevin Paeth, Layla Inés Davis, Milena Esherick, Jason Brand, Inês Amaro, and Tony Rousmaniere. 2025. Independent clinical evaluation of general-purpose llm responses to signals of suicide risk.
- Faria Binte Kader, Nafisa Hossain Nujat, Tasmia Binte Sogir, Mohsinul Kabir, Hasan Mahmud, and Md Kamrul Hasan. 2023. “when words fail, emojis prevail”: A novel architecture for generating sarcastic sentences with emoji using valence reversal and semantic incongruity. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 334–351, Toronto, Canada. Association for Computational Linguistics.
- Mihir Kale and Abhinav Rastogi. 2020. Text-to-text pre-training for data-to-text tasks. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 97–102, Dublin, Ireland. Association for Computational Linguistics.
- Rogers Kaliisa, Ryan Shaun Baker, Barbara Wasson, and Paul Prinsloo. 2025. The coming but uneven storm: How ai regulation will impact ai and learning analytics research in different countries. *Journal of Learning Analytics*, 12(2):140–157.
- Antonios Kalloniatis and Panagiotis Adamidis. 2024. Computational humor recognition: a systematic literature review. *Artificial Intelligence Review*, 58(2):43.
- Masahiro Kaneko, Danushka Bollegala, and Naoaki Okazaki. 2023. The impact of debiasing on the performance of language models in downstream tasks is underestimated. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for*

BIBLIOGRAPHY

- Computational Linguistics (Volume 2: Short Papers)*, pages 29–36, Nusa Dua, Bali. Association for Computational Linguistics.
- Immanuel Kant. 1790. *Critique of Judgment*. Barnes & Noble.
- Justine T. Kao, Roger Levy, and Noah D. Goodman. 2016. A computational model of linguistic humor in puns. *Cognitive Science*, 40(5):1270–1285.
- Ayush Kaushal and Kyle Mahowald. 2022. What do tokens know about their characters and how do they know it? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2487–2507, Seattle, United States. Association for Computational Linguistics.
- Sedrick Scott Keh, Steven Y. Feng, Varun Gangal, Malihe Alikhani, and Eduard Hovy. 2023. PANCETTA: Phoneme aware neural completion to elicit tongue twisters automatically. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 491–504, Dubrovnik, Croatia. Association for Computational Linguistics.
- Heather Kember, Kathryn Connaghan, and Rupal Patel. 2017. Inducing speech errors in dysarthria using tongue twisters. *International journal of language & communication disorders*, 52(4):469–478.
- Aimee Kendall Roundtree. 2024. Public perception of ai: A review. In *HCI International 2024 – Late Breaking Papers*, pages 72–87, Cham. Springer Nature Switzerland.
- Mary Ogbuka Kenneth, Foaad Khosmood, and Abbas Edalat. 2024. Systematic literature review: Computational approaches for humour style classification.

BIBLIOGRAPHY

- Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation.
- Seth Ketron and Nancy Spears. 2021. Sound-symbolic signaling of online retailer sizes: The moderating effect of shopping goals. *Journal of Retailing and Consumer Services*, 58:102245.
- Anas Anwarul Haq Khan, Tanik Saikh, Arpan Phukan, and Asif Ekbal. 2024. Hope ‘the paragraph guy’ explains the rest : Introducing MeSum, the meme summarizer. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6654–6668, Miami, Florida, USA. Association for Computational Linguistics.
- Reza Khanmohammadi, Mitra Sadat Mirshafiee, Yazdan Rezaee Jouryabi, and Seyed Abolghasem Mirroshandel. 2023. Prose2poem: the blessing of transformers in translating prose to persian poetry. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(6):1–18.
- Hwa-Yeon Kim, Jong-Hwan Kim, and Jae-Min Kim. 2022. Fast bilingual grapheme-to-phoneme conversion. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, pages 289–296.
- Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An Open Source Language Model Specialized in Evaluating Other Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4334–4353, Miami, Florida, USA. Association for Computational Linguistics.

BIBLIOGRAPHY

- Jürgen Klausenburger. 1970. *French Prosodics and Phonotactics : An Historical Typology.*, 1st ed. edition. Beihefte Zur Zeitschrift Für Romanische Philologie Series. Walter de Gruyter GmbH, Tübingen.
- W. Köhler. 1929. *Gestalt psychology.* Gestalt psychology. Liveright, Oxford, England.
- Klaus Krippendorff. 2019. Content analysis: An introduction to its methodology.
- Paul Kroeger. 2018. *Analyzing meaning: An introduction to semantics and pragmatics.* Textbooks in Language Sciences. Language Science Press.
- Justin Kruger, Derrick Wirtz, Leaf Van Boven, and Todd Altermatt. 2004. The effort heuristic. *Journal of Experimental Social Psychology*, 40(1):91–98.
- Vikas Kumar and Manju Lata. 2025. *Chatbots and Beyond: Exploring the Future of Conversational Technology in Education.* John Wiley & Sons, Ltd.
- Peter Ladefoged. 1996. *Elements of acoustic phonetics.* A Phoenix book, P433. University of Chicago Press, Chicago.
- Liisi Laineste. 2002. Take it with a grain of salt: The kernel of truth in topical jokes. *Folklore: Electronic Journal of Folklore*, 21:7–25.
- Pierre-Carl Langlais, Carlos Rosas Hinostroza, Mattia Nee, Catherine Arnett, Pavel Chizhov, Eliot Krzystof Jones, Irène Girard, David Mach, Anastasia Stasenko, and Ivan P. Yamshchikov. 2025. Common corpus: The largest collection of ethical data for llm pre-training.
- Migle Laukyte. 2019. Ai as a legal person. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law, ICAIL '19*, page 209–213, New York, NY, USA. Association for Computing Machinery.

BIBLIOGRAPHY

- Tiffany J. Lawless, Conor J. O’Dea, Stuart S. Miller, and Donald A. Saucier. 2020. Is it really just a joke? gender differences in perceptions of sexist humor. *Humor (Berlin, Germany)*, 33(2):291–315.
- Brigitte Le Roux, Henry Rouanet, Mike Savage, and Alan Warde. 2008. Class and cultural division in the uk. *Sociology*, 42(6):1049–1071.
- Hao-Ping (Hank) Lee, Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks, and Nicholas Wilson. 2025. The impact of generative ai on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI ’25, New York, NY, USA. Association for Computing Machinery.
- Hsin-Pei Lee, Jhih-Sheng Fang, and Wei-Yun Ma. 2019. iComposer: An automatic songwriting system for Chinese popular music. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 84–88, Minneapolis, Minnesota. Association for Computational Linguistics.
- Yi-Chieh Lee, Naomi Yamashita, Yun Huang, and Wai Fu. 2020. "i hear you, i feel you": Encouraging deep self-disclosure through a chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI ’20, page 1–12, New York, NY, USA. Association for Computing Machinery.
- Zhihong Lei, Xingyu Na, Mingbin Xu, Ernest Pusateri, Christophe Van Gysel, Yuanyuan Zhang, Shiyi Han, and Zhen Huang. 2025. Contextualization of asr with llm using phonetic retrieval-based augmentation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

BIBLIOGRAPHY

- Wai Yie Leong. 2025. Ethical and philosophical perspectives on artificial intelligence-generated art. *International Journal of Social Sciences and Artistic Innovations*, 5.
- G. Lessard. 1992. Computational modelling of linguistic humour : Tom swifties. *Selected Papers from the 1992 Association for Literary and Linguistic Computing (ALLC) and the Association for Computers and the Humanities (ACH) Joint Annual Conference*, pages 175–178.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.
- Xinjian Li, Florian Metze, David R Mortensen, Shinji Watanabe, and Alan W Black. 2022a. Zero-shot learning for grapheme to phoneme conversion with language ensemble. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2106–2115.
- Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Xingran Chen, Hanzhi Yin, Chenghao Xiao, Chenghua Lin, Anton Ragni, Emmanouil Benetos, Norbert Gyenge, Roger Dannenberg, Ruibo Liu, Wenhua Chen, Gus Xia, Yemin Shi, Wenhao Huang, Zili Wang, Yike Guo, and Jie Fu. 2024. MERT: Acoustic music understanding model with large-scale self-supervised training. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*.

BIBLIOGRAPHY

- Yucheng Li, Frank Guerin, and Chenghua Lin. 2022b. The secret of metaphor on expressing stronger emotion. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 39–43.
- Yucheng Li, Shun Wang, Chenghua Lin, Frank Guerin, and Loic Barrault. 2023a. FrameBERT: Conceptual metaphor detection with frame embedding learning. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1558–1563.
- Zhi Li and Yin Zhang. 2023. Cultural concept adaptation on multimodal reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 262–276, Singapore. Association for Computational Linguistics.
- Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023b. Synthetic data generation with large language models for text classification: Potential and limitations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10443–10461.
- Marcio Lima Inácio and Hugo Gonçalo Oliveira. 2023. Towards generation and recognition of humorous texts in Portuguese. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 26–36, Dubrovnik, Croatia. Association for Computational Linguistics.
- Marcio Lima Inácio, Gabriela Wick-pedro, and Hugo Goncalo Oliveira. 2023. What do humor classifiers learn? an attempt to explain humor recognition models. In *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 88–98, Dubrovnik, Croatia. Association for Computational Linguistics.

