Processing Semantic Outliers and Low Frequency Phenomena with Large Language Models

Dylan Phelps

Supervisor: Dr. Maria-Cruz Villa-Uriol

Supervisor: Prof. Aline Villavicencio

A report submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

in the

School of Computer Science

Declaration

I, Dylan Phelps, 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 through acknowledgement of collaboration, the work presented is entirely my own.

Abstract

Distributional semantics is fundamental to modern language models and is the basis for learning linguistic vector representations. This leads to errors when encountering phenomena that deviate from standard statistical patterns. This thesis examines two domains presenting semantic outliers: idiomatic expressions and semantic changes in Alzheimer's Disease (AD) speech.

The first part of this work tests two hypotheses around idiomatic expressions: whether high-quality embeddings can be trained using low-resource techniques, and whether modern LLMs have out-of-the-box representations that allow them to surpass smaller fine-tuned models. The latter part explores whether semantic content alone gives sufficient diagnostic signal for AD detection when isolated from surface features.

Idiomatic expressions are non-compositional multiword expressions like 'break a leg', where meaning cannot be derived from component words. We apply existing low resource techniques to idiomaticity tasks. Expression embeddings trained using 1-150 contexts and knowledge injection through Pattern Exploit Training both lead to improvements in English idiomaticity detection, though show weaker results in Portuguese and Galician. Multi-billion parameter LLMs are evaluated on multiple idiomaticity detection datasets, with the best models achieving high F1 scores (> 0.86), and analyses revealing strong idiomatic understanding.

AD cognitive impairment manifests as semantic deterioration, including comprehension deficits, semantic paraphasias, and increased generic term usage. We develop an LLM-based pipeline that transforms speech across multiple languages by translating transcripts into English, generating summaries, and creating narrative storyboards. Validation using BLEU, chrF, and semantic similarity shows low surface-form overlap but high semantic preservation. Classifiers trained on transformed transcripts show minimal performance changes (± 0.1 macro F1), confirming that semantic changes alone provide sufficient diagnostic information.

These findings demonstrate that distributional semantic outliers can both challenge LLMs with idiomaticity, whilst presenting opportunities for AD detection.

Acknowledgments

Firstly, my thanks go out to my supervisors, Maria-Cruz and Aline, for their guidance and support throughout the last 4 years. Despite a number of challenges and changes to the scope of my PhD, their wisdom and depth of knowledge have got me back on track, and created new paths when old ones vanished.

Alongside their support, I have been fortunate enough to work with a number of other collaborators here at Sheffield and further afield, primarily through Aline's MWE research group. Ed, Harish, Tom, Maggie, Emily, Rodrigo, Wei, and Marco have all provided interesting discussion, collaboration and inspiration for all the works in this thesis and beyond. Additionally, members of the NLP group have provided numerous insightful discussions through the reading group.

Outside of academia, I'm so grateful to have had so many friends who have made the past 4 years so enjoyable. My flatmates at 12C: Harry, George, Will, Alice, Beth, and Chris, who have brought so much joy to my life, mostly by forcing me to actually leave my room through board-games, karaoke or just chats in the kitchen. I also need to thank my 'Cambridge Friends': Katie, Matt, Henry, Livvy, Claire, Ben, Tom, and Tom. Despite not seeing them as often as I'd like, I'm always excited to reunite with them; they all remain some of my best friends.

I think it would be remiss of me to mention the only part of my life that may have taken up more time than my PhD over the last 4 years: Sheffield University Performing Arts Society (SUPAS). The skills, experiences, and most importantly friends I have made through SUPAS will stay with me for life. I struggle to comprehend the sheer volume of gratitude I have our little group of techies: Sian, Charlie, Maddy, Em, Bethany and Ollie.

Finally, my Mum and Dad who have unconditionally supported me through everything I've done in my life, nurtured my love of learning, and who get most of the credit for the person I am today.

This work was supported by the Healthy Lifespan Institute (HELSI) at The University of Sheffield and is funded by the Engineering and Physical Sciences Research Council [grant number EP/T517835/1]. A particular thank you for the administrative support provided throughout the process by Gemma at HELSI.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

As language models gain widespread adoption beyond academic Natural Language Processing (NLP) circles, their success stems from remarkable advances in model capabilities and downstream utility. Yet despite these advances, the models used widely today remain fundamentally grounded in the distributional semantics hypothesis.

Distributional semantics represents a fundamental approach to computational models of word meaning, built upon the distributional hypothesis first articulated by Harris, 1954 and later popularised by Firth, 1957's famous axiom that "you shall know a word by the company it keeps." This theoretical framework proposes that the semantic content of lexical units can be derived from their patterns of co-occurrence with other units within a large enough corpus, i.e. words that appear in similar contexts tend to have similar meanings.

This framework has proven highly successful as a theoretical foundation for representing meaning within computational models of language. Distributional Semantic Models (DSMs) which use neural networks to model the statistical distribution and produce vector representations of word meaning have come to dominate the field of NLP. Over the last decade, DSMs using progressively more complex architectures (Mikolov et al., 2013; Devlin et al., 2019; Radford et al., 2019), ultimately producing Large Language Models (LLMs) that achieve impressive results across numerous tasks such as translation (Dabre et al., 2020; Zhu et al., 2024), summarisation (Liu and Lapata, 2019; Zhang et al., 2025), and general language understanding (Wang et al., 2019; Qiu et al., 2020; Wang et al., 2024). LLMs have also been applied across domains such as mathematical reasoning (Ahn et al., 2024), law (Lai et al., 2024), and clinical/biomedical applications (Spasic and Nenadic, 2020; Lee et al., 2020; Nori et al., 2023).

However, the distributional hypothesis has important limitations when confronted with out-of-distribution items (Lenci and Sahlgren, 2023). This could be due to phenomena that are under-represented in the data, such as unseen words and patterns; or cases where modelling assumptions do not hold. Since distributional models rely on observing linguistic items across different contexts to infer meaning, they struggle when certain contexts are absent from training data.

1.1 Research Directions

This thesis examines two of these cases, idiomatic expressions and semantic deterioration in the speech of patients with Alzheimer's Disease (AD). This thesis proposes that idiomaticity detection and dementia-related language changes share a fundamental characteristic: both represent departures from standard distributional semantic expectations.

Idiomatic expressions are a class of multiword expressions (MWEs) that derive their meaning not from the composition of their component's meanings, but from conventional usage patterns and are instead learnt as distinct lexical units (Nunberg et al., 1994). For example, idiomatic expressions like 'kick the bucket' (meaning 'to die') cannot be understood through their component words, as the literal meanings of 'kick,' 'the,' and 'bucket' provide no indication of the figurative meaning. Idiomatic expressions are individually rare (Moon, 1998), which means they exert insufficient influence on their component words' distributional representations. DSMs therefore struggle to detect when idiomaticity is being used and as a consequence struggle to create high quality representations capturing the semantic nuance (Sag et al., 2002; Shwartz and Dagan, 2019).

Similarly, the speech of individuals with dementia often exhibits semantic disruptions that deviate from typical distributional patterns. These disruptions manifest as semantic paraphasias (substituting related but incorrect words, such as "fork" for "spoon"), increased use of generic terms such as substituting nouns for pronouns, and the use of empty or vague terms that lack specific semantic content. Such deviations challenge distributional models because the altered word usage patterns no longer reflect conventional semantic relationships that these models rely upon to infer meaning.

Departing from the statistical distribution has different implications for the two cases explored in this thesis. For idiomatic expressions, the failure to accurately represent intended meaning can impair model performance in downstream tasks. Language models have been repeatedly shown to perform worse in sentiment analysis (Williams et al., 2015), machine translation (Dankers et al., 2022), and reasoning (Chakrabarty et al., 2022a) in the presence of idioms. Conversely, for dementia detection, these distributional outliers present a linguistic marker of cognitive decline (Forbes-McKay and Venneri, 2005; Petti et al., 2020). Being able to isolate and specifically identify the semantic changes will improve the ability of language models to serve as automated tools for assisting in AD detection.

This research exists at the intersection of computational linguistics, cognitive science, and clinical applications, and evaluates how large language models represent semantic outliers, and proposing novel methodologies to improve their detection and processing. By examining these models' handling of non-literal language and subtle cognitive changes, this work delivers both theoretical insights into semantic representation and practical frameworks for idiom processing and early dementia detection.

Figure 1.1 illustrates how the two seemingly disparate research domains of idiomaticity and semantic changes in AD are unified by their shared characteristic as distributional semantic outliers, and how this thesis addresses each through complementary research questions.

1.2 Research Aims and Objectives

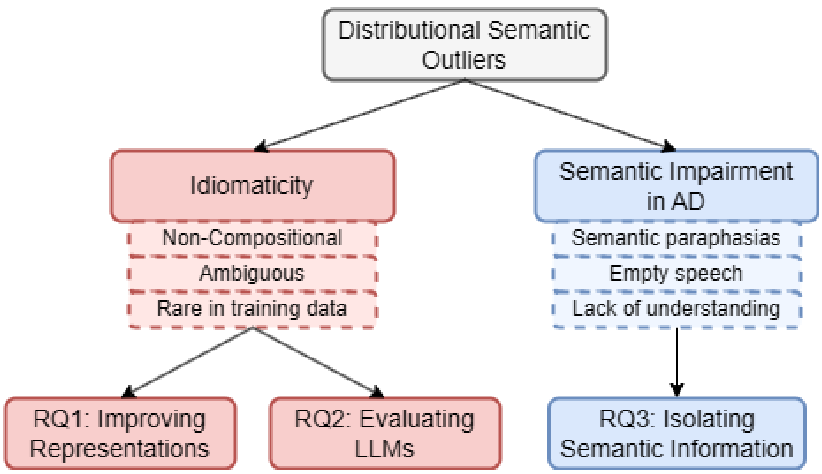


Figure 1.1: Research Questions

This thesis focusses on two types of distributional outliers: idiomaticity through idiomaticity representation and detection tasks, and automatic detection of cognitive decline from the transcribed speech of AD patients. As such, the research questions are split between these two types, with research questions 1 and 2 focusing on idiomaticity and research question 3 focusing on AD. Over the course of the thesis work, 5 publications were produced to answer the research questions.

RQ1: How can low resource techniques be used to enhance representations of idiomatic expressions, and can better representations be used to improve idiomaticity detection? This research question seeks to explore how the performance of pre-trained language models on idiomaticity tasks can be improved whilst only using a small amount of easily accessible (i.e. not manually selected) additional data. Given that individual idiomatic expressions occur infrequently in corpora it may not always be possible to assemble large amounts of examples for each expression, meaning low resource techniques are necessary.

Two promising low-resource approaches for addressing this challenge are single-token embedding training and pattern-based learning. This thesis hypothesises that training a single token to represent the entire MWE is a good candidate for low resource techniques, as this has been shown to work for rare words (Schick and Schütze, 2020). Additionally, Pattern Exploit Training (PET; Schick and Schütze (2021a)), which reformulates classification tasks as cloze-style questions, has demonstrated effectiveness in few-shot settings by leveraging the inherent knowledge of pre-trained language models. Both techniques can produce high-quality representations which may enhance performance on idiomaticity detection tasks.

This hypothesis is tested by training and evaluating new single-token embeddings for each MWE in the SemEval-2022 Task 2 tasks (Tayyar Madabushi et al., 2022) for idiom representation and detection, alongside evaluating PET's pattern-based approach.

Addressing RQ1, Publication I uses BERTRAM to train single-token embeddings from between 1-150 randomly chosen examples of each MWE. Evaluation on SemEval-2022 task 2b demonstrates state-of-the-art (SOTA) performance, with our approach ranking 1st and 2nd place on the task's shared leaderboard.

Publication II evaluates both techniques on the idiomaticity detection through SemEval-2022 task 2a. The results show that both single-token embeddings and PET provide performance improvements on English language data, whilst having little effect on Portuguese and Galician.

Overall, the results from both publications confirm the hypothesis that high quality representations of idiomatic expressions can be trained from small amounts of data. Further, these low resource techniques can improve downstream tasks such as idiomaticity detection.

RQ2: Do multi-billion parameter generative models understand idiomaticity out-of-the-box, and do they compare to smaller model fine-tuned models on idiomaticity detection? This research investigates whether multi-billion parameter generative models can accurately represent non-compositional language without any specific fine-tuning. The hypothesis is that these massive models, with their substantial increase in scale and training data can outperform smaller models on established benchmarks for understanding idiomatic language.

This hypothesis is examined by evaluating a range of multi-billion parameter models without task-specific training on established idiomaticity detection datasets, using both zero-shot and few-shot prompting techniques. Furthermore, by examining the outputs of modern reasoning models, which produce a 'chain-of-thought' (CoT) before a final response, it is possible to evaluate the additional knowledge which the model may use.

Addressing RQ2, Publication III demonstrates that large language models achieve substantial performance on idiomaticity detection tasks without any task-specific fine-tuning. However, comparison with smaller fine-tuned models reveals that zero-shot performance, while impressive, does not quite match the peak performance achieved by specialized fine-tuned approaches.

Publication IV extends this analysis by incorporating models specifically designed for CoT reasoning. This evaluation examines whether explicit reasoning processes yield superior performance and provides insights into how these models internally represent and process idiomatic expressions through analysis of their reasoning outputs. The results show that on multiple datasets larger models with more than 32B parameters begin to reach the SOTA results achieved by fine-tuned models, with manual analysis of the CoT outputs finding a high level of knowledge and reasoning ability

Collectively, the findings suggest that modern LLMs are able to represent idiomaticity well in their internal weights, and thus, achieve strong performance on idiomaticity detection.

RQ3: Can language models detect AD from speech when relying solely on semantic information rather than surface-level linguistic features? The final research question investigates methods to isolate semantic content from spontaneous speech transcripts and automatically detect AD using just this information.

This research question addresses a fundamental issue in automatic approaches to AD detection: whether the diagnostic signal stems from intrinsic semantic degradation, or from surface-level linguistic changes that may be more easily influenced by factors such as education, socioeconomic status, or individual speaking styles. Understanding this distinction will increase our theoretical understanding of how dementia affects language processing, and allow for the development of more robust automatic diagnostic tools.

We hypothesize that semantic content alone contains sufficient information for AD detection, and that LLMs present a unique opportunity to isolate the semantic information, whilst neutralising the diagnostic effects of surface forms.

Addressing RQ3, Publication V introduces a novel pipeline that uses LLMs to transform speech transcripts, altering surface-level features while preserving semantic content.

The effectiveness of the transformations is validated by using similarity metrics to show low surface-form overlap, but high semantic preservation. The transformations are further validated by showing that BERT models trained using the transformed data perform similarly to those trained on the original. These findings provide evidence that semantic degradation can be used as a strong diagnostic signal in systems for automatically detecting AD from spontaneous speech.

The results suggest that fundamental semantic changes can be observed in Alzheimer's disease, with important implications for developing more equitable diagnostic tools that are less dependent on sociolinguistic factors and more focused on the core cognitive changes associated with dementia.

1.3 Thesis Overview

This thesis follows a thesis-by-publication format, with the main body comprised of published papers and manuscripts submitted for review. The following section overviews the chapters of the thesis and provides publication details for the content chapters.

Chapter 2 provides an overview of the relevant background not included in the publications themselves. This first includes an overview of distributional semantic models, before focusing on the limitations and previous work applying those models to idiomaticity and AD detection from speech.

Chapter 3: Learning idiom representations using BERTRAM (Phelps, 2022) was a submission to the SemEval-2022 Task 2 shared task, with this work focusing on subtask b, idiomaticity representation. Evaluated on the final test set this system placed 1st and 2nd in subtask 2a and 2b respectively. This paper was published in the *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*

I am the sole author of this paper, as such all aspects of the paper were completed by myself.

Chapter 4: Sample Efficient Approaches for Idiomaticity Detection (Phelps et al., 2022) was published in the *Proceedings of the 18th Workshop on Multiword Expressions @LREC2022*.

As the lead author, I completed the majority of the experiments and analyses presented, with some additional work on PET being performed by Xuan-Rui Fan. The paper was conceived by myself and Harish Tayyar-Madabushi, with writing support from the other authors.

Chapter 5: Sign of the Times: Evaluating the use of Large Language Models for Idiomaticity Detection (Phelps et al., 2024) was published in the *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024* and was the **winner of the workshop's Best Paper Award**.

This was a collaborative work in which I was the lead author, presenting the initial idea, and leading on analysis and paper writing with assistance from the other authors. The working running the experiments was split between myself, Thomas Pickard, Edward Gow-Smith and Maggie Mi.

Chapter 6: Stands to Reason: Investigating the Effect of Reasoning on Idiomaticity Detection (Phelps et al., 2025) will be submitted for review between thesis submission and the viva.

Again, this was a collaborative work, where I presented the idea, performed the experiments, organised and took part in the manual labelling, and lead on the analysis and writing up. Alongside assisting with analysis and writing, Rodrigo Wilkens performed much of the word count analysis, and Edward Gow-Smith and Thomas Pickard also took part in manual labelling.

The version of the paper presented in the thesis differs significantly from the version submitted for review. A section using linguistic metrics to analyse reasoning output was removed, as the work was completed almost entirely by Maggie Mi. A new section on adding definitions to prompts has been added in its place, of which I am the sole author.

Chapter 7: Beyond surface form: A pipeline for semantic analysis in Alzheimer's Disease detection from spontaneous speech has been submitted for review at the *PLoS One Journal.*

The idea for this paper was conceived by myself, Rodrigo Wilkens and Aline Villavicencio, with Rodrigo Wilkens assisting with the initial data preparation and exploratory experiments. The experiments, results, and analysis presented in the paper were all completed and written up by myself. Lilian Hubner and Barbara Malcorra provided domain knowledge of AD. All other authors provided feedback throughout the process and on various drafts of the paper.

Chapter 8 concludes the thesis by detailing how the work presented has addressed the research questions and giving suggested directions for future work.

Chapter 2

Background

2.1 Distributional Semantics

Modern language models and much of modern NLP is built upon the distributional hypothesis, ‘words that appear in the same contexts tend to have similar meanings’ (Harris, 1954). This hypothesis has been operationalised in NLP by representing words as vectors in a high-dimensional space, with the geometric relationships between these vectors capturing the semantic relationships between the corresponding words. Words that frequently appear in similar contexts are positioned closer together in this semantic space, reflecting their semantic similarity.

Generally, language models built upon the distributional hypothesis, or DSMs, take a large corpus of text as input and generate a numeric representation for each word in the input vocabulary using a statistical method (Lenci, 2018). This results in vectors, or embeddings, for either words or documents, which represent meaning in a mathematical format. This enables the use of geometric methods to calculate similarities between words or documents, providing a quantitative measure of how closely related their meanings are. The most common similarity metric for measuring the distance between vectors in a semantic vector space is cosine similarity (Bullinaria and Levy, 2007).

Lenci and Sahlgren (2023) describe 3 generations of DSMs: *count* DSMs which create embeddings through counting co-occurrences of words in a corpus, *predict* DSMs which use neural networks trained to predict words, and *contextual* DSMs which use deeper neural networks to generate different embeddings for each occurrence of a word changing based on the words in the context, rather than the single embedding for each word generated with count and predict DSMs.

Early success with DSMs were primarily *count* DSMs, which created vector representations by performing simple statistical operations over matrix representations of corpora. Latent semantic analysis (Deerwester et al., 1990; Landauer and Dumais, 1997) performs rank-lowering through singular-value decomposition on a term-document matrix for a corpus to produce vector representations of documents in the corpus. The rank reduction causes related terms to be mapped together even if they never co-occurred, revealing latent semantic structure. Another early method, Hyperspace Analogue for Language (HAL; Lund and Burgess (1996)) produced word vectors for a corpus by counting co-occurrence of words within a sliding window.

Alongside these statistical methods, *predict* models using neural networks were being developed by several researchers. Initial neural network models were trained to predict the next word in the sequence, and usually contained an embedding layer, whose weights composed the representations. Early work (Elman, 1990) used recurrent neural networks, whereas more prominent later work used feed-forward neural networks (Bengio et al., 2003; Mnih and Hinton, 2008). Alternatively, some work focused on the use of discriminative networks, where the training task was to assign higher scores to n-grams occurring in a corpus than randomly generated n-grams (Collobert and Weston, 2008).

Word2Vec Mikolov et al. (2013) introduced two new neural network DSMs, Skip-Gram and Continuous Bag-of-Words (CBOW), which became well known and widely used in both the NLP community and more generally. Both DSMs were similar to previous feed-forward networks (Bengio et al., 2003), but without a non-linear hidden layer and with both an input and output embedding layer. Additionally, previous works were intended to predict next words in a sequence, with representations being a byproduct of this process, word2vec, instead targeted high quality representations.

The two variants differ in training objective, for CBOW the model predicts a target word given the surrounding context, whereas Skip-Gram predicts the surrounding words from a given target.

Word2Vec improved over previous models in various ways: hierarchical softmax, representing frequent words as shorter binary codes, subsampling frequent words, and phrase detection, which allowed common n-grams to be represented as single tokens. Additionally, previous *predict* (and *count*) DSMs had been limited on corpus size by the computation needed to train the model. As well as the above optimisations, word2vec introduced negative sampling, which alters the training objective to be discriminative with only a small number of negative examples. Negative sampling allowed linear scaling with respect to corpus size, as a larger corpus no longer meant a larger vocabulary to compare each example to.

Further to the efficiency improvements, Mikolov et al. (2013) also showed that the representations generated by word2vec displayed compositional properties, allowing geometric operations to be performed on word embeddings. The most famous example of this is a case of the parallelogram model of analogy (Rumelhart and Abrahamson, 1973; Peterson et al., 2020), 'king' - 'man' + 'woman' = 'queen'.

Multiple variations on word2vec have been proposed in subsequent years (Melamud et al., 2016; Li et al., 2017). FastText Bojanowski et al. (2017) uses character n-grams as the base tokens of the model rather than words in an effort to represent rare words that are not in the vocabulary of the original corpora. Alternatively, Global Vectors (GloVe) (Pennington et al., 2014) uses a *count* based methodology to more directly capture the semantic relationships, by training the model to produce word vectors where the dot product of any two word vectors relates to their co-occurrence probability in a corpus.

A major limitation of all of the above DSMs is that they only produce one embedding per word in the vocabulary. This is suboptimal for polysemous words, as all the senses of the word must be captured in a single vector representation. This has a particularly large impact on words with one dominant sense, which is highly represented in the embedding, whilst other, rarer, senses are dominated.

Therefore, the most recent generation of DSMs, *contextual* DSMs, instead produce a new context-specific embedding for each word token in a given sentence. These models typically use a multilayered neural network with fixed input embeddings that are contextualised as they pass through the network. Embeddings from Language Models (ELMo) (Peters et al., 2018) used bidirectional Long Short-Term Memory (LSTM) layers to generate context-sensitive representations, demonstrating that the same word could have different embeddings depending on its linguistic environment. Being bidirectional also means that embeddings produced by ELMo can incorporate context from both sides of the word. This helped to establish the principle that meaning is fundamentally contextual rather than static.

2.1.1 Large Language Models

The Transformer (Vaswani et al., 2017) revolutionised the field by using the attention mechanism (Bahdanau et al., 2015) as the primary computational component, moving away from the recurrent and feed-forward networks used to this point. This architecture enables contextual DSMs to capture long-range dependencies and generate dynamic word representations that vary based on surrounding context, rather than the static embeddings produced by earlier *predict* models. The self-attention mechanism allows these models to weigh the relevance of different words in a sequence when computing representations, leading to more nuanced contextualised semantic representation that helped to address the problems of polysemy.

The transformer architecture (shown in Figure 2.1) is made up of two stacks, each consisting of n identical layers. The first stack, the encoder, is used to map the input embeddings to an internal contextualised representation, whilst the second stack, the decoder, has a separate input embedding layer, but also takes the contextualised embeddings from the encoder to output a probability distribution over the vocabulary to predict the next token.

Since its introduction, the transformer has become the de-facto architecture for SOTA language models, with variants using just the encoder stack to create embeddings (Devlin et al., 2019), and either just the decoder stack (Radford et al., 2018; Touvron et al., 2023) or both stacks (Raffel et al., 2020; Lewis et al., 2020) to create generative next-word prediction models. The architecture has been shown to scale well to billions of parameters and trillions of tokens of training data, leading to what are referred to as Large Language Models (LLMs).

2.1.2 Encoder Models

Bidirectional Encoder Representations from Transformers (BERT, Devlin et al. (2019)) is one popular encoder-only implementation of the transformer architecture, specifically trained to make high quality contextualised embeddings that can be used downstream in a variety of tasks (Rogers et al., 2020). The inputs to the model are sub-word tokens from a word-piece tokeniser (Wu et al., 2016). Tokenisation breaks text into processable units, and sub-word tokenization handles out-of-vocabulary words by decomposing them into known components while maintaining a compact vocabulary (Sennrich et al., 2016). This allows BERT to tokenize the entire English language with

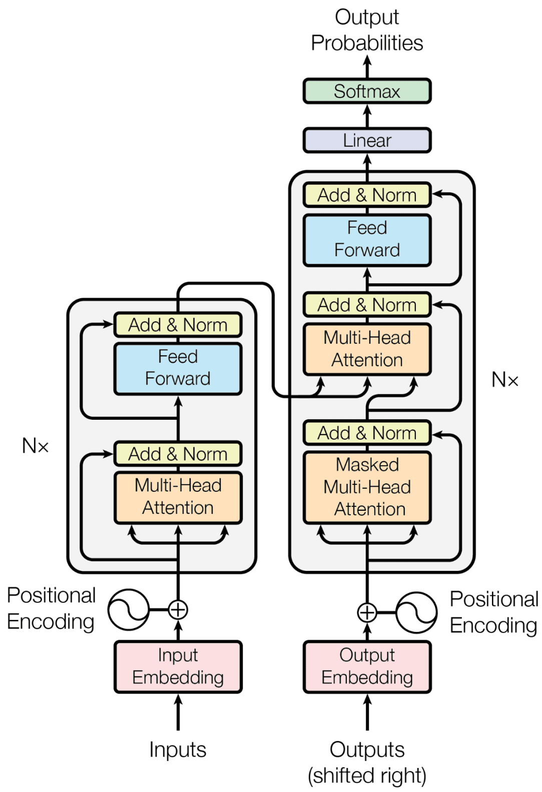


Figure 2.1: The encoder and decoder stacks of the transformer, made up from a number of layers (a single layer shown in the grey box for each stack). Figure taken from Vaswani et al. (2017).

a vocabulary size of 30,000, ensuring computational efficiency while avoiding unknown token issues.

BERT is trained using a self-supervised regime on two tasks. Firstly, masked language modelling (MLM), where 15% of the input is replaced with '[MASK]' tokens which the model attempts to predict. Secondly, next sentence prediction (NSP) where two sentences are passed to the model which should predict whether the second sentence follows the first sentence or whether it was another randomly chosen sentence. BERT is then trained on large amounts of unlabelled general text data taken from the internet (pre-training) and can be further trained on specific tasks or domains to achieve higher performance on those downstream tasks and domains (fine-tuning).

BERT has been incredibly successful at many NLP tasks, and has become the de-facto SOTA embedding model, with many works exploring its high performance (Rogers et al., 2020). Because of this success, many variants of BERT have been proposed, which reduce the size of the model, tweak the pre-training tasks, or improve performance on certain subsets of tasks. RoBERTa (Liu et al., 2019) is a variant of BERT, which retains the architecture and the model size, but is trained on 10x more data (160GB vs 16GB), and drops the NSP pretraining task. RoBERTa has been shown to achieve 2-20% more performance than BERT on a variety of downstream tasks. DistilBERT (Sanh et al., 2020) is a distilled version of BERT, with only half the number of parameters. This smaller model size means it is easier to fine-tune and deploy, whilst retaining about 97% of BERT's performance across a range of tasks.

Another variation with a significantly different training procedure to BERT is ELECTRA (Clark et al., 2020), which is trained as a discriminator rather than a generator. A smaller language model performs the MLM task on a sentence, the main ELECTRA model must then predict which words were originally in the sentence, and which words were masked and subsequently replaced. This task is show to be more efficient than MLM, as ELECTRA must discriminate every token in the input, rather than predicting just 15% of tokens.

2.1.3 Generative Models

Most recently, decoder only generative models have become the most popular form of DSMs. However, instead of explicitly producing embeddings, these models have implicit embeddings as internal representations and are instead trained in the task of next word prediction, just like some of the earliest DSMs (Elman, 1990; Bengio et al., 2003).

Generative Pre-trained Transformer (GPT; Radford et al. (2018)), was the first popular implementation of this generation of generative LLMs. GPT is an autoregressive language model that processes text unidirectionally, using only preceding context to predict the next token in a sequence. During generation, the model produces one token at a time, with each newly generated token being appended to the input context for subsequent predictions, enabling coherent long-form text generation.