BIBLIOGRAPHY

- Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Hongzhan Lin, Ziyang Luo, Jing Ma, and Long Chen. 2023. Beneath the surface: Unveiling harmful memes with multimodal reasoning distilled from large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9114–9128, Singapore. Association for Computational Linguistics.
- Hairong Liu, Mingbo Ma, Liang Huang, Hao Xiong, and Zhongjun He. 2019. Robust neural machine translation with joint textual and phonetic embedding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3044–3049, Florence, Italy. Association for Computational Linguistics.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26286–26296.
- Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. Llava-next: Improved reasoning, ocr, and world knowledge.
- Nayu Liu, Wenjing Han, Guangcan Liu, Da Peng, Ran Zhang, Xiaorui Wang, and Huabin Ruan. 2022a. ChipSong: A controllable lyric generation system for Chinese popular song. In *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022)*, pages 85–95, Dublin, Ireland. Association for Computational Linguistics.
- Nelson F. Liu, Gina-Anne Levow, and Noah A. Smith. 2018. Discovering phonemes with sparse regularization. In *Proceedings of the Second Workshop on Sub-*

BIBLIOGRAPHY

- word/Character LLevel Models*, pages 49–54, New Orleans. Association for Computational Linguistics.
- Ruixue Liu, Shaozu Yuan, Aijun Dai, Lei Shen, Tiangang Zhu, Meng Chen, and Xiaodong He. 2022b. Few-shot table understanding: A benchmark dataset and pre-training baseline. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3741–3752, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Elena Lloret, Laura Plaza, and Ahmet Aker. 2018. The challenging task of summary evaluation: an overview. *Language Resources and Evaluation*, 52:101–148.
- Tyler Loakman, Joseph James, and Chenghua Lin. 2025a. Seeing isn’t hearing: Benchmarking vision language models at interpreting spectrograms. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 79–86, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.
- Tyler Loakman, Yucheng Li, and Chenghua Lin. 2024. With ears to see and eyes to hear: Sound symbolism experiments with multimodal large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2849–2867, Miami, Florida, USA. Association for Computational Linguistics.
- Tyler Loakman and Chenghua Lin. 2024. ReproHum #0087-01: Human evaluation reproduction report for generating fact checking explanations. In *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval) @ LREC-COLING 2024*, pages 255–260, Torino, Italia. ELRA and ICCL.

BIBLIOGRAPHY

- Tyler Loakman, Aaron Maladry, and Chenghua Lin. 2023a. The iron(ic) melting pot: Reviewing human evaluation in humour, irony and sarcasm generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6676–6689, Singapore. Association for Computational Linguistics.
- Tyler Loakman, Chen Tang, and Chenghua Lin. 2023b. TwistList: Resources and baselines for tongue twister generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 579–589, Toronto, Canada. Association for Computational Linguistics.
- Tyler Loakman, Chen Tang, and Chenghua Lin. 2025b. Train and constrain: Phonologically informed tongue twister generation from topics and paraphrases. *Computational Linguistics*, 51(2):415–466.
- Tyler Loakman, William Thorne, and Chenghua Lin. 2025c. Comparing apples to oranges: A dataset & analysis of LLM humour understanding from traditional puns to topical jokes. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 9502–9518, Suzhou, China. Association for Computational Linguistics.
- Tyler Loakman, William Thorne, and Chenghua Lin. 2025d. Who’s laughing now? an overview of computational humour generation and explanation. In *Proceedings of the 18th International Natural Language Generation Conference*, pages 780–794, Hanoi, Vietnam. Association for Computational Linguistics.
- Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The ai scientist: Towards fully automated open-ended scientific discovery.
- Ximing Lu, Sean Welleck, Peter West, Liwei Jiang, Jungo Kasai, Daniel Khashabi, Ronan Le Bras, Lianhui Qin, Youngjae Yu, Rowan Zellers, Noah A. Smith, and Yejin Choi. 2022. NeuroLogic a*esque decoding: Constrained text generation with

BIBLIOGRAPHY

- lookahead heuristics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 780–799.
- Fuli Luo, Shun Yao Li, Pengcheng Yang, Lei Li, Baobao Chang, Zhifang Sui, and Xu Sun. 2019. Pun-GAN: Generative adversarial network for pun generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3388–3393, Hong Kong, China. Association for Computational Linguistics.
- Zheheng Luo, Chenhan Yuan, Qianqian Xie, and Sophia Ananiadou. 2025. EMPEC: A comprehensive benchmark for evaluating large language models across diverse health-care professions. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9945–9958, Vienna, Austria. Association for Computational Linguistics.
- Christopher W. Lynn and Danielle S. Bassett. 2020. How humans learn and represent networks. *Proceedings of the National Academy of Sciences*, 117(47):29407–29415.
- Jiayuan Ma, Hongbin Na, Zimu Wang, Yining Hua, Yue Liu, Wei Wang, and Ling Chen. 2025. Detecting conversational mental manipulation with intent-aware prompting. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9176–9183, Abu Dhabi, UAE. Association for Computational Linguistics.
- Brooke N. Macnamara, Ibrahim Berber, M. Cenk Çavuşoğlu, Elizabeth A. Krupinski, Naren Nallapareddy, Noelle E. Nelson, Philip J. Smith, Amy L. Wilson-Delfosse, and Soumya Ray. 2024. Does using artificial intelligence assistance accelerate skill decay and hinder skill development without performers’ awareness? *Cognitive Research: Principles and Implications*, 9(1):46.

BIBLIOGRAPHY

- Krishanu Maity, Prince Jha, Sriparna Saha, and Pushpak Bhattacharyya. 2022. A multitask framework for sentiment, emotion and sarcasm aware cyberbullying detection from multi-modal code-mixed memes. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 1739–1749, New York, NY, USA. Association for Computing Machinery.
- Wingyun Mak and Brian D. Carpenter. 2007. Humor comprehension in older adults. *Journal of the International Neuropsychological Society*, 13(4):606–614.
- Aaron Maladry, Els Lefever, Cynthia Van Hee, and Véronique Hoste. 2024. The limitations of irony detection in dutch social media. *Language Resources and Evaluation*, 58(4):1355–1386.
- Zofia Malisz, Jan Foremski, and Małgorzata Kul. 2024. PRODIS - a speech database and a phoneme-based language model for the study of predictability effects in Polish. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13068–13073, Torino, Italia. ELRA and ICCL.
- Shrikant Malviya, Rohit Mishra, and Uma Shanker Tiwary. 2016. Structural analysis of hindi phonetics and a method for extraction of phonetically rich sentences from a very large hindi text corpus. In *2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 188–193.
- Enrique Manjavacas, Mike Kestemont, and Folgert Karsdorp. 2019. Generation of hip-hop lyrics with hierarchical modeling and conditional templates. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 301–310.

BIBLIOGRAPHY

- Muhammad Arslan Manzoor, Sarah Albarri, Ziting Xian, Zaiqiao Meng, Preslav Nakov, and Shangsong Liang. 2023. Multimodality representation learning: A survey on evolution, pretraining and its applications. *ACM Trans. Multimedia Comput. Commun. Appl.*, 20(3).
- Marie-Helen Maras and Alex Alexandrou. 2019. Determining authenticity of video evidence in the age of artificial intelligence and in the wake of deepfake videos. *The International Journal of Evidence & Proof*, 23(3):255–262.
- Xavier Marjou. 2021. OTEANN: Estimating the transparency of orthographies with an artificial neural network. In *Proceedings of the Third Workshop on Computational Typology and Multilingual NLP*, pages 1–9, Online. Association for Computational Linguistics.
- Nina Markl and Stephen Joseph McNulty. 2022. Language technology practitioners as language managers: arbitrating data bias and predictive bias in ASR. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6328–6339, Marseille, France. European Language Resources Association.
- Nestor Maslej, Loredana Fattorini, Raymond Perrault, Yolanda Gil, Vanessa Parli, Njenga Kariuki, Emily Capstick, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, Toby Walsh, Armin Hamrah, Lapo Santarlaschi, Julia Betts Lotufo, Alexandra Rome, Andrew Shi, and Sukrut Oak. 2025. Artificial intelligence index report 2025.
- Ryo Masumura, Naoki Makishima, Mana Ihori, Akihiko Takashima, Tomohiro Tanaka, and Shota Orihashi. 2020. Phoneme-to-grapheme conversion based large-scale pre-

BIBLIOGRAPHY

- training for end-to-end automatic speech recognition. In *Interspeech 2020*, pages 2822–2826.
- Yuhi Matogawa, Yusuke Sakai, Taro Watanabe, and Chihiro Taguchi. 2024. Japanese rule-based grapheme-to-phoneme conversion system and multilingual named entity dataset with international phonetic alphabet. In *Proceedings of the 21st SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 77–86.
- Richard T. McCoy and Robert Frank. 2018. Phonologically informed edit distance algorithms for word alignment with low-resource languages. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2018*, pages 102–112.
- Deborah McCutchen and Charles A. Perfetti. 1982. The visual tongue-twister effect: Phonological activation in silent reading. *Journal of Verbal Learning and Verbal Behavior*, 21(6):672–687.
- A. Peter McGraw and Caleb Warren. 2010. Benign violations: Making immoral behavior funny. *Psychological Science*, 21(8):1141–1149.
- A. Peter McGraw, Caleb Warren, Lawrence E. Williams, and Bridget Leonard. 2012. Too close for comfort, or too far to care? finding humor in distant tragedies and close mishaps. *Psychological Science*, 23(10):1215–1223. PMID: 22941877.
- Stephen McGregor, Matthew Purver, and Geraint Wiggins. 2016. Process based evaluation of computer generated poetry. In *Proceedings of the INLG 2016 Workshop on Computational Creativity in Natural Language Generation*, pages 51–60, Edinburgh, UK. Association for Computational Linguistics.

BIBLIOGRAPHY

- Oliver McLaughlin, Jack Merullo, and Arjun Khurana. 2025. I have no mouth, and i must rhyme: Uncovering internal phonetic representations in llama 3.2. In *ICML 2025 Workshop on Assessing World Models*.
- Nicholas Meade, Elinor Poole-Dayana, and Siva Reddy. 2022. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1878–1898, Dublin, Ireland. Association for Computational Linguistics.
- J. A. Meaney. 2020. Crossing the line: Where do demographic variables fit into humor detection? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 176–181, Online. Association for Computational Linguistics.
- J. A. Meaney, Steven Wilson, Luis Chiruzzo, Adam Lopez, and Walid Magdy. 2021. SemEval 2021 task 7: HaHackathon, detecting and rating humor and offense. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 105–119, Online. Association for Computational Linguistics.
- Thomas Metzinger. 2021. Artificial suffering: An argument for a global moratorium on synthetic phenomenology. *Journal of Artificial Intelligence and Consciousness*, 08(01):43–66.
- Shivam Mhaskar, Nirmesh Shah, Mohammadi Zaki, Ashishkumar Gudmalwar, Pankaj Wasnik, and Rajiv Shah. 2024. Isometric neural machine translation using phoneme count ratio reward-based reinforcement learning. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3966–3976, Mexico City, Mexico. Association for Computational Linguistics.

BIBLIOGRAPHY

- George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.
- Tristan Miller, Christian Hempelmann, and Iryna Gurevych. 2017. SemEval-2017 task 7: Detection and interpretation of English puns. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 58–68, Vancouver, Canada. Association for Computational Linguistics.
- Abhijit Mishra, Tarun Tater, and Karthik Sankaranarayanan. 2019. A modular architecture for unsupervised sarcasm generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6144–6154, Hong Kong, China. Association for Computational Linguistics.
- Anirudh Mittal, Yufei Tian, and Nanyun Peng. 2022. AmbiPun: Generating humorous puns with ambiguous context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1053–1062, Seattle, United States. Association for Computational Linguistics.
- Brent Mittelstadt, Sandra Wachter, and Chris Russell. 2023. To protect science, we must use llms as zero-shot translators. *Nature Human Behaviour*, 7(11):1830–1832.
- Ishani Mondal, Kabir Ahuja, Mohit Jain, Jacki O’Neill, Kalika Bali, and Monojit Choudhury. 2022. Global readiness of language technology for healthcare: What would it take to combat the next pandemic? In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4320–4335, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

BIBLIOGRAPHY

- E.A. Moravcsik. 2006. *An Introduction to Syntax: Fundamentals of Syntactic Analysis*. Bloomsbury Academic.
- John Morreall. 2024. Philosophy of Humor. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Fall 2024 edition. Metaphysics Research Lab, Stanford University.
- David R. Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori Levin. 2016. PanPhon: A resource for mapping IPA segments to articulatory feature vectors. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3475–3484, Osaka, Japan. The COLING 2016 Organizing Committee.
- Kosuke Motoki, Toshiki Saito, Jaewoo Park, Carlos Velasco, Charles Spence, and Motoaki Sugiura. 2020. Tasting names: Systematic investigations of taste-speech sounds associations. *Food Quality and Preference*, 80:103801.
- Srija Mukhopadhyay, Adnan Qidwai, Aparna Garimella, Pritika Ramu, Vivek Gupta, and Dan Roth. 2024. Unraveling the truth: Do VLMs really understand charts? a deep dive into consistency and robustness. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16696–16717, Miami, Florida, USA. Association for Computational Linguistics.
- Abhishek Muzumdar, Gaurav Kurian, Arjun Rawat, and Priyesh Bhardwaj. 2025. The dead internet theory: A survey on artificial interactions and the future of social media. *Asian Journal of Research in Computer Science*, 18(1):67–73.
- Karim Nader, Paul Toprac, Suzanne Scott, and Samuel Baker. 2024. Public understanding of artificial intelligence through entertainment media. *AI & SOCIETY*, 39(2):713–726.

BIBLIOGRAPHY

- Toshiaki Nakazawa, Hideki Nakayama, Chenchen Ding, Raj Dabre, Shohei Higashiyama, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, Chenhui Chu, Akiko Eriguchi, Kaori Abe, Yusuke Oda, and Sadao Kurohashi. 2021. Overview of the 8th workshop on Asian translation. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 1–45, Online. Association for Computational Linguistics.
- Sarvesh Narayanan, Jayasimman J, and Shiva Ganesh V. 2024. Clef 2024 joker task 2: Using roberta and bert-uncased for humour classification according to genre and technique. In *CEUR Workshop Proceedings*, volume 3740 of *GEUR Workshop Proceedings*. Notebook for the Joker Lab at CLEF 2024.
- Boris New, Christophe Pallier, Marc Brysbaert, and Ludovic Ferrand. 2004. Lexique 2 : A new french lexical database. *Behavior Research Methods, Instruments, & Computers*, 36(3):516–524.
- Hoang H Nguyen, Khyati Mahajan, Vikas Yadav, Julian Salazar, Philip S. Yu, Masoud Hashemi, and Rishabh Maheshwary. 2025. Prompting with phonemes: Enhancing LLMs’ multilinguality for non-Latin script languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11975–11994, Albuquerque, New Mexico. Association for Computational Linguistics.
- Khoi P. N. Nguyen and Vincent Ng. 2024. Computational meme understanding: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21251–21267, Miami, Florida, USA. Association for Computational Linguistics.

BIBLIOGRAPHY

- Linh The Nguyen, Thinh Pham, and Dat Quoc Nguyen. 2023. XPhoneBERT: A Pre-trained Multilingual Model for Phoneme Representations for Text-to-Speech. In *Proceedings of the 24th Annual Conference of the International Speech Communication Association (INTERSPEECH)*.
- Alan K. S. Nielsen and Drew Rendall. 2013. Parsing the role of consonants versus vowels in the classic takete-maluma phenomenon. *Canadian Journal of Experimental Psychology / Revue canadienne de psychologie expérimentale*, 67(2):153–163.
- Anton Nijholt, Andreea Niculescu, Alessandro Valitutti, and Rafael Enrique Banchs. 2017. Humor in human-computer interaction: A short survey. In *IFIP TC13 International Conference on Human-Computer Interaction*.
- Nikola I. Nikolov, Eric Malmi, Curtis Northcutt, and Loreto Parisi. 2020. Rapformer: Conditional rap lyrics generation with denoising autoencoders. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 360–373, Dublin, Ireland. Association for Computational Linguistics.
- Hyungrae Noh. 2025. Beyond the responsibility gap: Distributed non-anthropocentric responsibility in the ai era. *Topoi*.
- Lukas Nolte and Sven Tomforde. 2025. A helping hand: A survey about ai-driven experimental design for accelerating scientific research. *Applied Sciences*, 15(9).
- Valeria Occelli, Gianluca Esposito, Paola Venuti, Giuseppe Maurizio Arduino, and Massimiliano Zampini. 2013. The takete—maluma phenomenon in autism spectrum disorders. *Perception*, 42(2):233–241. PMID: 23700961.
- Dieke Oele and Kilian Evang. 2017. BuzzSaw at SemEval-2017 task 7: Global vs. local context for interpreting and locating homographic English puns with sense embeddings.

BIBLIOGRAPHY

- In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 444–448, Vancouver, Canada. Association for Computational Linguistics.
- Katherine Ognyanova, David Lazer, Ronald E. Robertson, and Christo Wilson. 2020. Misinformation in action: Fake news exposure is linked to lower trust in media, higher trust in government when your side is in power. *The Harvard Kennedy School Misinformation Review*, 1:1–19.
- J Oh, K Choi, and Hitoshi Isahara. 2006. A comparison of different machine transliteration models. *Journal of Artificial Intelligence Research*, 27:119–151.
- Ken D. O’Halloran. 2020. A tongue-twister to translation? increased complexity of genioglossus movement during wakefulness in persons with obstructive sleep apnoea. *The Journal of Physiology*, 598(3):435–436.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, et al. 2023. Gpt-4 technical report.
- Silviu Oprea, Steven Wilson, and Walid Magdy. 2021. Chandler: An explainable sarcastic response generator. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 339–349, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Silviu Vlad Oprea, Steven Wilson, and Walid Magdy. 2022. Should a chatbot be sarcastic? understanding user preferences towards sarcasm generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7686–7700, Dublin, Ireland. Association for Computational Linguistics.
- Aitor Ormazabal, Mikel Artetxe, Manex Agirrezabal, Aitor Soroa, and Eneko Agirre. 2022. PoeLM: A meter- and rhyme-controllable language model for unsupervised

BIBLIOGRAPHY

- poetry generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3655–3670, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.
- Selcen Ozturkcan and Ayşe Aslı Bozdağ. 2025. Responsible ai in marketing: Ai booing and ai washing cycle of ai mistrust. *International Journal of Market Research*, 67(6):696–722.
- Victor Manuel Palma Preciado, Grigori Sidorov, Liana Ermakova, Anne-Gwenn Bosser, Tristan Miller, and Adam Jatowt. 2024. Overview of the clef 2024 joker task 2: Humour classification according to genre and technique. In *CEUR Workshop Proceedings*, volume 3740 of *GEUR Workshop Proceedings*. Presented at the 15th Conference and Labs of the Evaluation Forum (CLEF 2024).
- Vaishnavi Pamulapati, Gayatri Purigilla, and Radhika Mamidi. 2020. A novel annotation schema for conversational humor: Capturing the cultural nuances in kanyasulkam. In *Proceedings of the 14th Linguistic Annotation Workshop*, pages 34–47, Barcelona, Spain. Association for Computational Linguistics.
- Birong Pan, Yongqi Li, Weiyu Zhang, Wenpeng Lu, Mayi Xu, Shen Zhou, Yuanyuan Zhu, Ming Zhong, and Tiejun Qian. 2025. A survey on training-free alignment of large language models. In *Findings of the Association for Computational Linguistics*:

BIBLIOGRAPHY

- EMNLP 2025*, pages 4445–4461, Suzhou, China. Association for Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Jeongsik Park, Khoi P. N. Nguyen, Terrence Li, Suyesh Shrestha, Megan Kim Vu, Jerry Yining Wang, and Vincent Ng. 2024. MemeIntent: Benchmarking intent description generation for memes. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 631–643, Kyoto, Japan. Association for Computational Linguistics.
- Mihir Parmar, Swaroop Mishra, Mor Geva, and Chitta Baral. 2023. Don’t blame the annotator: Bias already starts in the annotation instructions. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1779–1789, Dubrovnik, Croatia. Association for Computational Linguistics.
- Ananya Passi and S. P. Arun. 2022. The bouba–kiki effect is predicted by sound properties but not speech properties. *Attention, Perception, & Psychophysics*.
- Pat Pataranutaporn, Chayapatr Archiwaranguprok, Samantha W. T. Chan, Elizabeth Loftus, and Pattie Maes. 2025. Synthetic human memories: Ai-edited images and videos can implant false memories and distort recollection. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI ’25, New York, NY, USA. Association for Computing Machinery.