GPT-2 (Radford et al., 2019) showed that by scaling the number of parameters, from 117 million to 1.5 billion, and the amount of the training data used, you can achieve linear performance improvement. This finding was further substantiated up to 175 billion parameters by GPT-3 (Brown et al., 2020), which additionally demonstrated that few-shot learning, wherein a small number of task-specific examples are provided in context rather than through additional training, can effectively adapt model behaviour to new domains and tasks. These 'scaling laws' (Kaplan et al., 2020) have become another foundation in modern language modelling, consistently demonstrating that larger models and datasets yield predictable improvements in performance across diverse tasks. This comes at the cost of increased computational requirements (Hoffmann et al., 2022), environmental impact (Strubell et al., 2019; Faiz et al., 2024), and training data demands (Villalobos et al., 2024).

A key feature of these models is their ability to perform in-context learning, where they can acquire new skills and adapt to novel tasks through prompts and examples provided at inference time, without requiring parameter updates or architectural modifications (Liu et al., 2023b; Sahoo et al., 2025). Prompts serve as natural language instructions that guide the model's output generation, enabling users to specify tasks, provide context, set formatting constraints, and even influence the model's reasoning approach through strategic prompt design (Vatsal and Dubey, 2024; Chen et al., 2025). As models scale beyond certain parameter thresholds, it has been argued that they exhibit emergent abilities that are not present in smaller models and have not explicitly been trained for but spontaneously arise at scale (Wei et al., 2022a). However, other work has shown that such emergent abilities may just be explained by a combination of in-context learning and linguistic knowledge (Lu et al., 2024), or a result of metric choice (Schaeffer et al., 2023).

Another recent innovation in prompting methodology has been the utilisation of chain-of-thought (CoT) prompting (Wei et al., 2022b), whereby models are instructed to output their 'thinking' before producing an answer. This technique is thought to work as it allows the model to add information to the context where it can more efficiently used for prediction. Whilst having been shown to excel in mathematical, programming, and commonsense reasoning tasks (Lu et al., 2023; Chu et al., 2024; Li et al., 2025), CoT prompting can also improve performance on more general language tasks (Wang et al., 2024). Models trained specifically to better utilise CoT prompting have subsequently become SOTA general LLMs (DeepSeek-AI et al., 2025; OpenAI, 2025b).

2.2 Idiomaticity

Multiword expressions (MWEs) are collocations of multiple words (Sag et al., 2002) which occur more often than by chance in a given corpus (Carpuat and Diab, 2010). There are multiple different categories of MWEs (Constant et al., 2017) such as idioms (e.g. *pain in the neck*), light verb constructions (e.g. *do a dance*), verb particle constructions (e.g. *get down*), named entities (e.g. *Sheffield University*), domain specific terminology (e.g. *Alzheimer's Dementia*), amongst others.

A key feature of MWEs is that they exist on a spectrum of non-compositionality, or idiomaticity (Sag et al., 2002). Semantic transparency refers to what degree the meaning of an MWE can be discerned from the meaning of its components. With expressions such as *olive oil* being fully transparent, and others such as *red herring* being fully opaque and requiring external knowledge or context to comprehend (Libben, 1998). Theories suggesting that humans process idioms as either a single lexical unit (the Idiom Principle; Sinclair (1991)) or as both component and as a single item concurrently (Titone and Connine, 1999; Caillies and Butcher, 2007).

DSMs are built on the Principal of Compositionality (Partee, 1984; Mitchell and Lapata, 2010), making the assumption that through combining individual word or subword representations the meaning of the whole can be understood. MWEs challenge this assumption, and as such, have been shown to affect the performance of models in downstream tasks (Sag et al., 2002; Acosta et al., 2011; Dankers et al., 2022).

Another complicating aspect of MWEs, and idiomatic expressions specifically, is the scope for ambiguity (Cronk et al., 1993). Most idioms have a literal sense, that, although perhaps uncommon, will inevitably appear within a large enough corpus.

'Breaking the ice' can be used both to describe overcoming awkwardness in a social situation or literally smashing frozen water; disambiguation can be done using the context surrounding the expression. Detecting which sense of these 'potentially idiomatic expressions (PIEs)' (Haagsma et al., 2019) has been used is the task of idiomaticity detection.

This ambiguity and variation in what is defined as an MWE or an idiomatic expression means that comprehensive resources are difficult to compile. Some online resources designed for language learners exist[1], however, their quality is hard to verify. This issue is one we explore further in Publication 4.

2.2.1 Idiomaticity Representation

Static embedding models (word2vec, Mikolov et al. (2013); fasttext, Bojanowski et al. (2017)) are particularly vulnerable to these challenges, as a single embedding for each word much represent all of its senses. For example, given 'red herring' a single vector must capture the meaning of both a primary coloured fish and a misleading clue. Contextual models (BERT, Devlin et al. (2019); ELMo, Peters et al. (2018)) have the ability to instead create different embeddings for each sense depending on the sense.

Effectively evaluating the quality of idiomatic expressions has been found to be quite difficult. Some work has attempted to explicitly measure this by probing the representations, often comparing the embeddings to the embeddings of paraphrases (Garcia et al., 2021b; Klubička et al., 2023). Other work has focused on 'downstream' tasks, where it is expected good representations are needed for good performance.

Static embeddings have shown good performance in predicting the level of compositionality in given expressions (Salehi et al., 2015), often matching the performance of contextual embeddings (Cordeiro et al., 2019). Idiomaticity detection is another downstream task which involves distinguishing between idiomatic and literal uses of expressions, and can be seen as a form of word sense disambiguation (Stevenson and Wilks, 2005). A number of datasets have been produced for this task (Cook et al., 2008; Haagsma et al., 2020; Tayyar Madabushi et al., 2022; Mi et al., 2025). Early contextualised embeddings performed better on these datasets than static embeddings (Hashempour and Villavicencio, 2020), but still claimed overall weak performance (Garcia et al., 2021a). Later work has shown a marked improvement when fine-tuning

[1] www.theidioms.com, for example

models like BERT for idiomaticity detection, with performance approaching saturation (Chu et al., 2022; Bigoulaeva et al., 2022; Zeng and Bhat, 2021).

There remains open questions about how to improve the representations of idiomatic expressions within transformer language models. Some work has experimented with creating single token embeddings (Garcia et al., 2021a), but this is limited to use within compositionality prediction. Other work has looked at other techniques, such as contrastive learning (He et al., 2024). Ultimately, this remains a gap in the literature, which we seek to address in Publications 1 and 2.

With the advances in multi-billion parameter generative models and innovations in prompting techniques, such as chain-of-thought reasoning, it is possible the problems of detecting and representing idiomaticity may have also been solved. Therefore, we also seek to determine whether the advancements in language skills from the latest generation of models also translate to improved idiomatic representations (see Publication 3) and if CoT prompting has any effect (see Publication 4).

2.2.2 Idiomaticity Datasets

Throughout the thesis, we utilise several existing idiomaticity datasets for evaluation of the proposed methodologies. This section provides an overview of the datasets as well as the evaluation task and metrics used with them. An example entry from each dataset is shown in Table 2.1, apart from SemEval-2022 Task 2b, which is shown in Table 3.1.

SemEval-2022 Tasks

Two of the most relevant datasets for this work are from the SemEval-2022 Task 2 shared task (Tayyar Madabushi et al., 2022). The tasks were designed to address the lack of high quality datasets to evaluate language models ability to detect when idiomicity is being used (Task 2a), and the quality of word and phrase representations when idiomaticity is being used (Task 2b). Both tasks are multi-lingual, with English and Portuguese data being included in the training set, additionally including Galician data in the test sets. Our work using the datasets conforms to the evaluation metrics described in the task.

Dataset	Expression	Example	Label
Task 2a	glass ceiling	Vice President Harris, the daughter of immigrants from Jamaica and India, shattered that major glass ceiling and then got right to work alongside President Biden.	Idiomatic
	glass ceiling	But Mr Patterson agrees the standout selling point is undoubtedly the pool and retractable glass ceiling.	Compositional
Flute	old as the hills	She had been old as the hills for as long as I could remember. She had been relatively young for as long as I could remember.	Contadiction
	dead in the water	That opportunity may be dead in the water. That opportunity may be completely over.	Entailment
MAGPIE	above board	Everything was very legitimate and above board, but we had very specific requirements	Idiomatic
	watering hole	There was a fairly decent - sized watering hole right in the centre of the field.	Compositional
DICE	against the grain	Out of duty she had caved in, but it still went against the grain.	Idiomatic
	against the grain	Carpenters recommend not to sand against the grain as it can damage the wood.	Compositional

Table 2.1: One example from each class for each idiomaticity detection dataset used within the thesis. FLUTE shows both the premise and hypothesis sentence for each example.

Task 2a is an idiomaticity detection task involving binary classification, where the model predicts whether a noun compound in a given sentence is used literally or idiomatically. Classification quality is evaluated using macro F1 as the primary metric. The shared task provides two evaluation settings: (1) a 'zero-shot' setting where noun compounds in the training and test sets are disjoint, and (2) a 'one-shot' setting where the training set includes one idiomatic and one literal example for each noun compound that appears in the test set.

Task 2b is a semantic text similarity (STS) task where the model assigns similarity scores (0 to 1) to sentence pairs. To evaluate idiomatic representations, the task uses specially constructed STS data: each sentence containing a noun compound idiom is transformed into two pairs by replacing the idiom with (1) a literal word-for-word paraphrase and (2) an idiomatic paraphrase. For example, 'big picture' becomes 'large image' (literal) and 'whole situation' (idiomatic). The evaluation assumes that the original sentence should receive a high similarity score (≈ 1) when paired with the idiomatic paraphrase, while both the original and the idiomatic paraphrase should receive low scores when compared against the literal paraphrase. Model performance is measured using Spearman rank correlation between predicted and gold standard scores.

Figurative Language Understanding through Textual Explanations (FLUTE)

FLUTE (Chakrabarty et al., 2022b) is a dataset that frames figurative language understanding as a natural language inference (NLI) task, where the model determines whether a hypothesis follows from a given premise. The dataset covers four types of figurative language: sarcasm, simile, metaphor, and idioms. In this work, we focus exclusively on the idiom split. For each example in the idiom set, the hypothesis is a modified version of the premise in which the idiomatic expression is replaced with a paraphrase generated using GPT-3 with human-in-the-loop verification. The model is evaluated using macro F1 over its predictions of 'entailment' and 'contradiction'.

MAGPIE

MAGPIE (Haagsma et al., 2020) is a dataset of 56,622 instances of potentially idiomatic expressions, labelled by crowd-sourced workers as either idiomatic (70%), literal (28%), or other (1%). We use the randomly split test set as a binary classification task where the model predicts whether a given instance is literal or idiomatic, similar to SemEval-2022 Task 2a. The dataset contains only English data, and macro F1 is used as the primary evaluation metric.

DICE

DICE (Mi et al., 2025) introduces a binary classification dataset for idiomaticity detection that addresses issues in previous datasets highlighted by Boisson et al. (2023). Specifically, DICE generates challenging literal examples that force models to attend more closely to context. These examples are generated using GPT-4, with expert annotators selecting high-quality instances. The dataset also introduces a new metric, strict consistency, which defines accuracy at the type level: a model must correctly classify all instances of a given expression to receive credit. Using this data creation process and evaluation metric, the authors demonstrate that modern LLMs achieve substantially lower scores on DICE compared to other idiomaticity detection datasets.

2.3 Dementia and Language

Approximately 982,000 people in the UK have a form of Dementia, of which Alzheimer's Disease (AD) is the most common, a number which is projected to rise to 1.4 million by 2040 (Alzheimer's Society, 2025). This anticipated increase will strain existing diagnostic resources and create demand for efficient screening tools.

Current AD diagnosis involves neuropsychological batteries, involving a range of diagnostic tests, including language-based tasks. One such task is 'picture description', where patients are asked to describe scenes depicted in standardized images (Goodglass and Kaplan, 1983). Clinicians analyse these responses for linguistic markers of cognitive decline before proceeding to additional testing.

Previous work has shown that cognitive decline due to AD can manifest as changes in speech patterns, including reduced semantic fluency, simplified syntax, and altered discourse coherence (Bayles et al., 1987; Lyons et al., 1994; Tomoeda et al., 1996). These changes represent a significant deviation from the typical distributions of language patterns seen in general corpora. Therefore, language models trained on large corpora offer a promising approach for automated detection of early cognitive decline markers (Luz et al., 2020; Balagopalan et al., 2020).

2.3.1 Effect of AD on language

Dementia affects several types of memory, especially episodic and semantic ones. Language can also be affected (Weiner et al., 2008; Ahmed et al., 2013), as besides being a complex symbolic system, it interacts with other cognitive constructs, like memory types, executive functions, and attention. For instance, studies on connected speech have shown that low speech connectedness in AD is associated with poorer semantic memory performance, which, in turn, impacts episodic memory capacity (Malcorra et al., 2021). Psycholinguistic studies have shown that certain psycholinguistic features are sensitive to reduced semantic ability in Alzheimer's, namely word concreteness, age of acquisition, animacy, and frequency, to name some of these features (Räling et al., 2017).

Semantic memory loss is a marker of early Alzheimer's disease-related neurodegeneration in older adults (Vonk et al., 2020). Moreover, there is a correlation between clinical dementia measures such as the Mini-Mental State Exam (MMSE) (Folstein et al., 1975; Tombaugh and McIntyre, 1992) with various linguistic aspects(Fraser et al., 2016; Kavé and Dassa, 2018; Malcorra et al., 2024). For instance, the study developed by Kavé and Dassa (2018) showed a correlation between MMSE scores with lexical features and information units in a narrative production based on pictures in their AD group but not in the control group.

In addition, AD might also impact syntax. Some studies (Saffran et al., 1989; Thompson et al., 1997) have identified patients who produce agrammatic speech, whilst, conversely, (Kempler et al., 1987) showed that syntactic ability can be preserved. It is widely discussed whether the effects on syntax seen in AD patients are due to memory and semantic impairment (Reilly et al., 2011). It is not an easy task to try to split apart what relates to syntax versus semantics when it comes to sentence and discourse processing (Ehrlich et al., 1997; Sajjadi et al., 2012). Furthermore, cognitive deficits, like those in semantic memory, may interfere with both syntactic and semantic processing at different levels and also depending on the disease stage (Nasiri et al., 2022).

2.3.2 Language model-based AD classification

Language impairment caused by AD leads to deviations from statistical patterns observed within general corpora. For this reason, language models, which represent the distribution of the underlying data they are trained on, can be used to identify these deviations and be used as a diagnostic tool for AD.

Automatic language assessment methods have been refining their ability to assess language for AD, and much of this work has focused on extracting linguistic features, such as lexical frequency and noun-pronoun ratio, and training classifiers based on these. More recently, however, we have seen the rise of transformer (Vaswani et al., 2017) language models such as BERT (Devlin et al., 2019) and GPT (Radford et al., 2019; Brown et al., 2020), in addition to their high level of performance in general tasks, these models have also demonstrated promising results when applied to tasks in the medical domain (Lee et al., 2020).

These models have subsequently been applied to automatic AD detection tasks, with many studies using the Alzheimer's Dementia Recognition through Spontaneous Speech Challenge (ADReSS) shared task dataset (Luz et al., 2020). ADReSS presents a dataset consisting of audio recordings and transcripts of spontaneous speech in response to the Cookie Theft picture description task (Goodglass and Kaplan, 1983). The dataset consists of an equal number of entries from people with AD and people without AD, and is balanced across gender and age for both cohorts, to avoid bias and allowing for easy comparison of classification models using accuracy or F1 score as a metric. Work on this task using fine-tuned language models, specifically BERT, has been shown to produce good performance (Balagopalan et al., 2020; Agbavor and Liang, 2022), generally better than classifiers based on using the audio data (Cummins et al., 2020), and those utilizing linguistic features (Balagopalan et al., 2021).

With the advent of LLMs such as GPT-4 (OpenAI et al., 2024a), some initial work has applied these models, directly out-of-the-box, to AD detection from spontaneous speech tasks, with limited success (B T and Chen, 2024; Yang et al., 2023). Furthermore, some research, which is most similar to the work in Chapter 7, has attempted to extract semantic features from the ADReSS transcripts by prompting GPT-4 to score each transcript on a number of linguistically relevant metrics, such as discourse impairment and semantic paraphasias (Heitz et al., 2025). Overall, they find that the using GPT-4 derived metrics alongside traditional metrics for classification improves performance.

The majority of the work discussing semantics in dementia uses proxy measures such as noun-pronoun ratio or lexical frequency to detect semantic changes. Whilst these measures have been widely and successfully used, they can also be impacted by other factors, such as class, education level, and language fluency. Therefore, there is a disconnect between measuring surface form features and actual comprehension deficits. The problem of isolating the semantic features from more fragile surface form features and variations remains unresolved.

Chapter 3

Publication I: Learning idiom representations using BERTRAM

Dylan Phelps[1,2]

1 Healthy Lifespan Institute, The University of Sheffield, Sheffield, United Kingdom

2 School of Computer Science, The University of Sheffield, Sheffield, United Kingdom

Abstract

This paper describes our system for SemEval-2022 Task 2 Multilingual Idiomaticity Detection and Sentence Embedding sub-task B. We modify a standard BERT sentence transformer by adding embeddings for each idiom, which are created using BERTRAM and a small number of contexts. We show that this technique increases the quality of idiom representations and leads to better performance on the task. We also perform analysis on our final results and show that the quality of the produced idiom embeddings is highly sensitive to the quality of the input contexts.

3.1 Introduction

Idiomatic expressions present a challenge to Large Language Models (LLMs) as their meaning cannot necessarily be derived from the composition of their component tokens, a trait that LLMs often exploit to create representations of multiword expressions. The lack of compositionality leads to poor representations for idiomatic expressions and in turn poor performance in downstream tasks whose data includes them.

SemEval-2022 task 2b (Tayyar Madabushi et al., 2022) encourages the creation of better representations of idiomatic expressions across multiple languages by presenting a **Semantic Text Similarity (STS)** task in which correct STS scores are required whether or not either sentence contains an idiomatic expression. The sub-task requires the creation of a self-consistent model in which a sentence including an idiomatic expression and one containing its literal meaning ('*swan song*' and '*final performance*') are exactly similar to each other and equally similar to any other sentence.

To achieve this goal, we investigate whether due to the similarity between idioms and rare-words Schick and Schütze's BERT for Attentive Mimicking (BERTRAM; Schick and Schütze (2020)) model, which was designed for use with rare-words, can be used to explicitly learn high-quality embeddings for idiomatic expressions. We also investigate how many examples of each idiom are required to create embeddings that perform well on the task, as well as how the quality of contexts fed to the BERTRAM model effects the representations and performance on the task.

Evaluating our model on the task shows that externally trained idiom embeddings significantly increase the performance on STS data containing idioms while maintaining high performance on general STS data. This improved performance gained an overall spearman rank score of 0.6402 and first place (of six entries) on the pre-train setting, and an overall spearman rank score of 0.6504 and second place (of five entries) on the fine-tune setting.[1]

3.2 Background

Adopting the idiom principle (Sinclair, 1991) to produce a single token representation for MWEs has been used widely within static embedding distributional semantic models (Mikolov et al., 2013; Cordeiro et al., 2019). Within contextualised representation models, Hashempour and Villavicencio, 2020 show that the contextualised representations produced by context2vec (Melamud et al., 2016) and BERT (Devlin et al., 2019) models can be used to differentiate between idiomatic and literal uses of MWEs. However, the MWEs are only represented by one token in the input, before being broken into many tokens using BERTs word piece tokenizer. Tayyar Madabushi et al., 2021 add a token to the BERT embedding matrix and shows that this method

[1]The code for creating the embeddings and the modified baseline system code can be found on GitHub: https://github.com/drsphelps/semeval-task-2.

improves representations through increased performance on their proposed STS task. The embeddings they add to BERT are randomly initialised, however, and only trained during the fine-tun step on limited data.

3.2.1 BERTRAM

BERT for Attentive Mimicking (BERTRAM) (Schick and Schütze, 2020), originally developed to improve representations of rare words, builds upon attentive mimicking (Schick and Schütze, 2019) to create embeddings, within existing embedding spaces, for tokens that incorporate both form and context information from a small number of example contexts. During training the model attempt to recreate embeddings for common words with the existing embedding in the model treated as the 'gold embedding', a process known as mimicking. Form embeddings are then learnt using trained n-gram character embeddings, before being passed with a context into a BERT model. The output of the BERT model forms the embedding for that specific context. To incorporate knowledge from many contexts an attention layer is applied over the outputs for each context to get the final embedding. There exist other models to produce effective embeddings from a small number of contexts (Zhao et al., 2018; Pinter et al., 2017), however, BERTRAM is the only model that is non-bag-of-words and incorporates both form and context information when creating the embedding.

Rare words are unsurprisingly defined by how uncommon they are within datasets. This leads to problems when using LLMs on tasks involving rare words as the word pieces they are broken down into have not been influenced enough during pre-training to accurately represent them. Similarly, idiomatic phrases represent a small proportion of the usage of their constituent words, the idioms in the development set for this task represent an average of 4.9% of the usage of their constituent words. Therefore, the embeddings for constituent words are not significantly effected by the usage of idioms in the training data, leading to the model failing to understand the idiomatic expressions. Further similarities between idioms and rare-words include the variance in compositionality, for example, *unicycle* can be partially understood from its word pieces, whereas *kumquat* cannot.

3.3 Methodology

3.3.1 Embedding Creation

Due to the similarities between rare words and idioms, we use BERTRAM to create representations for idiomatic expressions. A separate BERTRAM model is used for each nof the tasks languages. For English, we use the pre-trained model provided with the original paper. For Portuguese and Galician we train BERTRAM models with BERTimbau Base (Souza et al., 2020) and Bertinho-Base (Vilares et al., 2021) respectively used as the base transformers. The Portuguese and Galician BERTRAM models that we train are trained using almost the same training regime outlined for the English model in the original paper, 3 epochs of context only training, 10 epochs of form only training and 3 epochs of combined training. Due to time and compute restrictions, we do not use One-Token Approximation to expand the number of gold standard representations that can be used for attentive mimicking. The Portuguese and Galician splits of the cc100 dataset (Conneau et al., 2020; Wenzek et al., 2020) are used to train the models, with the entire split being used for Galician, and a 10GB subset used for Portuguese.

Contexts for each of the idioms found in the task data can then be created using these models. Examples are retrieved from the relevant split in the cc100 dataset using a grep command [2] that retrieves the entire line that the instance of the idiom is found on. We investigate how changing the number of contexts used to create each embeddings changes our performance on the task by creating embeddings for each idiom with between 1-250 examples in intervals.

3.3.2 Model Architecture

For predicting the similarity scores, a separate model is used for each of the languages BERT-Base (Devlin et al., 2019) for English, BERTimbau for Portuguese, and Bertinho-Base for Galician. The created BERTRAM embeddings for each of the idioms found within the task are added into the embedding matrix of the relevant model. These models are used within a Sentence BERT (Reimers and Gurevych, 2019) setup, implemented using the SentenceTransformers library, which consists of a siamese net-

[2] *grep -i " $val" -m250 en.txt > $val.data*, where $val is the idiom of interest

work structure that uses mean squared error over the cosine similarities of the input sentences as it's loss function. This allows us to use the contextualised embedding outputs of our BERT networks to find cosine similarity between a given pair of sentences.

3.4 Task Dataset

The sub-task evaluates the quality of idiomatic representation using an STS task, where idioms are replaced with literal and semantic similar paraphrases. When the idiom is replaced by a semantically similar phrase, the similarity score should remain 1. When each component word is replaced literally the similarity between the new sentence and both the idiomatic and semantically similar sentences should be the same. Examples of the data can be found in Table 3.1. This data composes the 'idiom STS' split.

Usage	**Example in Sentence**
Idiomatic	Blockchains, fundamentally, are banking because what they're doing is allowing the transaction of value across networks ... they're doing it in an orthogonally different way," he said Wednesday in what may be his **swan song** in public office.
Literal	Blockchains, fundamentally, are banking because what they're doing is allowing the transaction of value across networks ... they're doing it in an orthogonally different way," he said Wednesday in what may be his **bird song** in public office.
Semantically Similar	Blockchains, fundamentally, are banking because what they're doing is allowing the transaction of value across networks ... they're doing it in an orthogonally different way," he said Wednesday in what may be his **final performance** in public office.

Table 3.1: Example sentences for the Idiomatic STS data. Idiomatic and Semantically similar should be given an STS score of 1, and be given the same score when compared to the literal use.

The data are split into 3 languages: English, Portuguese, and Galician; this allows an assessment of whether the methods used generalise across languages. The dataset also includes 'general STS' data which is used to evaluate whether the idiom specific tuning done to improve performance on the idiomatic STS split has an effect on overall performance.

English and Portuguese are the primary languages and general STS data, from STSBenchmark (Cer et al., 2017) and ASSIN2 (Real et al., 2020) for English and Portuguese respectively, and idiom STS data for both languages are included in the train, dev, eval and test sets. A very small amount (50 examples) of Galician data, comprised of idiom STS data, is also included in the test set.

The task is split into two settings, pre-train and fine-tune. The pre-train setting does not allow for the use of STS score annotated data which includes idioms, whereas any data can be used in the fine-tune setting.

The evaluation metric used in this task is the correlation between the predicted similarities and the gold standard ones, calculated using Spearman's Rank Correlation Coefficient. The Spearman's Rank is calculated for the general STS data and the idiom STS data separately, however, the Spearman's Rank for the entire dataset is used in the final evaluation.

3.4.1 Pre-train Setting

For the pre-train setting, we use the general STS data in English and Portuguese to train the respective models. Due to a lack of available STS data for Galician, it is trained on the Portuguese data, as there is a high level of similarity between the two languages.

Evaluating the models on the dev split, we investigate the optimal number of epochs for the English and Portuguese models. The results (shown in Figure 3.1) show that 45 epochs are optimal for Portuguese and 35 for English. Due to a lack of dev split data for Galician we use the result from the Portuguese model as they are trained on the same data.

3.4.2 Fine-tune Setting

For the fine-tune setting we start with the models from the pre-train setting, and further train them on the Idiom STS data provided as part of the task.

Again we investigate the optimal number of epochs of training on this data (results shown in Figure 3.2). We find that the overall spearman rank is highest after just a single epoch of training, with further training considerably reducing the performance on the general STS data, and thus on the overall STS score. However, further training, up to 50 epochs, continues to increase the performance of the model on Idiom STS data. Therefore, depending on the application and required trade-off, the model can be tuned to either perform better on general STS data or idiom STS data.

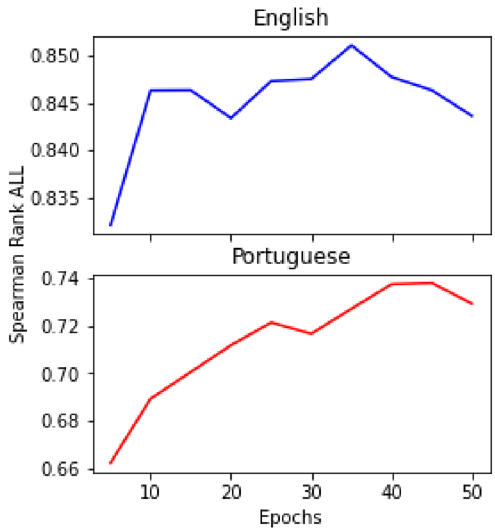


Figure 3.1: Overall Spearman Rank performance on the development set for the English and Portuguese models at different epochs during training

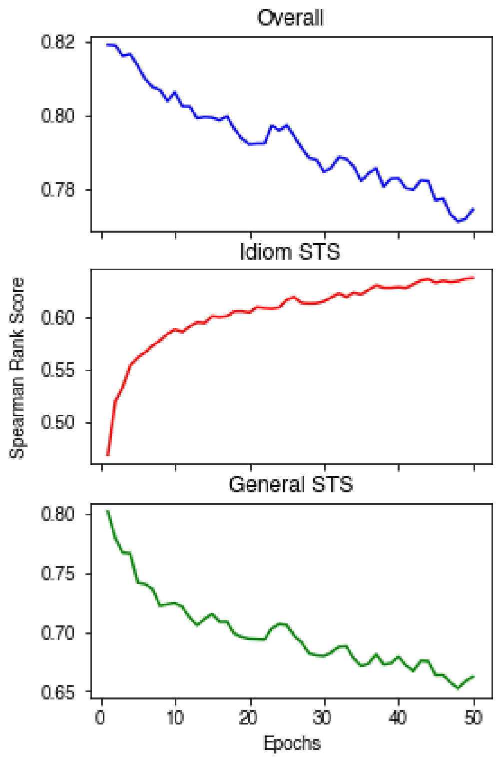


Figure 3.2: Overall and Idiom STS Only Spearman Rank on the development set whilst training on the Idiom STS data

3.4.3 Number of Examples

We also tune the number of examples given for each idiom on the development data. Using BERTRAM we train embeddings for each of the idioms using a range of differ-

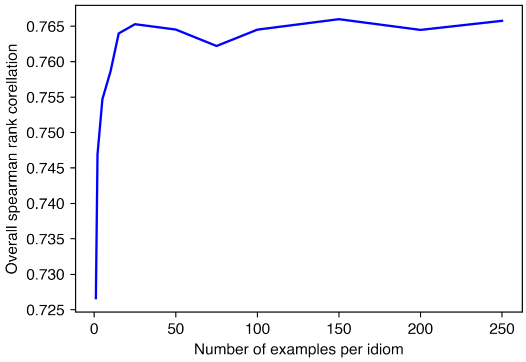


Figure 3.3: Overall Spearman Rank correlation score on the development set with different numbers of examples used to create the idiom embeddings with BERTRAM.

ent numbers of examples from 1-250. The performance of each set of embeddings is evaluated by training the whole system for 10 epochs followed by evaluation on the dev set. Figure 3.3 shows the results of this experiment. The performance increases quickly from 1-15 examples before flattening out. The absolute highest performance is achieved at 150 examples, and so this is the value we use going forward.