BIBLIOGRAPHY

- Mohammed Johar Pathariya, Pratyush Basavraj Jalkote, Aniket Maharudra Patil, Abhishek Ashok Sutar, and Rajashree L. Ghule. 2024. Tunes by technology: A comprehensive survey of music generation models. In *2024 International Conference on Cognitive Robotics and Intelligent Systems (ICC - ROBINS)*, pages 506–512.
- Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent Disagreements in Human Textual Inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.
- Karl Pearson. 1895. Note on Regression and Inheritance in the Case of Two Parents. *Proceedings of the Royal Society of London Series I*, 58:240–242.
- Ivo Petrov, Jasper Dekoninck, Lyuben Baltadzhiev, Maria Drencheva, Kristian Minchev, Mislav Balunovic, Nikola Jovanović, and Martin Vechev. 2025. Proof or bluff? evaluating LLMs on 2025 USA math olympiad. In *2nd AI for Math Workshop @ ICML 2025*.
- Saša Petrović and David Matthews. 2013. Unsupervised joke generation from big data. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 228–232, Sofia, Bulgaria. Association for Computational Linguistics.
- Pew Research Center. 2025. How the U.S. Public and AI Experts View Artificial Intelligence. Report, Pew Research Center.
- Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017, Mexico City, Mexico. Association for Computational Linguistics.

BIBLIOGRAPHY

- Louise H. Phillips, Roy Allen, Rebecca Bull, Alexandra Hering, Matthias Kliegel, and Shelley Channon. 2015. Older adults have difficulty in decoding sarcasm. *Developmental Psychology*, 51(12):1840–1852.
- Thomas Pickard, Tyler Loakman, and Mugdha Pandya. 2023. shefnlp at SemEval-2023 task 10: Compute-efficient category adapters. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1069–1075, Toronto, Canada. Association for Computational Linguistics.
- Barbara Plank. 2022. The “problem” of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Plato. 1892. *Philebus*. Clarendon Press, Oxford. First composed c. 360-347 BCE.
- Artem Ploujnikov and Mirco Ravanelli. 2022. Soundchoice: Grapheme-to-phoneme models with semantic disambiguation. In *Interspeech 2022*, pages 486–490.
- Sheila Pontis, Ann Blandford, Elke Greifeneder, Hesham Attalla, and David Neal. 2017. Keeping up to date: An academic researcher’s information journey. *Journal of the Association for Information Science and Technology*, 68(1):22–35.
- Andrei Popescu-Belis, Àlex R. Atrio, Bastien Bernath, Etienne Boisson, Teo Ferrari, Xavier Theimer-Lienhard, and Giorgos Vernikos. 2023. GPoeT: a language model trained for rhyme generation on synthetic data. In *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 10–20.
- Matt Post and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of the 2018 Conference*

BIBLIOGRAPHY

- of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1314–1324, New Orleans, Louisiana. Association for Computational Linguistics.
- Peter Potash, Alexey Romanov, and Anna Rumshisky. 2015. GhostWriter: Using an LSTM for automatic rap lyric generation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1919–1924, Lisbon, Portugal. Association for Computational Linguistics.
- Peter Potash, Alexey Romanov, and Anna Rumshisky. 2018. Evaluating creative language generation: The case of rap lyric ghostwriting. In *Proceedings of the Second Workshop on Stylistic Variation*, pages 29–38, New Orleans. Association for Computational Linguistics.
- Marc Prats, Sandra Smid, and Monica Ferrin. 2024. Lack of trust in institutions and political engagement: An analysis based on the 2021 oecd trust survey. *OECD Working Papers on Public Governance*, (75).
- PwC. 2025. The fearless future: 2025 global ai jobs barometer. Report, PwC.
- Mahta Fetrat Qharabagh, Zahra Dehghanian, and Hamid R. Rabiee. 2025a. Fast, not fancy: Rethinking G2P with rich data and statistical models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22382–22408, Suzhou, China. Association for Computational Linguistics.
- Mahta Fetrat Qharabagh, Zahra Dehghanian, and Hamid R Rabiee. 2025b. Llm-powered grapheme-to-phoneme conversion: Benchmark and case study. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

BIBLIOGRAPHY

- Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 Technical Report.
- Ella Rabinovich, Hila Gonen, and Suzanne Stevenson. 2020. Pick a fight or bite your tongue: Investigation of gender differences in idiomatic language usage. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5181–5192, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.
- Kristina Radivojevic, Matthew Chou, Karla Badillo-Urquiola, and Paul Brenner. 2024a. Human perception of llm-generated text content in social media environments.
- Kristina Radivojevic, Nicholas Clark, and Paul Brenner. 2024b. Llms among us: Generative ai participating in digital discourse. *Proceedings of the AAAI Symposium Series*, 3(1):209–218.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer

BIBLIOGRAPHY

- learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.
- Elnaz Rahmati and Hossein Sameti. 2024. GE2PE: Persian end-to-end grapheme-to-phoneme conversion. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3426–3436, Miami, Florida, USA. Association for Computational Linguistics.
- V S Ramachandran and Edward M Hubbard. 2001. Synaesthesia — a window into perception, thought and language. *Journal of Consciousness Studies*, 8(12):3–34.
- Ananth Ramakrishnan A, Sankar Kuppan, and Sobha Lalitha Devi. 2009. Automatic generation of Tamil lyrics for melodies. In *Proceedings of the Workshop on Computational Approaches to Linguistic Creativity*, pages 40–46, Boulder, Colorado. Association for Computational Linguistics.
- C. Ramakristanaiah, P. Namratha, Rajendra Kumar Ganiya, and Midde Ranjit Reddy. 2021. A survey on humor detection methods in communications. In *2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, pages 668–674.
- Aida Ramezani and Yang Xu. 2023. Knowledge of cultural moral norms in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 428–446, Toronto, Canada. Association for Computational Linguistics.
- Victor Raskin. 1985. *Semantic mechanisms of humor*. Synthese language library ; 24. Reidel, Dordrecht.
- Walter Redfern. 1987. Puns. *Journal of English Linguistics*, 20(1):158–158.

BIBLIOGRAPHY

- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.
- Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401.
- Pascual Restrepo. 2025. *We Won't Be Missed: Work and Growth in the AGI World*. University of Chicago Press.
- Sam Ribeiro, Giulia Comini, and Jaime Lorenzo-Trueba. 2023. Improving grapheme-to-phoneme conversion by learning pronunciations from speech recordings. In *Interspeech 2023*, pages 999–1003.
- Right to Warn Coalition. 2024. A right to warn about advanced artificial intelligence.
- Graeme Ritchie. 2001. Current Directions in Computational Humour. *Artificial Intelligence Review*, 16(2):119–135.
- Graeme Ritchie. 2005. Computational mechanisms for pun generation. In *Proceedings of the Tenth European Workshop on Natural Language Generation (ENLG-05)*, Aberdeen, Scotland. Association for Computational Linguistics.
- Graeme D Ritchie, Ruli Manurung, Helen Pain, Annalu Waller, Rolf Black, and Dave O'Mara. 2007. A practical application of computational humour. In *Proceedings of the Fourth International Joint Conference on Computational Creativity (Goldsmith's, London)*, pages 91–98.
- Hannes Ritschel and Elisabeth André. 2018. Shaping a social robot's humor with natural language generation and socially-aware reinforcement learning. In *Proceedings*

BIBLIOGRAPHY

- of the Workshop on NLG for Human–Robot Interaction*, pages 12–16, Tilburg, The Netherlands. Association for Computational Linguistics.
- Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, et al. 2024. Gemma 2: Improving Open Language Models at a Practical Size.
- Patricia Rockwell and Evelyn M. Theriot. 2001. Culture, gender, and gender mix in encoders of sarcasm: A self-assessment analysis. *Communication Research Reports*, 18(1):44–52.
- Henry Rogers. 2013. *The sounds of language : an introduction to phonetics*, 1st ed. edition. Learning about Language. Routledge, London.
- José Carlos Rosales Núñez, Djamé Seddah, and Guillaume Wisniewski. 2019. Phonetic normalization for machine translation of user generated content. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 407–416, Hong Kong, China. Association for Computational Linguistics.
- Eleonora Rosati. 2019. Copyright as an obstacle or an enabler? a european perspective on text and data mining and its role in the development of ai creativity*. *Asia Pacific Law Review*, 27(2):198–217.
- Eleonora Rosati. 2025. Copyright exceptions and fair use defences for ai training done for “research” and “learning,” or the inescapable licensing horizon. *European Journal of Risk Regulation*, 16(3):961–984.
- Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. 2010. Automatic keyword extraction from individual documents. In *Text Mining*, pages 1–20. John Wiley & Sons, Ltd.