3.5 Results

The final results for our system on the test data can be seen in Table 3.2. These scores show significant improvement over the baseline system and led to our system being placed first for the pre-train setting, and second for the fine-tune setting.

Fine-tuning has a much lower effect on the performance of the system when evaluated on the test set than compared with the dev and evaluation sets, with only a small, but significant, rise in overall correlation. Performance rises by only 0.0198 and 0.022 for English and Portuguese respectively, and unlike on dev data we do not see a uniform increase on the SR Idiom score.

3.5.1 Galician Performance

The performance we achieve on the Galician idiom data is much lower than what is seen on the English and Portuguese data. As we didn't have access to any development data for Galician further investigation will be needed to identify the causes of this discrepancy. Due to the smaller amount of Galician data in the cc100 corpus, some idioms did not have the full 150 examples that were used to create the embeddings for

Setting	Language(s)	SR ALL	SR Idiom	SR STS
Pre-Train	EN	0.7445	0.4422	0.8709
Pre-Train	PT	0.7087	0.4806	0.8010
Pre-Train	GL	0.2924	0.2924	-
Pre-Train	**All**	**0.6402**	**0.4030**	**0.8641**
Pre-Train	*EN*	*0.5958*	*0.2488*	*0.8300*
Pre-Train	*PT*	*0.5584*	*0.2761*	*0.7745*
Pre-Train	*GL*	*0.1976*	*0.1976*	-
Pre-Train	*All*	*0.4810*	*0.2263*	*0.8311*
Fine-Tune	EN	0.7643	0.4861	0.8344
Fine-Tune	PT	0.7307	0.4643	0.7908
Fine-Tune	GL	0.2859	0.2859	-
Fine-Tune	**All**	**0.6504**	**0.4124**	**0.8188**
Fine-Tune	*EN*	*0.6684*	*0.4109*	*0.6210*
Fine-Tune	*PT*	*0.6026*	*0.4090*	*0.5523*
Fine-Tune	*GL*	*0.3842*	*0.3842*	-
Fine-Tune	*All*	*0.5951*	*0.3990*	*0.5961*

Table 3.2: Final Spearman Rank (SR) scores of the system on the test set, split into idiom Semantic Text Similarity (STS), general STS, and all datasets. Aggregated results for all languages in bold. Results for the baseline system, also broken down into languages, are in italics.

the English and Portuguese idioms. Additionally, there was no Galician STS data to train the final model on, and even though Portuguese and Galician are very similar, the small difference may lead to differences in the performance.

3.5.2 Error Analysis and Data Issues

To perform analysis on the quality of the created representations we calculate the Spearman's Rank Correlation for each of the idioms in the development set individually. Any idioms with less than 5 occurrences in the development data are removed, as significant correlation scores cannot be achieved with such a low sample size.

When evaluating the performance of the idioms individually, we can see that some of the idiomatic expressions perform much worse than average. For example the spearman rank for score for 'fish story' is just 0.190 when the embedding is trained on 10 random examples.

Analysis of these errors shows that the lower performance can, at least in part, be attributed to different phrase senses in the automatically collected examples. Taking our above example '*fish story*', 3 different phrase senses can be observed in the original randomly selected examples: a tall tale, a literal story about fish, and as a proper noun in the title of the film ‘A Fish Story’. This leads to a divergence in the contexts in the examples, and the contexts for the idiomatic uses, leading to worse embeddings for the idiomatic phrases. Examples of the out-of-scope uses (i.e. proper nouns, misuses, spurious co-occurrences) are given in Table 3.3.

Expression	Category	Example of out-of-scope use
Fish Story	Proper noun	If you haven’t seen **Fish Story** yet, you may want to stop reading this review after this paragraph
Banana Republic	Proper noun	I stopped by **Banana Republic** on my lunch break to pick up a new blazer for work
Panda Car	Spurious	So if you lost your Punto car key or are just looking for a replacement **Panda car** key, give us a call

Table 3.3: Categories of out-of-scope expressions in the randomly sampled CC-100 data tested for filtering in the ablation study.

We can explore this further by producing a manually collected gold standard example set, for the English language subset of the MWEs. Taking the original 250 examples for each idiom, we select 10 gold standard examples. To avoid overfitting our embeddings to this task, we only manually remove examples where the MWE is being used as a proper noun (e.g. the film ’A Fish Story’), or the idiom is being misused, leaving in correct literal and idiomatic uses of the phrase. After removing the proper noun and misused cases, 10 random examples are selected to form our ’gold standard’ example set.

We then compare the spearman scores achieved when the embeddings are trained with the gold standard examples, to scores when the representations are produced using 10 random examples when both models are evaluated on the English split of development set. The results for selected MWEs with the randomly selected (auto) and manually chosen (manual) contexts can be seen in Table 3.4.

The manually selected examples lead to an increase in performance on the Idiom STS data split from 0.406 to 0.450. A small increase from 0.841 to 0.848 overall on the English split can also be observed, however this performance is limited by the general STS score which is unaffected by our manual selection. Particularly large improvements in spearman rank coefficient can be seen on MWEs with multiple meanings (panda car, banana republic, fish story, etc.). Surprisingly, we actually see the performance on some MWEs fall, however this can likely be attributed to the random selection of examples, and variance in the contexts used for each idiom, especially on the MWEs which did not have many usages removed as they are only used in the idiomatic form (eager beaver, chain reaction, etc.).

MWE	Auto	Manual	Change
panda car	0.399	0.851	0.452
banana republic	0.391	0.753	0.362
...	...	...	...
fish story	0.190	0.304	0.114
...	...	...	...
chain reaction	0.356	0.240	-0.116
eager beaver	0.491	0.352	-0.159

Table 3.4: Improvement in correlation, measured using Spearman's Rank Coefficient, when trained on manually chosen examples vs. automatically collected ones.

3.6 Conclusion

We build our system by augmenting BERT models for each language with single token embeddings learnt using BERTRAM. BERTRAM is used due to its high performance on rare words, which share many properties with idioms such as non-compositionality and being rare examples of component pieces. Our results, and subsequent ranking at first place (of six entries) in the pre-train setting and second place (of five entries) in the fine-tune setting, show that BERTRAM can learn high-quality word embeddings for idioms and that this leads to better performance on downstream tasks. Our error analysis shows that BERTRAM is sensitive to the quality of examples it is shown, and that performance can be improved even further by manually selecting a gold set of contexts for each idiom. Future work could look at the differences in performance between the Portuguese and Galician models with the goal of increasing performance on Galician, and perform more analysis to explore the discrepancy in performance between individual idioms further.

Chapter 4

Publication II: Sample Efficient Approaches for Idiomaticity Detection

Dylan Phelps[1,2], Xuan-Rui Fan[2], Edward Gow-Smith[2], Harish Tayyar Madabushi[2], Carolina Scarton[2], Aline Villavicencio[2, 3]

1 Healthy Lifespan Institute, The University of Sheffield, Sheffield, United Kingdom

2 School of Computer Science, The University of Sheffield, Sheffield, United Kingdom

3 Institute of Data Science and Artificial Intelligence, The University of Exeter, Exeter, United Kingdom

Abstract

Deep neural models, in particular Transformer-based pre-trained language models, require a significant amount of data to train. This need for data tends to lead to problems when dealing with idiomatic multiword expressions (MWEs), which are inherently less frequent in natural text. As such, this work explores *sample efficient* methods of idiomaticity detection. In particular, we study the impact of Pattern Exploit Training (PET), a few-shot method of classification, and BERTRAM, an efficient method of creating contextual embeddings, on the task of idiomaticity detection. In addition, to further explore generalisability, we focus on the identification of MWEs not present in the training data. Our experiments show that while these methods improve performance on English, they are much less effective on Portuguese and Galician, leading to an overall performance about on par with vanilla mBERT. Regardless, we believe sample efficient methods for both identifying and representing potentially idiomatic MWEs

are very encouraging and hold significant potential for future exploration.

4.1 Introduction and Motivation

The handling of idiomaticity is an important part of natural language processing, due to the ubiquity of idiomatic multiword expressions (MWEs) in natural language (Sag et al., 2002). As such, it is an area where the performance of state-of-the-art Transformer-based models has been investigated (Yu and Ettinger, 2020; Garcia et al., 2021b; Nandakumar et al., 2019), with the general finding being that, through pre-training alone, these models have limited abilities at handling idiomaticity. However, these models are extremely effective at transfer learning through fine-tuning, and thus are able to perform much better on supervised idiomatic tasks (Fakharian and Cook, 2021; Kurfalı and Östling, 2020), where significant amounts of labelled data is provided.

Unfortunately, individual MWEs tend to occur infrequently in natural text, making it harder to train models to capture the idiomatic meaning due to the lack of available training data. As such, it is important to be able to find methods of identifying potentially idiomatic MWEs using relatively less data. To address this question, we focus on *sample efficient* methods for the task, taking two perspectives. The first is an evaluation of a few-shot method on the task of zero-shot idiomaticity detection. In particular, we evaluate Pattern Exploit Training (PET) (Schick and Schütze, 2021a), which has been shown to be an effective few-shot method on other tasks (Schick and Schütze, 2021b). The second is an evaluation of the effectiveness of better representations of MWEs, created using a sample efficient strategy, namely BERTRAM (Schick and Schütze, 2020). Both of these are explored in the zero-shot context, where training data does not include MWEs present in the test data. So as to ensure reproducibility and to enable others to build upon this work, we make the programme code and models publicly available[1].

4.1.1 Research Questions and Contributions

Given the need for sample efficient methods when dealing with idiomaticity, this work is aimed at exploring the following questions:

[1] https://github.com/drsphelps/idiom-bertram-pet

- How effective are few-show methods on the task of zero-shot idiomaticity detection? We evaluate Pattern Exploit Training (PET) (Schick and Schütze, 2021a), which has been shown to be an effective few-shot method on other tasks (Schick and Schütze, 2021b).

- Given that prior work has shown pre-trained language models do not adequately capture multiword expressions, in particular those which are idiomatic, how effective is improving their representations on the task of detecting idiomaticity? In particular, we use BERTRAM (Schick and Schütze, 2020) as a sample efficient strategy for creating representations of MWEs.

From our experiments, we find that both BERTRAM and PET are able to outperform mBERT (Devlin et al., 2019) significantly on the English portion of the test data, which is a promising result. However, both of these models perform worse overall due to their significantly lower performance on Portuguese. We explore potential reasons for this poor performance on non-English languages: for PET our patterns are all in English and a multilingual model is used instead of a language specific one. However, an error analysis (Section 4.5.1) suggests that these are not the reasons for the lower performance on non-English languages. In BERTRAM, however, a monolingual model is used for each language, which might have contributed to the drop in performance. We believe that these results point to the need for further exploration in languages other than English.

Additionally, our exploration using BERTRAM is, to the best of our knowledge, the first work to explore the relation between the representation and detection of idiomaticity.

The rest of this paper is structured as follows. We begin in Section 4.2 by presenting a quick overview of work related to MWE identification, before presenting more details of the methods we make use of in this work. We then provide an overview of the data and task we use for our evaluation in Section 4.3, before presenting the methods in Section 4.4. We then present our results and a discussion of what these results imply in Section 4.5, before concluding in Section 4.6.

4.2 Related Work

Despite idiomaticity detection being a problem that has been widely explored (Constant et al., 2017), the impact of better MWE representations, especially within contextualised models, has not been well studied. To this end we use BERT for Attentive Mimicking (BERTRAM) (Schick and Schütze, 2020), which has been shown to perform well on idiom representation tasks (Phelps, 2022), to evaluate the effect idiom representations have on detection. Additionally, we apply a few-shot learning technique Pattern Exploit Training (PET) (Schick and Schütze, 2021a), to assess whether the relatively new paradigm of few-shot learning can be applied to this task successfully.

4.2.1 PET

PET (Schick and Schütze, 2021a,b) is a semi-supervised training method that improves performance in few-shot settings by integrating task descriptions into examples.

A Pattern is used to map each example into a cloze-style question with masked out tokens, for example *'X. It was [MASK]'*, where X is the input example, could be used for a sentiment classification task. A Verbaliser maps the task classes into outputs from the masked language model (MLM), for example positive/negative labels map to the words 'good'/'bad' in the MLM's vocabulary (label tokens), and is combined with the pattern to form a Pattern Verbaliser Pair (PVP). The probability of each class is then calculated using softmax over the logits for each label token.

For each PVP, an MLM can be fine-tuned on the small amount of labelled data. Knowledge is distilled from multiple PVPs by combining the predictions on the unlabelled data and using it as a larger labelled dataset to train another classifier. This allows for multiple patterns and verbalisers to be used without having to choose the best performer for each task, which may also change depending on the data split.

iPET

iPET (Schick and Schütze, 2021a) is a variation where each PVP's model is trained iteratively using a gradually increasing training set made up of labelled examples from another model's predictions in the previous iteration. Despite using the same PVPs and MLMs, iPET has been shown to improve the performance on a number of tasks (Schick and Schütze, 2021b).

4.2.2 BERTRAM

BERTRAM (Schick and Schütze, 2020) is a model for creating embeddings for new tokens within an existing embedding space, from a small number of contexts. To create an embedding for a token with a number of contexts, a form embedding is first created using embeddings trained for each of the n-grams in the token. This form embedding is then passed as an input, alongside the embeddings for words in the context, into a BERT model. An attention layer is then applied over the contextualised embedding output from BERT for each context to create the final embedding for the token.

The model is trained using embeddings for common words as 'gold standard' embeddings, with the distance from the embedding created by the model and the 'gold standard' embedding being used as the loss function.