BIBLIOGRAPHY

- Allen Roush, Sanjay Basu, Akshay Moorthy, and Dmitry Dubovoy. 2022. Most language models can be poets too: An AI writing assistant and constrained text generation studio. In *Proceedings of the Second Workshop on When Creative AI Meets Conversational AI*, pages 9–15, Gyeongju, Republic of Korea. Association for Computational Linguistics.
- James Route, Steven Hillis, Isak Czeresnia Etinger, Han Zhang, and Alan W Black. 2019. Multimodal, multilingual grapheme-to-phoneme conversion for low-resource languages. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 192–201, Hong Kong, China. Association for Computational Linguistics.
- Mark A. Runco. 2023. Ai can only produce artificial creativity. *Journal of Creativity*, 33(3):100063.
- Michael J Ryan, William Held, and Diyi Yang. 2024. Unintended impacts of LLM alignment on global representation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16121–16140, Bangkok, Thailand. Association for Computational Linguistics.
- Valentino Sabbatino, Enrica Troiano, Antje Schweitzer, and Roman Klinger. 2022. “splink” is happy and “phrouth” is scary: Emotion intensity analysis for nonsense words. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 37–50, Dublin, Ireland. Association for Computational Linguistics.
- Jivnesh Sandhan, Amruta Barbadikar, Malay Maity, Pavankumar Satuluri, Tushar Sandhan, Ravi M Gupta, Pawan Goyal, and Laxmidhar Behera. 2025. Aesthetics of Sanskrit poetry from the perspective of computational linguistics: A case study analy-

BIBLIOGRAPHY

- sis on śikṣāṣṭaka. In *Computational Sanskrit and Digital Humanities - World Sanskrit Conference 2025*, pages 15–36, Kathmandu, Nepal. Association for Computational Linguistics.
- Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M. Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Tali Bers, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask Prompted Training Enables Zero-Shot Task Generalization.
- Edward Sapir. 1929. A study in phonetic symbolism. *Journal of Experimental Psychology*, 12(3):225–239.
- Stephanie Schoch, Diyi Yang, and Yangfeng Ji. 2020. “this is a problem, don’t you agree?” framing and bias in human evaluation for natural language generation. In *Proceedings of the 1st Workshop on Evaluating NLG Evaluation*, pages 10–16, Online (Dublin, Ireland). Association for Computational Linguistics.
- Annika M Schoene and Cansu Canca. 2025. ‘for argument’s sake, show me how to harm myself!’: Jailbreaking llms in suicide and self-harm contexts.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms.
- Eric Schwitzgebel and Mara Garza. 2020. Designing ai with rights, consciousness, self-respect, and freedom. In *Ethics of Artificial Intelligence*. Oxford University Press.

BIBLIOGRAPHY

- Neil Selwyn. 2022. The future of ai and education: Some cautionary notes. *European Journal of Education*, 57(4):620–631.
- Neil Selwyn and Beatriz Gallo Cordoba. 2022. Australian public understandings of artificial intelligence. *AI & SOCIETY*, 37(4):1645–1662.
- Sergio Servantez, Joe Barrow, Kristian Hammond, and Rajiv Jain. 2024. Chain of logic: Rule-based reasoning with large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2721–2733, Bangkok, Thailand. Association for Computational Linguistics.
- Weiqiao Shan, Yang Li, Yuhao Zhang, Yingfeng Luo, Chen Xu, Xiaofeng Zhao, Long Meng, Yunfei Lu, Min Zhang, Hao Yang, Tong Xiao, and JingBo Zhu. 2025. Enhancing speech large language models with prompt-aware mixture of audio encoders. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19305–19320, Suzhou, China. Association for Computational Linguistics.
- Chen Shani, Alexander Libov, Sofia Tolmach, Liane Lewin-Eytan, Yoelle Maarek, and Dafna Shahaf. 2022. “alexa, do you want to build a snowman?” characterizing playful requests to conversational agents. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI EA ’22, New York, NY, USA. Association for Computing Machinery.
- Natalie Shapira, Oren Kalinsky, Alexander Libov, Chen Shani, and Sofia Tolmach. 2023. Evaluating humorous response generation to playful shopping requests. In *ECIR 2023*.
- Rahul Sharma, Kunal Dhawan, and Balakrishna Pailla. 2021. Phonetic word embeddings.

BIBLIOGRAPHY

- Zhonghao Sheng, Kaitao Song, Xu Tan, Yi Ren, Wei Ye, Shikun Zhang, and Tao Qin. 2020. Songmass: Automatic song writing with pre-training and alignment constraint. In *AAAI Conference on Artificial Intelligence*.
- Henry Shevlin. 2024. All too human? identifying and mitigating ethical risks of social ai. *Law, Ethics & Technology*, 1(2).
- Boaz Shmueli, Jan Fell, Soumya Ray, and Lun-Wei Ku. 2021. Beyond fair pay: Ethical implications of NLP crowdsourcing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3758–3769, Online. Association for Computational Linguistics.
- Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. 2025. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. In *NeurIPS*.
- Dong Shu, Haiyan Zhao, Jingyu Hu, Weiru Liu, Ali Payani, Lu Cheng, and Mengnan Du. 2025. Large vision-language model alignment and misalignment: A survey through the lens of explainability. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1713–1735, Suzhou, China. Association for Computational Linguistics.
- David M. Sidhu, Angeliki Athanasopoulou, Stephanie L. Archer, Natalia Czarnecki, Suzanne Curtin, and Penny M. Pexman. 2023. The maluma/takete effect is late: No longitudinal evidence for shape sound symbolism in the first year. *PLOS ONE*, 18(11):1–23.

BIBLIOGRAPHY

- David M. Sidhu and Gabriella Vigliocco. 2023. I don’t see what you’re saying: The maluma/takete effect does not depend on the visual appearance of phonemes as they are articulated. *Psychonomic Bulletin & Review*, 30(4):1521–1529.
- David M. Sidhu, Chris Westbury, Geoff Hollis, and Penny M. Pexman. 2021. Sound symbolism shapes the english language: The maluma/takete effect in english nouns. *Psychonomic Bulletin & Review*, 28(4):1390–1398.
- Gary F. Simons, Abbey L. L. Thomas, and Chad K. K. White. 2022. Assessing digital language support on a global scale. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4299–4305, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Mona Sloane and Elena Wüllhorst. 2025. A systematic review of regulatory strategies and transparency mandates in ai regulation in europe, the united states, and canada. *Data and Policy*, 7:e11.
- E A Smith and R J Senter. 1967. Automated readability index. *AMRL-TR. Aerospace Medical Research Laboratories (6570th)*, pages 1–14.
- Aaron Smuts. 2010. The ethics of humor: Can your sense of humor be wrong? *Ethical theory and moral practice*, 13(3):333–347.
- Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2025. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth International Conference on Learning Representations*.
- Anders Søgaard. 2022. Should we ban English NLP for a year? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5254–5260, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

BIBLIOGRAPHY

- Victoria Somoff. 2014. Four is not fourteen: Tongue twister patterns and the unmastery of language. *Western Folklore*, 73(2/3):195–215.
- Charles Spence. 2012. Managing sensory expectations concerning products and brands: Capitalizing on the potential of sound and shape symbolism. *Journal of consumer psychology*, 22(1):37–54.
- Herbert Spencer. 1875. The physiology of laughter. In *Illustrations of universal progress: A series of discussions.*, pages 194–209. D Appleton & Company.
- M. R. G. Spiller. 1992. *The Development of the Sonnet: An Introduction*. Routledge.
- Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to summarize from human feedback. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20*, Red Hook, NY, USA. Curran Associates Inc.
- Oliviero Stock and Carlo Strapparava. 2005. HAHAAcronym: A computational humor system. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 113–116, Ann Arbor, Michigan. Association for Computational Linguistics.
- Suzy J. Styles and Lauren Gawne. 2017. When does maluma/takete fail? two key failures and a meta-analysis suggest that phonology and phonotactics matter. *i-Perception*, 8(4). PMID: 28890777.
- Prasetyawan Sugiharto, Yan Santoso, and Maila Shofyana. 2022. Teaching english pronunciation using tongue twister. *Acitya: Journal of Teaching and Education*, 4(1):189–197.

BIBLIOGRAPHY

- Jiao Sun, Anjali Narayan-Chen, Shereen Oraby, Alessandra Cervone, Tagyoung Chung, Jing Huang, Yang Liu, and Nanyun Peng. 2022a. ExpUNations: Augmenting puns with keywords and explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4590–4605, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Jiao Sun, Anjali Narayan-Chen, Shereen Oraby, Shuyang Gao, Tagyoung Chung, Jing Huang, Yang Liu, and Nanyun Peng. 2022b. Context-situated pun generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4635–4648, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Yusen Sun, Liangyou Li, Qun Liu, and Dit-Yan Yeung. 2023. SongRewriter: A Chinese song rewriting system with controllable content and rhyme scheme. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12863–12880, Toronto, Canada. Association for Computational Linguistics.
- Mukuntha Narayanan Sundararaman, Ayush Kumar, and Jithendra Vepa. 2021. Phone-meBERT: Joint language modelling of phoneme sequence and asr transcript. In *Interspeech 2021*, pages 3236–3240.
- Raymond Hendy Susanto, Shamil Chollampatt, and Liling Tan. 2020. Lexically constrained neural machine translation with Levenshtein transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3536–3543, Online. Association for Computational Linguistics.
- Ashima Suvarna, Harshita Khandelwal, and Nanyun Peng. 2024. PhonologyBench: Evaluating phonological skills of large language models. In *Proceedings of the 1st*

BIBLIOGRAPHY

- Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 1–14, Bangkok, Thailand. Association for Computational Linguistics.
- Chen Tang, Chenghua Lin, Henglin Huang, Frank Guerin, and Zhihao Zhang. 2022a. EtriCA: Event-triggered context-aware story generation augmented by cross attention. In *Findings of the Association for Computational Linguistics: EMNLP 2022*.
- Chen Tang, Tyler Loakman, and Chenghua Lin. 2024a. A cross-attention augmented model for event-triggered context-aware story generation. *Computer Speech & Language*, 88:101662.
- Chen Tang, Hongbo Zhang, Tyler Loakman, Chenghua Lin, and Frank Guerin. 2023a. Enhancing dialogue generation via dynamic graph knowledge aggregation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4604–4616.
- Chen Tang, Hongbo Zhang, Tyler Loakman, Chenghua Lin, and Frank Guerin. 2023b. Terminology-aware medical dialogue generation. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.
- Chen Tang, Hongbo Zhang, Tyler Loakman, Bohao Yang, Stefan Goetze, and Chenghua Lin. 2024b. CADGE: Context-aware dialogue generation enhanced with graph-structured knowledge aggregation. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 371–383, Tokyo, Japan. Association for Computational Linguistics.
- Chen Tang, Zhihao Zhang, Tyler Loakman, Chenghua Lin, and Frank Guerin. 2022b. NGEF: A graph-based event planning framework for story generation. In *Proceedings of AACL-IJCNLP*, pages 186–193, Online. Association for Computational Linguistics.

BIBLIOGRAPHY

- Russell Taylor, Benjamin Herbert, and Michael Sana. 2025. Pun intended: Multi-agent translation of wordplay with contrastive learning and phonetic-semantic embeddings.
- Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context.
- Stina Teilmann-Lock and Andrej Savin. 2025. Beyond the ai-copyright wars: towards european dataset law? *Computer Law & Security Review*, 58:106190.
- Maria Teleki, Vedangi Bengali, Xiangjue Dong, Sai Tejas Janjur, Haoran Liu, Tian Liu, Cong Wang, Ting Liu, Yin Zhang, Frank Shipman, and James Caverlee. 2025. A survey on LLMs for story generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 13954–13966, Suzhou, China. Association for Computational Linguistics.
- Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguerre-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications.

BIBLIOGRAPHY

- Yuanhe Tian, Renze Lou, Xiangyu Pang, Lianxi Wang, Shengyi Jiang, and Yan Song. 2022a. Improving English-Arabic transliteration with phonemic memories. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3262–3272, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Yufei Tian, Anjali Narayan-Chen, Shereen Oraby, Alessandra Cervone, Gunnar Sigurdsson, Chenyang Tao, Wenbo Zhao, Yiwen Chen, Tagyoung Chung, Jing Huang, and Nanyun Peng. 2023. Unsupervised melody-to-lyrics generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9235–9254, Toronto, Canada. Association for Computational Linguistics.
- Yufei Tian and Nanyun Peng. 2022. Zero-shot sonnet generation with discourse-level planning and aesthetics features. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3587–3597, Seattle, United States. Association for Computational Linguistics.
- Yufei Tian, Divyanshu Sheth, and Nanyun Peng. 2022b. A unified framework for pun generation with humor principles. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3253–3261, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Yufei Tian, Arvind krishna Sridhar, and Nanyun Peng. 2021. HypoGen: Hyperbole generation with commonsense and counterfactual knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1583–1593, Punta Cana, Dominican Republic. Association for Computational Linguistics.

BIBLIOGRAPHY

- Aleksey Tikhonov and Ivan P. Yamshchikov. 2018. Sounds wilde. phonetically extended embeddings for author-stylized poetry generation. In *Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 117–124, Brussels, Belgium. Association for Computational Linguistics.
- Nitasha Tiku. 2022. The google engineer who thinks the company’s ai has come to life. *The Washington Post*.
- Riccardo Tommasini, Filip Ilievski, and Thilini Wijesiriwardene. 2023. IMKG: The Internet Meme Knowledge Graph. In Catia Pesquita, Ernesto Jimenez-Ruiz, Jamie McCusker, Daniel Faria, Mauro Dragoni, Anastasia Dimou, Raphael Troncy, and Sven Hertling, editors, *The Semantic Web*, volume 13870, pages 354–371. Springer Nature Switzerland, Cham.
- Joe Toplyn. 2023. Witscript 3: A hybrid ai system for improvising jokes in a conversation.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*.
- Paul Trichelair, Ali Emami, Adam Trischler, Kaheer Suleman, and Jackie Chi Kit Cheung. 2019. How reasonable are common-sense reasoning tasks: A case-study on the Winograd schema challenge and SWAG. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3382–3387, Hong Kong, China. Association for Computational Linguistics.
- Sean Trott. 2024. Can large language models help augment english psycholinguistic datasets? *Behavior Research Methods*.
- Yulia Tsvetkov, Sunayana Sitaram, Manaal Faruqui, Guillaume Lample, Patrick Littell, David Mortensen, Alan W Black, Lori Levin, and Chris Dyer. 2016. Polyglot neural

BIBLIOGRAPHY

- language models: A case study in cross-lingual phonetic representation learning. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1357–1366, San Diego, California. Association for Computational Linguistics.
- S. Tu, X. Cao, X. Yun, K. Wang, G. Zhao, and J. Qiu. 2014. A new association evaluation stage in cartoon apprehension: Evidence from an erp study. *Journal of Behavioral and Brain Science*, 4:75–83.
- Lewis Tunstall, Edward Emanuel Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro Von Werra, Cl  mentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M Rush, and Thomas Wolf. 2024. Zephyr: Direct distillation of LM alignment. In *First Conference on Language Modeling*.
- Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2022. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.
- Dmitri Usnadze. 1924. Ein experimenteller beitrag zum problem der psychologischen grundlagen der namengebung. pages 24–43.
- Mathieu Valette. 2024. What does perspectivism mean? an ethical and methodological countercriticism. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 111–115, Torino, Italia. ELRA and ICCL.
- Alessandro Valitutti, Hannu Toivonen, Antoine Doucet, and Jukka M. Toivanen. 2013. “let everything turn well in your wife”: Generation of adult humor using lexical

BIBLIOGRAPHY

- constraints. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 243–248, Sofia, Bulgaria. Association for Computational Linguistics.
- Tim Van de Cruys. 2020. Automatic poetry generation from prosaic text. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2471–2480, Online. Association for Computational Linguistics.
- Sven van den Beukel and Lora Aroyo. 2018. Homonym detection for humor recognition in short text. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 286–291, Brussels, Belgium. Association for Computational Linguistics.
- Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan. Association for Computational Linguistics.
- Stephan van der Zwaard, Arie-Willem de Leeuw, L. (Rens) A. Meerhoff, Sue C. Bodine, and Arno Knobbe. 2020. Articles with impact: insights into 10 years of research with machine learning. *Journal of Applied Physiology*, 129(4):967–979. PMID: 32790596.
- Cynthia Van Hee. 2017. *Can machines sense irony? : exploring automatic irony detection on social media*. Ph.D. thesis, Ghent University.
- Robert D. Van Valin. 2001. *An introduction to syntax*, first edition. edition. Cambridge University Press, Cambridge.
- S. Vaughn. 1904. *Limerick Lyrics: A Collection of Over 700 Choice Versifications, to which is Added a Number of Short Verses, from Many Sources*. T.J. Carey & Company.

BIBLIOGRAPHY

- Herman Veluwenkamp. 2025. What responsibility gaps are and what they should be. *Ethics and Information Technology*, 27(1):14.
- An Vo, Khai-Nguyen Nguyen, Mohammad Reza Taesiri, Vy Tuong Dang, Anh Totti Nguyen, and Daeyoung Kim. 2025. Vision language models are biased: Counting legs of an animal is surprisingly hard. In *2nd AI for Math Workshop @ ICML 2025*.
- Byron C Wallace, Laura Kertz, Eugene Charniak, et al. 2014. Humans require context to infer ironic intent (so computers probably do, too). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 512–516.
- Evan Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Allyson Ettinger, Michal Guerquin, David Heineman, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James Validad Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Jake Poznanski, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2025. 2 OLMo 2 furious (COLM’s version). In *Second Conference on Language Modeling*.
- Melanie Walsh, Anna Preus, and Maria Antoniak. 2024. Sonnet or not, bot? poetry evaluation for large models and datasets. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15568–15603, Miami, Florida, USA. Association for Computational Linguistics.

BIBLIOGRAPHY

- Yoshija Walter. 2025. Artificial influencers and the dead internet theory. *AI & SOCIETY*, 40(1):239–240.
- Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F. Chen. 2025a. AudioBench: A universal benchmark for audio large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4297–4316, Albuquerque, New Mexico. Association for Computational Linguistics.
- Boshi Wang, Xiang Deng, and Huan Sun. 2022. Iteratively prompt pre-trained language models for chain of thought. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2714–2730, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Chunfeng Wang, Peisong Huang, Yuxiang Zou, Haoyu Zhang, Shichao Liu, Xiang Yin, and Zejun Ma. 2023a. Liteg2p: A fast, light and high accuracy model for grapheme-to-phoneme conversion. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Minghan Wang, Hao Yang, Ying Qin, Shiliang Sun, and Yao Deng. 2020. Unified humor detection based on sentence-pair augmentation and transfer learning. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 53–59, Lisboa, Portugal. European Association for Machine Translation.
- Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024a. Math-shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the*

BIBLIOGRAPHY

- Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9426–9439, Bangkok, Thailand. Association for Computational Linguistics.
- Shun Wang, Yucheng Li, Chenghua Lin, Loic Barrault, and Frank Guerin. 2023b. Metaphor detection with effective context denoising. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1404–1409.
- Shun Wang, Ge Zhang, Han Wu, Tyler Loakman, Wenhao Huang, and Chenghua Lin. 2024b. MMTE: Corpus and metrics for evaluating machine translation quality of metaphorical language. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11343–11358, Miami, Florida, USA. Association for Computational Linguistics.
- Yang Wang, Chenghao Xiao, Chia-Yi Hsiao, Zi Yan Chang, Chi-Li Chen, Tyler Loakman, and Chenghua Lin. 2025b. Drivel-ology: Challenging LLMs with interpreting nonsense with depth. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23085–23107, Suzhou, China. Association for Computational Linguistics.
- Zhexu Wang, Yiping Liu, Yejie Wang, Wenyang He, Bofei Gao, Muxi Diao, Yanxu Chen, Kelin Fu, Flood Sung, Zhilin Yang, Tianyu Liu, and Weiran Xu. 2025c. OJBench: A Competition Level Code Benchmark For Large Language Models.
- Zihan Wang, Yunxuan Li, Yuexin Wu, Liangchen Luo, Le Hou, Hongkun Yu, and Jingbo Shang. 2024c. Multi-step problem solving through a verifier: An empirical analysis on model-induced process supervision. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7309–7319, Miami, Florida, USA. Association for Computational Linguistics.