4.3 Dataset and Task Description

In evaluating the models presented in this work we use the Task 2 of SemEval 2022: Multilingual Idiomaticity Detection and Sentence Embedding (Tayyar Madabushi et al., 2022). This task aims at stimulating the development and evaluation of improved methods for handling potentially idiomatic MWEs in natural language. While there exist datasets for evaluating models' ability to identify idiomaticity (Haagsma et al., 2020; Korkontzelos et al., 2013; Cordeiro et al., 2019; Garcia et al., 2021b; Shwartz and Dagan, 2019), these are often not particularly suited to investigating a) the transfer learning capabilities across different data set-ups b) the performance of pre-trained contextualised models.

The task consists of two subtasks: Subtask A, which is focused on the detection of idiomaticity, and Subtask B, which is focused on the representation of idiomaticity. In this work, we are interested in the task of idiomaticity classification, since we wish to investigate how our models can identify idiomaticity in text without having to generate semantic similarity scores. As such, we restrict our attention to Subtask A. We also want to see how our models perform when MWEs in the test data are disjoint from those in the training data, as we argue this means the models cannot easily leverage statistical information garnered from the training data, but must instead have some 'knowledge' of idiomaticity in general. As such, we also restrict our attention to the zero-shot setting of the SemEval task. The dataset consists of three languages: English, Portuguese and Galician. In the training data there are 3,327 entries in English, and 1,164 entries in Portuguese. There is no Galician training (or development) data in the zero-shot setting, to test the ability of models at cross-lingual transfer. In the test set, there are 916 English, 713 Portuguese, and 713 Galician examples, and macro F1 score is used as an evaluation metric.

It should be noted that the dataset provided by (Tayyar Madabushi et al., 2022) consists of four data splits: The training set, two development sets and the test set. Of the two development sets, the first - called the 'dev' split - includes gold labels and the second - called the 'eval' split - does not include gold labels but requires submission to the competition website. We report our results on the 'eval' set to maintain consistency with the SemEval task.

4.4 Methods

In this section we detail our use of PET, iPET and BERTRAM for the task of idiomaticity detection.

4.4.1 PET and iPET

During our experiments with PET and its variants, we define and test 5 Pattern Verbaliser Pairs, shown in Table 4.1, with example application for a given sentence shown in Table 4.2. P1 and P2 are generic prompts which do not give the model much more information about the example, whereas P3, P4, and P5 include the whole idiom within the prompt. We hypothesise that this will allow the model to understand which part of the example it should be focusing on. Each of the patterns we define is in English, even when the example sentence and idiom are in Portuguese or Galician — we will investigate the effect that this has on the final performance across the languages, as we hypothesise this may not have an impact given our use of a multilingual model.

For each PVP, we train a classification model using mBERT as the MLM. Furthermore, we train a standard PET model using all of the patterns. An iPET model is also trained, however to evaluate how using only generic prompts affects the results, we only train our iPET model using PVPs P1 and P2, for 2 iterations. Each of the model setups is trained 3 times using different random seeds, and the final distilled model is then used to produce the presented results.

Pattern Number	Pattern	Literal Token	Idiom Token
P1	X: _ _ _ _	literal	phrase
P2	(_ _ _ _) X	literal	phrase
P3	X. [IDIOM] is _ _ _ _ literal.	actually	not
P4	X. _ _ _ _, [IDIOM] is literal.	yes	no
P5	X. [IDIOM] is _ _ _ _ $[\text{IDIOM}]_2$	actually	not

Table 4.1: Pattern Verbaliser Pairs used in the task. X represents the example sentence, [IDIOM] is the idiom found in the example, and $[\text{IDIOM}]_2$ represents the 2nd component word of the idiom

Pattern	Examples
P1	The museum hung a big picture on the wall: **literal** Focus on the big picture of your argument: **phrase**
P2	(**literal**) The museum hung a big picture on the wall (**phrase**) Focus on the big picture of your argument
P3	The museum hung a big picture on the wall. Big picture is **actually** literal. Focus on the big picture of your argument. Big picture is **not** literal.
P4	The museum hung a big picture on the wall. **yes**, big picture is literal Focus on the big picture of your argument. **no**, big picture is literal
P5	The museum hung a big picture on the wall. Big picture is **actually** picture. Focus on the big picture of your argument. Big picture is **not** picture.

Table 4.2: An idiomatic and literal instance of 'big picture' formatted with each pattern, and shown with the expected verbaliser token (in bold) that the model should give a higher probability.

Additionally, we investigate how the number of labelled examples affects the achieved performance for each of the model setups discussed. We train the models using 10, 100, and 1000 labelled examples separately, with the examples chosen randomly across English and Portuguese, but with the split of idiomatic and literal uses being kept at 50/50. The PET and iPET models then have access to 3,000 unlabelled examples to use within their training tasks.

We evaluate each model setup and labelled example set size combination on the *eval* set, before choosing the best-performing combination for each PET variant to evaluate on the test set. The results from the *eval* set can be seen in Table 4.3. Here we see that PET-all trained on 1000 labelled examples performs best overall, beating the individual pattern models, a result also seen in the original paper (Schick and Schütze, 2021a). The lack of example specific prompts causes iPET to perform poorly when compared to the individual task specific patterns, and when compared to the best PET-all model.

Model	EN	PT	Overall
mBERT (Tayyar Madabushi et al., 2021)	0.7420	0.5519	0.6871
PET-all (10 labelled)	0.4365	0.2901	0.4267
PET-all (100 labelled)	0.5908	**0.5718**	0.5888
PET-all (1000 labelled)	**0.7820**	0.5619	**0.7164**
PET-P1 (1000 labelled)	0.6386	0.5507	0.6278
PET-P2 (1000 labelled)	0.6905	0.5495	0.6607
PET-P3 (1000 labelled)	0.7493	0.5474	0.6981
PET-P4 (1000 labelled)	0.7441	0.5315	0.6860
PET-P5 (1000 labelled)	0.7551	0.5680	0.7032
iPET (1000 labelled) [P1 & P2]	0.6701	0.5648	0.6522

Table 4.3: Macro F1 on the *eval* set, broken down into each language, for each of the models. Highest score for each language (or overall) shown in bold.

The highest scoring PET model (PET-all) and our iPET model are evaluated on the test dataset in Section 4.5.

4.4.2 BERTRAM

To evaluate the effect that improved idiom representations have on this idiom detection task, we use the same BERTRAM setup as presented in Phelps (2022), that was shown to give greatly improved performance over the baseline system for Subtask B, the task of representing idiomaticity. We use the same BERTRAM models: the English model presented in the original BERTRAM paper (Schick and Schütze, 2020), and the Portuguese and Galician models that were trained for Subtask B from data in the CC100 corpus. Unlike the English BERTRAM model, Phelps (2022) does not use one token approximation when training the Portuguese and Galician models. Embeddings for each of the idioms in the task datasets were generated with the appropriate BERTRAM model using 150 examples scraped from the CC100 dataset. 150 examples were chosen as this was shown to have the highest performance on Subtask B. It should be noted that the BERTRAM models were used to create representations of MWEs in the test set. While this does not require labelled data associated with MWEs (thus remaining a zero-shot task), it does require knowledge of which phrases need to have explicit representations created.

As we have separate BERTRAM models for each language that are trained to mimic embeddings from single language BERT models, we split the system and data into English, Portuguese and Galician. The English model uses BERT base (Devlin et al., 2019), and is trained on the 3,327 English training examples found in the training set. The Portuguese model uses BERTimbau (Souza et al., 2020), and Galician uses BERTinho (Vilares et al., 2021), and as there is no Galician training data available, both are trained on the 1,164 Portuguese examples. Each model has the MWEs from the relevant language added to its embedding matrix.

4.5 Results and Discussion

Table 4.4 presents the results of our best PET-based models alongside our BERTRAM-based model on the test set, as well as the mBERT system presented in (Tayyar Madabushi et al., 2022), for comparison. For each model we present the macro F1 score on the test set for each language, as well as the overall macro F1 score.

Model	EN	PT	GL	Overall
mBERT (Tayyar Madabushi et al., 2022)	0.7070	**0.6803**	0.5065	**0.6540**
BERTRAM	**0.7769**	0.5017	0.4994	0.6455
PET-all (10 labelled)	0.5197	0.2634	0.2090	0.4128
PET-all (100 labelled)	0.6777	0.5014	0.4902	0.5694
PET-all (1000 labelled)	0.7281	0.6253	**0.5110**	0.6446
iPET (1000 labelled) [P1 & P2]	0.6604	0.5676	0.4735	0.5879

Table 4.4: Macro F1 on the *test* set, broken down into each language, for each of the models. Highest score for each language (or overall) shown in bold.

An increase in performance over mBERT by our BERTRAM model is seen for the English split, with the score on the Galician split not seeing a significant change. The overall score for BERTRAM is brought down by a much lower score on the Portuguese data, however, meaning no overall increase in performance is seen. A similar picture is seen for the PET-all (1000 examples) model, with a higher macro F1 in both English and Galician, and a lower score in Portuguese, leading to an overall lower macro F1 across the entire test dataset. As found on the example data, the iPET model which was only trained on the non-example specific prompts (P1 and P2) performs very poorly.

The significant boost from using BERTRAM on English seems to indicate that the improved representations also lead to better classification, despite the lacklustre performance on Galician and Portuguese. We believe that this drop in performance is either because one-token approximation was not used in creating the non-English BERTRAM models, or because mBERT, trained on all three languages simultaneously, is trained on more data than each of our monolingual models. This lack of training data does not affect our English model as there is a more training data in English than in Portuguese and none at all in Galician. We perform a language specific error analysis to explore the causes of this drop in performance (Section 4.5.1).

It is interesting to note that pre-trained language models can identify idiomaticity in a zero-shot and sample efficient context *even when prior work has shown that they do not encode idiomaticity very well* (Garcia et al., 2021b). We believe that this implies that, while these models do not encode idiomaticity, they encode enough related information to be able to *infer* idiomaticity from relatively little data.

Unsurprisingly, 'highlighting' the phrase that is potentially idiomatic by adding the phrase to the pattern, as in patterns P3, P4 and P5 (see Table 4.1), significantly improves a model's ability to identify idiomaticity, which is consistent with results presented by(Tayyar Madabushi et al., 2021).

Research Questions The results presented herein suggest that few-shot learning methods are indeed effective on the task of idiomaticity detection despite the lower accuracy on Portuguese and Galician. Similarly, our results support the conclusion that improved MWE representations does have an impact on improved detection.

4.5.1 Error Analysis

The effectiveness of PET on the English split of the task suggests that pre-trained language models can effectively identify idiomatic MWEs in a sample efficient manner. However, the overall drop in performance on the task can be attributed to lower performance on non-English languages when compared to the results achieved by Tayyar Madabushi et al. (2021).

One possibility for the decrease in performance is the use of English prompts across all the languages. This leads to the inputs for English examples being monolingual and the inputs for non-English examples to be multilingual, which may cause confusion in the output logits for the verbalizer tokens from which PET draws it's predictions.

To investigate this further we translate one of our patterns, P4, into both Portuguese and Galician and evaluate the performance on the entire *test* split. P4 was chosen as it was one of the better performing patterns for English in our initial experiments (Table 4.3), and was easily translated into the two languages. The translations can be seen in Table 4.5.

Language	Pattern	Literal Token	Idiom Token
EN	X. _____, [IDIOM] is literal.	yes	no
PT	X. _____, [IDIOM] é literal.	sim	não
GL	X. _____, [IDIOM] é literal.	si	non

Table 4.5: The translations of P4 into Portuguese and Galician

As shown in Table 4.6, the use of Portuguese and Galician prompts does not increase the performance in the respective language. For Portuguese the model with Portuguese prompts achieves 0.6260 macro F1 compared to 0.6373 for that with English prompts. Galician shows similar results, with 0.5154 macro F1 for the model with prompts in Galician and 0.5365 for those in English.

Model	Prompt Language	EN	PT	GL	Overall
mBERT	N/A	0.7070	0.6803	0.5065	0.6540
PET-P4 (1000 labelled)	EN	0.7161	0.6373	0.5365	0.6581
PET-P4 (1000 labelled)	PT	0.6994	0.6260	0.4964	0.6283
PET-P4 (1000 labelled)	GL	0.7040	0.5997	0.5154	0.6279

Table 4.6: Macro F1 on the *test* set, broken down into each language, for PET using prompts in each of the task languages.

Additionally, we use multilingual BERT which was trained on a lot more English training data than Portuguese or Galician language. To investigate the impact of this on our results, we extract only the Portuguese section of the training and test data and compare the performance of multilingual BERT with Portuguese BERT (Souza et al., 2020). Surprisingly, we find that the there isn't a significant difference between the performance of multilingual BERT and Portuguese BERT, with overall macro F1 of 0.4541 and 0.4621, respectively.

4.6 Conclusions and Future work

This work presented our exploration of *sample efficient* methods for idiomaticity detection, crucial given the infrequent occurrence of specific MWEs in natural language text. Our experiments show that these methods are extremely promising and have great potential.

In future work, we intend to raucously evaluate and find solutions to the problem of lower performance on non-English test splits. We also intend to explore other variations of BERTRAM (e.g. one-token approximation) in bridging the performance gap between English and the other languages.

As noted earlier, we show that pre-trained language models can identify idiomaticity in a zero-shot and sample efficient context *even when prior work has shown that they do not encode idiomaticity very well.* As such, an important avenue of future exploration is the generalisation of these methods to develop models capable of identifying *the notion of idiomaticity*, much like humans are able to grasp that certain phrases are clearly non-compositional.

Chapter 5

Publication III: Sign of the Times: Evaluating the use of Large Language Models for Idiomaticity Detection

Dylan Phelps[1,2], Thomas Pickard[2], Maggie Mi[2], Edward Gow-Smith[2]. Aline Villavicencio[2,3]

1 Healthy Lifespan Institute, The University of Sheffield

2 School of Computer Science, The University of Sheffield

3 Institute of Data Science and Artificial Intelligence, The University of Exeter

Abstract

Despite the recent ubiquity of large language models and their high zero-shot prompted performance across a wide range of tasks, it is still not known how well they perform on tasks which require processing of potentially idiomatic language. In particular, how well do such models perform in comparison to encoder-only models fine-tuned specifically for idiomaticity tasks? In this work, we attempt to answer this question by looking at the performance of a range of LLMs (both local and software-as-a-service models) on three idiomaticity datasets: SemEval 2022 Task 2a, FLUTE, and MAGPIE. Overall, we find that whilst these models do give competitive performance, they do not match the results of fine-tuned task-specific models, even at the largest scales (e.g. for GPT-4). Nevertheless, we do see consistent performance improvements across model scale. Additionally, we investigate prompting approaches to improve performance, and discuss the practicalities of using LLMs for these tasks.

5.1 Introduction

Large, pre-trained language models (LLMs) are becoming increasingly popular in academic, industrial, and lay spheres due to their ability to perform well across a range of tasks in a zero-shot or few-shot prompting set-up, including question answering, common-sense reasoning (Yang et al., 2023; Team et al., 2024), and machine translation (Xu et al., 2024; Koshkin et al., 2024; Dabre et al., 2023). Despite this, there is yet to be an analysis of how well such models are able to handle potentially idiomatic language. Much previous work has shown that smaller, encoder-only transformer models have poor performance in identifying and representing idiomatic expressions when pre-trained on a large general dataset (Nandakumar et al., 2019; Garcia et al., 2021b). However, the performance of such models increase hugely when they are fine-tuned on a task-specific dataset containing a large number of idiomatic expressions (Tayyar Madabushi et al., 2021; Zeng and Bhat, 2021). This fine-tuning procedure, however, requires dedicated hardware and training, something that isn't possible with LLMs on an academic budget.

In this work, we benchmark the performance of several widely-used LLMs (using both software-as-a-service remote implementations and local instances) on three in-context idiomaticity detection datasets; the idiom portion of FLUTE (Chakrabarty et al., 2022b), MAGPIE (Haagsma et al., 2020), and SemEval 2022 Task 2a (Tayyar Madabushi et al., 2022). FLUTE and MAGPIE cover English (EN) only, while the SemEval dataset also includes expressions in Brazilian Portuguese (PT-BR) and Galician (GL).

Overall, our experiments show that large LLMs give competitive performance on idiomaticity datasets, which can be generally applied due to the lack of type specific fine-tuning, but nevertheless lag in general behind much-smaller finetuned encoder-only models. We also find that idiomaticity detection performance still scales with the number of parameters in the model. Finally, we discuss a number of considerations affecting the models' performance and the practicality of using them for idiomaticity detection, including the training dataset and the capability of the model to follow instructions given in the prompt.

5.2 Datasets

We investigate the performance of LLMs on three datasets consisting of potentially idiomatic expressions in context. The datasets are chosen to provide a diverse set of potentially idiomatic expressions which feature a range of morphological forms and variations across two different tasks: textual entailment and idiomaticity detection. 1,859 different English target expressions are represented across the three datasets. We focus on English, but the inclusion of Semeval 2022 Task 2a allows us to additionally explore performance across languages.

5.2.1 FLUTE

FLUTE (Chakrabarty et al., 2022b) frames the understanding of four kinds of figurative language (sarcasm, simile, metaphor and idioms) as a natural language inference (NLI) task, in which pairs of literal and figurative sentences are labelled as either entailing or contradicting one another. The sentence pairs are generated using a model-in-the-loop approach, with base text generated by GPT-3 which is then edited by crowdworkers and reviewed by experts.

For our analysis, we consider only the idiom section of the FLUTE dataset, which consists of 1,768 training examples across 479 idioms and a further 250 test examples across 69 idioms. No idiom appears in both the training and test sets.

Chakrabarty et al. (2022b) provide benchmark performance metrics using T5 models (Raffel et al., 2020) on the FLUTE training data, reporting 79.2% accuracy (0.791 macro F1). A FigLang22 shared task using the FLUTE dataset (Saakyan et al., 2022) attracted several entries, with the best-performing systems developed by (Gu et al., 2022) and (Bigoulaeva et al., 2022). The latter adopt a pipeline approach, improving the T5 baseline by sequentially fine-tuning on e-SNLI dataset (Camburu et al., 2018) and IMPLI (which incorporates figurative language) (Stowe et al., 2022), followed by the task dataset. Using the authors' published outputs, we calculate a macro F1 of 0.952 on the idiom portion of the FLUTE test set.

5.2.2 SemEval 2022 Task 2a

SemEval 2022 Task 2a (Tayyar Madabushi et al., 2022) is a binary classification idiomaticity detection task, in which a potentially idiomatic noun compound, as used in a given context sentence, must be labelled as either literal or idiomatic. The dataset includes compounds across a range of idiomaticity, including fully compositional (*insurance company*) as well as partially (*eager beaver*) and entirely opaque (*sugar daddy*) items. The task offers both “one-shot" and “zero-shot" settings; the former is evaluated with new context instances of previously-seen items, while the latter uses compounds not present in the training data for evaluation.

The test set for the task contains 50 compounds each in English (with 916 instances), Brazilian Portuguese (713 instances) and Galician (713 instances).

Table 5.1 shows the macro F1 scores in the zero-shot and one-shot settings for the baseline models (fine-tuned multilingual mBERT, per Tayyar Madabushi et al., 2021) and the best-performing entries to the shared task[1].

Setting	Reference	Language			
		EN	PT	GL	All
Zero-Shot	Best	0.902	0.828	0.928	0.890
	Baseline	0.707	0.680	0.507	0.654
One-Shot	Best	0.964	0.894	0.937	0.939
	Baseline	0.886	0.864	0.816	0.865

Table 5.1: Reference scores (macro F1) for SemEval 2022 Task 2a.

5.2.3 MAGPIE

MAGPIE (Haagsma et al., 2020) is a corpus of instances of potentially idiomatic expressions (PIEs – expressions which have multiple senses, including at least one with a high level of idiomaticity), in which each instance has been annotated as either idiomatic, literal, or other (proper noun, etc.) by a group of crowd-sourced workers. The PIEs in the dataset are chosen from three online dictionaries and so have a wide range of forms and frequencies.

The final dataset consists of 56,622 annotated instances, of which 70% are idiomatic, 28% are literal and 1% are other. In our experiments we use the test split of the randomly split dataset, which has 4,840 instances across 1,134 PIEs).

Haagsma et al. (2020) do not provide baseline models for the MAGPIE data, but several benchmarks are provided by Zeng and Bhat (2021).

[1]For the one-shot setting, the best-performing model is a fine-tuned multilingual XLM-RoBERTa, as described in Chu et al. (2022).

5.2.4 Construction Artifacts

Recent work by Boisson et al. (2023) has found that language models tuned for metaphor identification (in which they include idiomaticity detection) on artificially-constructed datasets (i.e. those not sampled from 'naturally-occurring' text) can perform well when the target expression or the surrounding context are hidden from the model, "in both cases close to the model with complete information".

As our experiments employ pre-trained LLMs without fine-tuning for the idiomaticity detection task, we anticipate that the concerns highlighted by Boisson et al. (2023) should not affect our findings. While the training regimes for many of the models we examine are not public, it seems likely that they have consumed large quantities of training data containing 'naturally distributed' idiomatic expressions.

It is also worth noting that we can not rule out the possibility that these LLMs' training data includes the training or test datasets under evaluation[2], and it is likely (for SemEval and MAGPIE) that the context sentences could have been 'seen' by the models during training (albeit without idiomaticity markers), as they are taken from online sources.

5.3 Models

To be able to compare results from a range of currently-available LLMs, we evaluate both software-as-a-service (SaaS) and local instances of open models. To maximise applicability of our findings to researchers, we focus on local instances that can be run on consumer-level hardware (targeting a machine with 32GB RAM and 12GB VRAM).

Table 5.2 summarises the models used in our experiments, including the parameter count (where available), cost to run for SaaS models, and whether the training dataset is multilingual.

[2]The SemEval test set is publicly available only without labels; FLUTE and MAGPIE are public.

Model	Params (billions)	Cost ($US per 1000 tokens)	Multilingual
GPT-3.5-turbo	Unknown	0.0005	Y
GPT-4-turbo	Unknown	0.01	Y
GPT-4	Unknown	0.03	Y
Gemini-1.0 Pro	Unknown	0.000125	Y
Llama2-7B-chat	7	N/A	N
Llama2-13B-chat	13	N/A	N
Llama2-70B-chat	70	N/A	N
Phi-2	2.5	N/A	N
Mistral-7B	7	N/A	N
Flan-T5-Small	0.08	N/A	Y
Flan-T5-Base	0.25	N/A	Y
Flan-T5-Large	0.78	N/A	Y
Flan-T5-XL	3	N/A	Y
Flan-T5-XXL	11	N/A	Y

Table 5.2: Characteristics of the models evaluated.

5.3.1 Software-as-a-service Models

OpenAI

OpenAI models are seen to be the current state of the art in SaaS models. GPT-4 (OpenAI et al., 2024a), their current largest model, has been shown to achieve or exceed human-level performance in a number of commonly used benchmarks. We evaluate GPT-3.5-turbo (gpt-3.5-turbo-0613), GPT-4-turbo (gpt-4-0125-preview) and GPT-4 (gpt-4) in this work. GPT-3.5 is a smaller model created as a test run during the development of GPT-4, and GPT-4-turbo is an optimised and more recent variant of GPT-4. The parameter counts for these models are not known, but it is assumed that GPT-4 is substantially larger than GPT-3.5.

Google

Google provides access to a number of models of varying size and price through its VertexAI API. In this work we evaluate the performance of the Gemini Pro 1.0 model. Gemini Pro is trained on a multimodal and multilingual dataset and its performance exceeds that of GPT-3.5 on a number of benchmarks (Team et al., 2024).

5.3.2 Local Models

Additionally, we evaluate the performance of popular open models that can be run locally. The models chosen are the Llama2 models, (Touvron et al., 2023) Llama2-7B-chat and Llama2-13B-chat, Phi-2 (Li et al., 2023; Abdin et al., 2023), and the CapybaraHermes[3] variant of Mistral-7B (Jiang et al., 2023).

To ensure that the models can be run on consumer-level hardware we use quantized variants of each model with 7B or more parameters. Quantization (Dettmers et al., 2022; Frantar et al., 2023) involves converting each parameter from full 16-bit floating point numbers to a set of 2^n discrete values. This massively reduces the size of the models so they can be run on a wider range of hardware, with a trade-off of lower performance. We use **Q5_K_S quantisation variants**, which use 5-bit quantization, provided by TheBloke on Huggingface[4]. 5 bit quantization has been shown to have minimal impact on the performance of the model[5].

To run the models we use the Huggingface transformers library (Wolf et al., 2020) for Phi-2 and llama.cpp[6] for all the quantized models.

5.3.3 Multilingual Models

We also explore the performance of multilingual models. In particular, we target our exploration to variants of the Flan-T5 models (Chung et al., 2024): Flan-T5-Small, Flan-T5-Base, Flan-T5-Large, Flan-T5-XL, and Flan-T5-XXL.

We are interested in how multilingual models' performance on idiomatic language-related tasks differs from monolingual ones. Moreover, we want to investigate the extent to which the performance is impacted by model size.

5.4 Results

Our main results across the three datasets (using our default prompts) are shown in Table 5.3. To make our results representative and generalisable, we ran the models multiple times, where not computation or cost prohibitive – all of the Flan models were run three times, whilst the Gemini Pro and GPT-3.5 models were run twice on

[3] `https://huggingface.co/argilla/CapybaraHermes-2.5-Mistral-7B`
[4] `https://huggingface.co/TheBloke`
[5] See `https://github.com/ggerganov/llama.cpp/pull/1684`.
[6] `https://github.com/ggerganov/llama.cpp`

	SemEval	FLUTE	MAGPIE
GPT-3.5-Turbo	0.645	0.820	0.559
GPT-4-turbo	0.668	0.936	0.860
GPT-4	0.636	0.936	0.896
Gemini 1.0 Pro	0.672	0.924	0.721
Phi-2	0.447	0.458	0.531
Llama2 (7B-chat)	0.479	0.373	0.314
Llama2 (13B-chat)	0.505	0.602	0.483
CapybaraHermes-2.5-Mistral-7B	0.539	0.812	0.587
Flan-T5-Small	0.333	0.333	0.203
Flan-T5-Base	0.390	0.764	0.213
Flan-T5-Large	0.424	0.872	0.290
Flan-T5-XL	0.452	0.956	0.456
Flan-T5-XXL (11.3B)	0.514	0.940	0.753
baseline	0.654	0.791	0.872
best	0.890	0.952	0.955

Table 5.3: Main results of our models across the three idiomaticity datasets. All results presented are macro F1 over the two classes. Baseline results are taken from Tayyar Madabushi et al. (2021) (BERT), Chakrabarty et al. (2022b) (T5) and Zeng and Bhat (2021) (BERT). 'Best' results are taken from Chu et al. (2022) (BERT, RoBERTa), Bigoulaeva et al. (2022) (T5) and Zeng and Bhat (2021) (BERT, *RNN*). Most baseline and best models are fine-tuned variants of pre-trained transformers models, the model family is given after each citation.

SemEval, which is particularly important for reducing the variance of the results when testing different prompting methods; all other models were run once only.

Comparing the results with the baseline and best-performing models, we can see that while the performance of large, contemporary LLMs may be higher than out-of-the-box encoder-only models, there is still a gap between them and the results which can achieved by encoders fine-tuned to the particular tasks. However, given the work of Boisson et al. (2023) on construction artifacts within datasets for idiomaticity detection, the ability of LLMs to disambiguate a wide-range of PIEs without additional fine-tuning shows the general ability of these models to detect idiomaticity, which may not have been achieved by fine-tuned encoders.

5.4.1 Model Scaling

With the exception of the Mistral-7B model, there is a significant gap in performance between the smaller, locally-run models and the larger SaaS models. We can also see the same trend for our Llama2 models, where the larger Llama2-13B model outperforms the smaller Llama2-7B one on all datasets and splits. From the results of the Flan-T5 model variants, as shown in Figure 5.1, there is a clear trend that increasing model size leads to improved performance. This trend appears to slow down somewhat after model size reaches around 3B parameters (Flan-T5-XL), though performance on the MAGPIE dataset continues to grow.