BIBLIOGRAPHY

- Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: A benchmark of linguistic minimal pairs for English. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 409–410, New York, New York. Association for Computational Linguistics.
- Kento Watanabe and Masataka Goto. 2020. Lyrics information processing: Analysis, generation, and applications. In *Proceedings of the 1st Workshop on NLP for Music and Audio (NLP4MusA)*, pages 6–12, Online. Association for Computational Linguistics.
- Kento Watanabe, Yuichiroh Matsubayashi, Satoru Fukayama, Masataka Goto, Kentaro Inui, and Tomoyasu Nakano. 2018. A melody-conditioned lyrics language model. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 163–172, New Orleans, Louisiana. Association for Computational Linguistics.
- Maurice Weber, Daniel Y. Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, Ben Athiwaratkun, Rahul Chalamala, Kezhen Chen, Max Ryabinin, Tri Dao, Percy Liang, Christopher Ré, Irina Rish, and Ce Zhang. 2024. Redpajama: an open dataset for training large language models. *NeurIPS Datasets and Benchmarks Track*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting

BIBLIOGRAPHY

- elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22*, Red Hook, NY, USA. Curran Associates Inc.
- Sheng-Lun Wei, Cheng-Kuang Wu, Hen-Hsen Huang, and Hsin-Hsi Chen. 2024. Unveiling selection biases: Exploring order and token sensitivity in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5598–5621, Bangkok, Thailand. Association for Computational Linguistics.
- Orion Weller, Nancy Fulda, and Kevin Seppi. 2020. Can humor prediction datasets be used for humor generation? humorous headline generation via style transfer. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 186–191, Online. Association for Computational Linguistics.
- Orion Weller and Kevin Seppi. 2019. Humor detection: A transformer gets the last laugh. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3621–3625, Hong Kong, China. Association for Computational Linguistics.
- Orion Weller and Kevin Seppi. 2020. The rJokes dataset: a large scale humor collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6136–6141, Marseille, France. European Language Resources Association.
- J. C. Wells. 1982a. *Accents of English: Beyond the British Isles*, volume 3 of *Accents of English*. Cambridge University Press, Cambridge.
- J. C. Wells. 1982b. *Accents of English: The British Isles*, volume 2 of *Accents of English*. Cambridge University Press, Cambridge.

BIBLIOGRAPHY

- John C. Wells. 1995. Computer-coding the ipa: a proposed extension of sampa.
- Chenxi Whitehouse, Monojit Choudhury, and Alham Aji. 2023. LLM-powered data augmentation for enhanced cross-lingual performance. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 671–686.
- Michael Wierzbicki and Richard David Young. 1978. The relation of intelligence and task difficulty to appreciation of humor. *The Journal of General Psychology*, 99(1):25–32.
- Angus R Williams, Liam Burke-Moore, Ryan Sze-Yin Chan, Florence E. Enock, Federico Nanni, Tvesha Sippy, Yi-Ling Chung, Evelina Gabasova, Kobi Hackenburg, and Jonathan Bright. 2025. Large language models can consistently generate high-quality content for election disinformation operations. *PLOS ONE*, 20(3):1–29.
- Carolyn E. Wilshire. 1999. The “tongue twister” paradigm as a technique for studying phonological encoding. *Language and Speech*, 42(1):57–82.
- Bodo Winter, Gary Lupyan, Lynn K. Perry, Mark Dingemanse, and Marcus Perlman. 2023. Iconicity ratings for 14,000+ english words. *Behavior Research Methods*.
- Jörg Wöckener, Thomas Haider, Tristan Miller, The-Khang Nguyen, Thanh Tung Linh Nguyen, Minh Vu Pham, Jonas Belouadi, and Steffen Eger. 2021. End-to-end style-conditioned poetry generation: What does it take to learn from examples alone? In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 57–66, Punta Cana, Dominican Republic (online). Association for Computational Linguistics.
- Lukas Wolf, Tiago Pimentel, Evelina Fedorenko, Ryan Cotterell, Alex Warstadt, Ethan Gotlieb Wilcox, and Tamar I. Regev. 2023. Quantifying the redundancy between prosody and text. In *Proceedings of the 2023 Conference on Empirical*

BIBLIOGRAPHY

- Methods in Natural Language Processing*, pages 9765–9784, Singapore. Association for Computational Linguistics.
- Min Ney Wong, Yanky Chan, Manwa L. Ng, and Frank F. Zhu. 2019. Effects of transcranial direct current stimulation over the broca’s area on tongue twister production. *International Journal of Speech-Language Pathology*, 21(2):182–188. PMID: 29642741.
- Ernest Vincent Wright. 2016. *Gadsby: A story of over 50,000 words without using the letter “E”*. Digital Ninjas Media, Inc.
- Chih-Chiang Wu and Wei-Peng Chen. 2022. What’s an ai name worth? the impact of ai etfs on their underlying stocks. *Finance Research Letters*, 46:102474.
- Shih-Hung Wu, Yu-Feng Huang, and Tsz-Yeung Lau. 2024a. Humour classification by fine-tuning llms: Cyut at clef 2024 joker lab subtask humour classification according to genre and technique. In *CEUR Workshop Proceedings*, volume 3740 of *GEUR Workshop Proceedings*. Notebook for the CYUT Lab at CLEF 2024.
- Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. 2022. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, 135:364–381.
- Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. 2024b. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1819–1862, Mexico City, Mexico. Association for Computational Linguistics.

BIBLIOGRAPHY

- Nan Xu and Xuezhe Ma. 2025. LLM the genius paradox: A linguistic and math expert’s struggle with simple word-based counting problems. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3344–3370, Albuquerque, New Mexico. Association for Computational Linguistics.
- Zhangchen Xu, Yang Liu, Yueqin Yin, Mingyuan Zhou, and Radha Poovendran. 2025. KodCode: A diverse, challenging, and verifiable synthetic dataset for coding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6980–7008, Vienna, Austria. Association for Computational Linguistics.
- Zhijun Xu, Siyu Yuan, Lingjie Chen, and Deqing Yang. 2024. “a good pun is its own reword”: Can large language models understand puns? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11766–11782, Miami, Florida, USA. Association for Computational Linguistics.
- Lanqing Xue, Kaitao Song, Duocai Wu, Xu Tan, Nevin L. Zhang, Tao Qin, Wei-Qiang Zhang, and Tie-Yan Liu. 2021. DeepRapper: Neural rap generation with rhyme and rhythm modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 69–81, Online. Association for Computational Linguistics.
- Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

BIBLIOGRAPHY

- Aditya Yadavalli, Tiago Pimentel, Tamar I Regev, Ethan Wilcox, and Alex Warstadt. 2025. What do prosody and text convey? characterizing how meaningful information is distributed across multiple channels.
- Ivan P. Yamshchikov, Viascheslav Shibaev, and Alexey Tikhonov. 2019. Dyr bul shchyl. proxying sound symbolism with word embeddings. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 90–94, Minneapolis, USA. Association for Computational Linguistics.
- Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Haizhou Zhao, Hang Xu, Haoze Sun, Hongda Zhang, Hui Liu, Jiaming Ji, Jian Xie, JunTao Dai, Kun Fang, Lei Su, Liang Song, Lifeng Liu, Liyun Ru, Luyao Ma, Mang Wang, Mickel Liu, MingAn Lin, Nuolan Nie, Peidong Guo, Ruiyang Sun, Tao Zhang, Tianpeng Li, Tianyu Li, Wei Cheng, Weipeng Chen, Xiangrong Zeng, Xiaochuan Wang, Xiaoxi Chen, Xin Men, Xin Yu, Xuehai Pan, Yanjun Shen, Yiding Wang, Yiyu Li, Youxin Jiang, Yuchen Gao, Yupeng Zhang, Zenan Zhou, and Zhiying Wu. 2023. Baichuan 2: Open large-scale language models.
- Bohao Yang, Chen Tang, and Chenghua Lin. 2024a. Improving medical dialogue generation with abstract meaning representations. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11826–11830.
- Bohao Yang, Chen Tang, Kun Zhao, Chenghao Xiao, and Chenghua Lin. 2024b. Effective distillation of table-based reasoning ability from LLMs. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5538–5550.

BIBLIOGRAPHY

- Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and Jingren Zhou. 2024c. AIR-bench: Benchmarking large audio-language models via generative comprehension. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1979–1998, Bangkok, Thailand. Association for Computational Linguistics.
- Shunyu Yao, Howard Chen, Austin W. Hanjie, Runzhe Yang, and Karthik R Narasimhan. 2024. COLLIE: Systematic construction of constrained text generation tasks. In *The Twelfth International Conference on Learning Representations*.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models.
- Xupeng Yao, Wenxiao Du, Li Sun, and Bin Hu. 2025. Automatic trailer generation for movies using convolutional neural network. *Scientific Reports*, 15(1):7819.
- Kenneth Yasuda. 1973. *The Japanese haiku : its essential nature, history, and possibilities in English, with selected examples*, tut book ed.. edition. Charles E. Tuttle, Rutland, Vermont.
- Zhangyue Yin, Qiushi Sun, Qipeng Guo, Zhiyuan Zeng, Xiaonan Li, Junqi Dai, Qinyuan Cheng, Xuanjing Huang, and Xipeng Qiu. 2024. Reasoning in flux: Enhancing large language models reasoning through uncertainty-aware adaptive guidance. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2401–2416, Bangkok, Thailand. Association for Computational Linguistics.

BIBLIOGRAPHY

- Zheng Xin Yong, Vineel Pratap, Michael Auli, and Jean Maillard. 2025. Effects of Speaker Count, Duration, and Accent Diversity on Zero-Shot Accent Robustness in Low-Resource ASR. In *Interspeech 2025*, pages 1148–1152.
- Eric Yorkston and Geeta Menon. 2004. A sound idea: Phonetic effects of brand names on consumer judgments. *Journal of Consumer Research*, 31(1):43–51.
- Khalil Yousef and Shelley Schmollgruber. 2024. Artificial intelligence in low- and middle-income countries: Reducing the gaps in health care, research, and education. *International Journal of Critical Care*, 18(2):1–3.
- Dingyao Yu, Kaitao Song, Peiling Lu, Tianyu He, Xu Tan, Wei Ye, Shikun Zhang, and Jiang Bian. 2023. MusicAgent: An AI agent for music understanding and generation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 246–255.
- Jiaxing Yu, Xinda Wu, Yunfei Xu, Tiejiao Zhang, Songruoyao Wu, Le Ma, and Kejun Zhang. 2025. Songglm: lyric-to-melody generation with 2d alignment encoding and multi-task pre-training. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI’25/IAAI’25/EAAI’25. AAAI Press.
- Zhiwei Yu, Jiwei Tan, and Xiaojun Wan. 2018. A neural approach to pun generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1650–1660, Melbourne, Australia. Association for Computational Linguistics.
- Zhiwei Yu, Hongyu Zang, and Xiaojun Wan. 2020. Homophonic pun generation with lexically constrained rewriting. In *Proceedings of the 2020 Conference on*

BIBLIOGRAPHY

- Empirical Methods in Natural Language Processing (EMNLP)*, pages 2870–2876, Online. Association for Computational Linguistics.
- Chunhong Yuan, Jingyi Tang, Yi Ding Cao, Tianshi Wei, and Wei Tao Shen. 2024a. The impact of artificial intelligence on economic development: A systematic review: The impact of artificial intelligence on economic development. *International Theory and Practice in Humanities and Social Sciences*, 1(1):130–143.
- Ruibin Yuan, Hanfeng Lin, Yi Wang, Zeyue Tian, Shangda Wu, Tianhao Shen, Ge Zhang, Yuhang Wu, Cong Liu, Ziya Zhou, Liumeng Xue, Ziyang Ma, Qin Liu, Tianyu Zheng, Yizhi Li, Yinghao Ma, Yiming Liang, Xiaowei Chi, Ruibo Liu, Zili Wang, Chenghua Lin, Qifeng Liu, Tao Jiang, Wenhao Huang, Wenhui Chen, Jie Fu, Emmanouil Benetos, Gus Xia, Roger Dannenberg, Wei Xue, Shiyin Kang, and Yike Guo. 2024b. ChatMusician: Understanding and generating music intrinsically with LLM. In *Findings of the Association for Computational Linguistics*, pages 6252–6271.
- Francisco Yus. 2013. *An inference-centered analysis of jokes: The Intersecting Circles Model of humorous communication*, pages 59–82.
- Mike Zajko. 2021. Conservative AI and social inequality: conceptualizing alternatives to bias through social theory. *AI & SOCIETY*, 36(3):1047–1056.
- Alessandro Zangari, Matteo Marcuzzo, Andrea Albarelli, Mohammad Taher Pilehvar, and Jose Camacho-Collados. 2025. Pun unintended: LLMs and the illusion of humor understanding. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27936–27971, Suzhou, China. Association for Computational Linguistics.
- Qingcheng Zeng and An-Ran Li. 2022. A survey in automatic irony processing: Linguistic, cognitive, and multi-X perspectives. In *Proceedings of the 29th International*

BIBLIOGRAPHY

- Conference on Computational Linguistics*, pages 824–836, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Guangyan Zhang, Kaitao Song, Xu Tan, Daxin Tan, Yuzi Yan, Yanqing Liu, Gang Wang, Wei Zhou, Tao Qin, Tan Lee, and Sheng Zhao. 2022. Mixed-Phoneme BERT: Improving BERT with Mixed Phoneme and Sup-Phoneme Representations for Text to Speech. In *Interspeech 2022*, pages 456–460.
- Le Zhang, Rongsheng Zhang, Xiaoxi Mao, and Yongzhu Chang. 2022a. QiuNiu: A Chinese lyrics generation system with passage-level input. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 76–82.
- Rongsheng Zhang, Xiaoxi Mao, Le Li, Lin Jiang, Lin Chen, Zhiwei Hu, Yadong Xi, Changjie Fan, and Minlie Huang. 2020a. Youling: an AI-assisted lyrics creation system. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 85–91, Online. Association for Computational Linguistics.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.
- Ying Zhang, Hidetaka Kamigaito, Tatsuya Aoki, Hiroya Takamura, and Manabu Okumura. 2021. Generic mechanism for reducing repetitions in encoder-decoder models. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1606–1615.
- Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020c. DIALOGPT : Large-scale generative

BIBLIOGRAPHY

- pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.
- Yue Zhang, Hongliang Fei, Dingcheng Li, and Ping Li. 2022b. PromptGen: Automatically generate prompts using generative models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 30–37, Seattle, United States. Association for Computational Linguistics.
- Zhehao Zhang, Weicheng Ma, and Soroush Vosoughi. 2024. Is GPT-4V (ision) all you need for automating academic data visualization? exploring vision-language models’ capability in reproducing academic charts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8271–8288, Miami, Florida, USA. Association for Computational Linguistics.
- Kun Zhao, Bohao Yang, Chenghua Lin, Wenge Rong, Aline Villavicencio, and Xiaohui Cui. 2023a. Evaluating open-domain dialogues in latent space with next sentence prediction and mutual information. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 562–574.
- Wenye Zhao, Qingbao Huang, Dongsheng Xu, and Peizhi Zhao. 2023b. Multi-modal sarcasm generation: Dataset and solution. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5601–5613, Toronto, Canada. Association for Computational Linguistics.
- Alex Zhavoronkov, Yan A Ivanenkov, Alex Aliper, Mark S Veselov, Vladimir A Aladinskiy, Anastasiya V Aladinskaya, Victor A Terentiev, Daniil A Polykovskiy, Maksim D Kuznetsov, Arip Asadulaev, Yury Volkov, Artem Zholus, Rim R Shayakhmetov, Alexander Zhebrak, Lidiya I Minaeva, Bogdan A Zagribelnyy, Lennart H Lee, Richard

BIBLIOGRAPHY

- Soll, David Madge, Li Xing, Tao Guo, and Alán Aspuru-Guzik. 2019. Deep learning enables rapid identification of potent DDR1kinase inhibitors. *Nature Biotechnology*, 37(9):1038–1040.
- Mingyu Zheng, Xinwei Feng, Qingyi Si, Qiaoqiao She, Zheng Lin, Wenbin Jiang, and Weiping Wang. 2024. Multimodal table understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9102–9124, Bangkok, Thailand. Association for Computational Linguistics.
- Tan Zhi-Xuan, Micah Carroll, Matija Franklin, and Hal Ashton. 2025. Beyond preferences in ai alignment. *Philosophical Studies*, 182(7):1813–1863.
- Mingyang Zhou, Yi Fung, Long Chen, Christopher Thomas, Heng Ji, and Shih-Fu Chang. 2023a. Enhanced chart understanding via visual language pre-training on plot table pairs. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1314–1326, Toronto, Canada. Association for Computational Linguistics.
- Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023b. Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations*.
- Jian Zhu, Cong Zhang, and David Jurgens. 2022. Byt5 model for massively multilingual grapheme-to-phoneme conversion. In *Interspeech 2022*, pages 446–450.
- Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He,

BIBLIOGRAPHY

- Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2025a. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models.
- Sujia Zhu, Guodao Sun, Qi Jiang, Meng Zha, and Ronghua Liang. 2020. A survey on automatic infographics and visualization recommendations. *Visual Informatics*, 4(3):24–40.
- Yingjie Zhu, Xuefeng Bai, Kehai Chen, Yang Xiang, Jun Yu, and Min Zhang. 2025b. Benchmarking and improving large vision-language models for fundamental visual graph understanding and reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30678–30701, Vienna, Austria. Association for Computational Linguistics.
- Yueyue Zhu, Jared Baca, Banafsheh Rekabdar, and Reza Rawassizadeh. 2023. A survey of ai music generation tools and models.
- Le Zhuo, Ruibin Yuan, Jiahao Pan, Yinghao Ma, Yizhi Li, Ge Zhang, Si Liu, Roger Dannenberg, Jie Fu, Chenghua Lin, Emmanouil Benetos, Wang Xue, and Yike Guo. 2023. Lyricwhiz: Robust multilingual lyrics transcription by whispering to chatgpt. International Society for Music Information Retrieval Conference (ISMIR).
- Vilém Zouhar, Calvin Chang, Chenxuan Cui, Nate B. Carlson, Nathaniel Romney Robinson, Mrinmaya Sachan, and David R. Mortensen. 2024. PWESuite: Phonetic word embeddings and tasks they facilitate. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13344–13355, Torino, Italia. ELRA and ICCL.

Error 404: Page Not Found

The requested resource (more thesis) was not found.

$/p^{\text{heid}_3}/ \rightarrow [\emptyset]$

$\backslash_{-}(\text{o O})_{-}/$