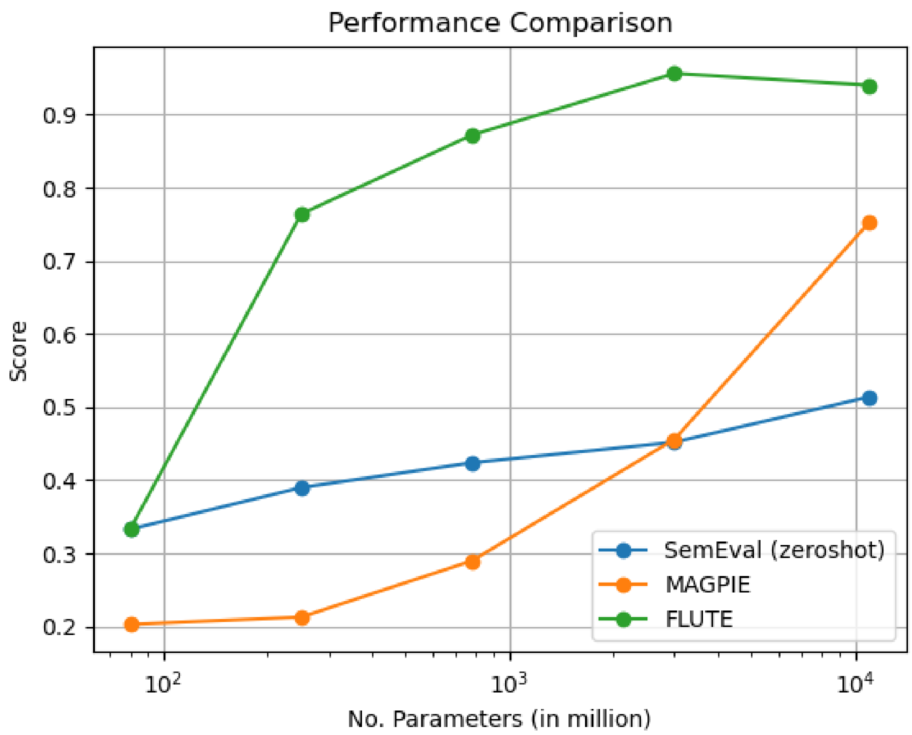


Figure 5.1: Performance on the three datasets for different Flan-T5 model sizes.

5.4.2 Prompts

Prompting has been shown to significantly effect the performance of LLMs (Zhao et al., 2021), however it is hard to decide before testing which prompts will lead to higher performance. As the prompt space is so large, with effective infinite word choice and prompt size, it is not possible at this time to systematically select the best prompt for a given task. One could theoretically define a limited search space based on varying a small number of keywords and combinations of techniques, however we leave this as future work. Instead, in this work we non-systematically test a number of prompts during our exploration phase, and present a prompt we found to be high quality as the ‘default’. We also present exploration of other prompting techniques below.

Due to the differing input formats required by the various models, we use slightly different prompts. Here, we show our default prompts used for the GPT models. For SemEval and MAGPIE, we use:

“Disambiguate whether the given expression is used idiomatically or literally in the given context, returning ’i’ if the expression is being used idiomatically or ’l’ if literally. Expression: <PIE>. Context: <target sentence>. Only return one letter (i or l).”

For the FLUTE entailment task, we use:

“Disambiguate whether the second sentence follows from the first, returning ’entailment’ if it does, and ’contradiction’ if not. Sentence 1: <premise sentence> Sentence 2: <hypothesis sentence>.”

5.4.3 Prompt Engineering

We investigate the effect of several prompt variations on performance for GPT-3.5-turbo on the English SemEval test set. As part of the OpenAI API, there are two prompts: “system” and “user”. We first tried using the system prompt to define the task for the model, but obtained better performance using only the user prompt – this aligns with the experiences of others that GPT-3.5 often doesn’t follow the system prompt well, unlike GPT-4[7].

We present our results for this in Table 5.4. Note that variation between runs using the same prompting strategy is high (up to 0.04 F1), which leads to difficulty in discerning the effect of changing the prompt.

	EN
Default	0.739
“Expert in language use”	0.635
“Expert in language use” + Idiomatic vs. Compositional	0.717
“Expert in Idiomatic Language”	0.538
No “Only return one letter (i or l).”	0.633

Table 5.4: Results (macro F1) on the English test set of SemEval with GPT-3.5-turbo using prompt engineering.

Expert impersonation is motivated by work which has shown that prompting LLMs to impersonate domain experts can lead to higher performance (Salewski et al., 2023).

[7]https://community.openai.com/t/what-is-the-difference-between-putting-the-ai-personality-in-sys
194938

As such, we tried two approaches; starting the prompt with "You are an expert in language use." or "You are an expert in idiomatic language.". However, we find that neither of these approaches lead to improved performance. Interestingly, replacing the word "Literal" with "Compositional" did seem to have a positive effect. We found that removing the instruction to explicitly return only one letter ('i' or 'l') led the model to occasionally return other outputs, which causes a drop in performance (as we treat such responses as invalid). For the English subset, this is the case for 3% of outputs (28 out of 916 examples).

Language Prompts

Since SemEval has test data in English, Portuguese, and Galician, we experiment with a) explicitly stating the language of the sentence in the prompt, and b) translating the prompt using a commercial machine translation tool. We perform this analysis for GPT-3.5-turbo, Gemini 1.0 Pro, and Flan-T5-XXL, with results shown in Table 5.5.

	GPT-3.5-turbo		Gemini 1.0		Flan-T5-XXL	
	PT	GL	PT	GL	PT	GL
Default	0.553	0.587	0.582	0.604	0.464	0.411
Language Prompt	0.554	0.604	0.561	0.640	0.479	0.457
Translated	0.541	0.512	0.549	0.665	0.573	0.477

Table 5.5: GPT 3.5-turbo, Gemini 1.0, and Flan-T5-XXL results for Portuguese and Galician on SemEval using multilingual prompts.

For Gemini 1.0 Pro and Flan-T5-XXL we see performance improvement for Galician under both of these approaches, with higher performance when translating the prompt. We hypothesise that both English and Portuguese are likely well-represented in the model training data, and LLMs in general work well in multilingual settings (Shi et al., 2022). However, Galician is likely to be both rare and potentially confused with Portuguese when the language is not specified, or when there is less text in that language available in the prompt. It would be interesting to experiment further with similar language pairs.

Not shown here is that we recorded reduced performance for English across all three models when specifying the language in the prompt (0.739 to 0.674 for GPT-3.5-turbo, 0.771 to 0.732 for Gemini 1.0 Pro, 0.716 to 0.706 for Flan-T5-XXL). It is possible that additional prompt tokens specifying the language may act as a 'distractor' when it is the *de facto* default, and the nature of the generative models means that we can anticipate variation in responses to identical prompts.

5.4.4 Few-shot Prompting

The "one-shot" setting of SemEval 2022 Task 2a (in which further examples of the target PIE in context are made available) allows for the investigation of passing examples to the model through the prompt. We thus experiment with doing so for GPT-3.5-turbo, Gemini 1.0 and Flan-T5-XXL. We try two configurations: passing one example per PIE (one-shot), and passing all the examples that are available in the dataset (few-shot)[8]. These results are shown in Table 5.6.

Model	Setting	EN	PT	GL	All
Gemini Pro 1.0	Zero-shot	**0.766**	0.590	0.600	0.672
	One-shot	0.706	0.625	0.711	0.688
	Few-shot	0.685	**0.642**	**0.745**	**0.693**
GPT-3.5-turbo	Zero-shot	**0.739**	**0.563**	**0.579**	**0.645**
	One-shot	0.645	0.542	0.553	0.594
	Few-shot	0.686	0.545	0.566	0.614
Flan-T5-XXL	Zeroshot	0.629	0.464	0.411	0.514
	Oneshot	0.810	0.665	0.732	0.749
	Fewshot	**0.845**	**0.713**	**0.828**	**0.805**
Best	One-shot	0.964	0.894	0.937	0.939

Table 5.6: Results on SemEval using few-shot prompting. *Best* results are taken from Chu et al. (2022), which uses a fine-tuned XLM-RoBERTa model.

Interestingly, the impact of few-shot prompting varies across the models. Flan-T5-XXL benefits the most from this, with stark and consistent performance improvements across the three settings and across all three languages – the overall macro F1 jumps from 0.580 in the Zero Shot setting to 0.805 in the Few Shot setting.

Further to this we analyse the performance of all size Flan-T5 models, and present a heatmap illustrating the impacts on performance stemming from zero-shot and few-shot scenarios in Table 5.7.

The smallest models benefited the most from seeing one or more examples before inference. In the best cases, performance in English improved by 0.432 in the one-shot setting and 0.516 in the few-shot setting. Interestingly, few-shot prompting can be seen

[8]Where available, the one-shot training data has one idiomatic example for each PIE, and one literal example. However, for some PIEs just one of these is present.

	Small	Base	Large	XL	XXL
Oneshot (EN)	0.432	0.079	0.199	0.348	0.182
Oneshot (PT)	0.388	0.011	0.227	0.228	0.202
Oneshot (GL)	0.526	0.049	0.053	0.185	0.321
Oneshot (ALL)	0.443	0.054	0.162	0.264	0.235
Fewshot (EN)	0.516	-0.003	0.332	0.404	0.217
Fewshot (PT)	0.391	0.000	0.093	0.285	0.249
Fewshot (GL)	0.576	0.000	0.137	0.354	0.417
Fewshot (ALL)	0.489	-0.001	0.227	0.352	0.291

Table 5.7: Enhancements in macro F1 scores (positive values) and declines (negative values) when compared to the performance in zero-shot conditions across all Flan-T5 models.

to improve performance across Portuguese and Galician examples in all model settings, apart from T5-FLAN-Base and Large where there is little, or no improvement. It appears that Flan-T5-Base seems to be least improved by prompting with examples, with a negative effect on performance in few-shot prompting settings. In the one-shot setting, improvement in model performance is minor. The Large, XL and XXL models also benefited from one- and few-shot prompting, with Flan-T5-XL seeing the most performance enhancement. It appears that whilst models follow "bigger is better" in zero-shot settings, they do not necessarily follow this pattern under one/few-shot prompting. In fact, the best performance in the few-shot setting is with T5-Small, which at only 80M parameters achieves an overall macro F1 of 0.821, the best performance of any of the models we have evaluated in this paper. This is in significant contrast to performance on MAGPIE and FLUTE, where zero-shot performance is very low. The model is likely learning some artefacts from the data such as predicting only one label for a given PIE in the SemEval dataset.

Gemini 1.0 Pro also achieves consistent (though smaller) performance improvements from Zero Shot to One Shot to Few Shot, but the performance for English reverses this pattern. We also see a big jump in performance between Zero Shot and One Shot for Galician, which we again attribute to the rarity of this language and its similarity with Portuguese.

GPT-3.5-turbo is hindered by providing examples. The reasons for this are unclear, but this may be linked to the inability shown by GPT-3.5 to follow system prompts. If the model is not successfully following longer prompts then they may effectively introduce noise and lead to worse performance, as we saw when comparing results with and without system prompts.

5.5 Discussion

5.5.1 Task Labelling

The majority of the models we examined achieved high performance on the FLUTE dataset. We attribute this to the nature of FLUTE's evaluation being distinct from MAGPIE and SemEval. For the latter two, the model is asked to label 'idiomatic' or 'literal' use of a given idiom, whereas, in the FLUTE STS task, the model is required to pick out the contradiction or entailment relationship between two sentences.

This means that a model might not necessarily require 'knowledge' of the target idiom to succeed, but could determine the relationship between the two sentences from other information, as facilitated by contextualised embeddings (Boisson et al., 2023). Moreover, the model is likely to have encountered similar tasks during its pre-training. Flan-T5 models are instruction-refined versions of T5 (Raffel et al., 2020; Chung et al., 2024), that have undergone exposure to over 1000 tasks during its fine-tuning process alone. Among these tasks are evaluations of entailment and contradiction judgments, akin to FLUTE, such as SNLI (Bowman et al., 2015), MNLI (Williams et al., 2018), CB (de Marneffe et al., 2019) and numerous other reasoning tasks (for details see Raffel et al., 2020; Chung et al., 2024).

5.5.2 Practicalities

In contrast with fine-tuned classification models, as prompted models are capable of open-ended generation, they may not output a response in the format requested. While the output may be readily interpretable by a human reader, this is not practical when evaluating large numbers of responses. Prompting for specific formats is easier for models which have undergone more instruction tuning (Ouyang et al., 2022; Rafailov et al., 2023), and is a key reason why the Mistral-7B model outperforms the Llama2 7B variant.

Prompted, generative models produce outputs which are subject to variation when they are repeatedly given the same prompt. While the user may have some control over this behaviour through 'temperature' parameters, this variability is inherent to generative models. When converting the outputs of such models to a labelling decision, this variability will also affect the results.

Despite their generally higher performance than the local models and their advantages when it comes to prototyping, there are a number of considerations specific to SaaS models which may be significant. These include:

1. Cost – The larger models have a higher per-1000-tokens cost, which may lead to some evaluations being cost-prohibitive. Evaluating GPT-4 on the (relatively small) SemEval test set, for example, costs $11. Running evaluation on this model, especially across multiple runs for prompt tuning, etc. may potentially price out researchers with lower budgets.

2. Safety Features – Commercial SaaS models frequently include features designed to limit models and users' capability to process or generate content which may cause harm. These features may also impact on researchers' ability to use the tools, as they produce what are effectively false positives. For example, when using the VertexAI API for experiments with Gemini Pro, the API consistently refused to generate responses for a small number of prompts. These included certain contexts for the expression *street girl* which referred to prostitution or sexualization, but also the FLUTE sentence pair "Your brother is mature and behaves in an adult manner. Your brother is a big baby." for the expression *to be a big baby*[9]. We treat any such responses as incorrect in our statistics.

3. Service Changes – Changes to the underlying model can be made by the third party at any time, and can significantly impact the performance of the models and the consistency of results. Whilst undertaking this work the default gpt-3.5-turbo model changed from one released in June 2023, to one released in January 2024.

4. Rate limits – For larger datasets, the rate limits of commercial APIs can become an issue. As it is still not fully released, for a significant amount of time during the creation of this work, the daily rate limit for GPT-4-turbo was lower than the number of tokens in MAGPIE, which prevented us from completing any evaluation runs for this model and dataset combination.

[9] Replacing the word 'adult' with 'grown-up' convinced the service to generate a response.

5.6 Conclusion

In this work we have evaluated the performance of various large language models on three idiomaticity datasets (SemEval 2022 Task 2a, FLUTE, and MAGPIE). We have investigated locally-run models up to 13B parameters, as well as significantly larger models (GPT-3.5, GPT-4, and Gemini 1.0 Pro) accessed through commercial APIs. We perform an extensive analysis of the impact of several factors on performance; model size, prompt engineering and few-shot prompting. In addition, we discuss considerations for practitioners wishing to use these models in their own work, with emphasis on cost and practicalities such as the variability of outputs and the impacts of decisions made by the companies operating these services. Our overall findings are as follows: 1) LLMs at the highest scale are able to achieve competitive results for idiomaticity detection, and performance on FLUTE in particular seems to have saturated, but these general models do not match the performance of (much-smaller) encoder models fine-tuned for the specific idiomaticity detection tasks of SemEval and MAGPIE. 2) The performance of prompted, generative LLMs seems to scale consistently with parameter count for these datasets, indicating the potential of even bigger models to achieve further increases in performance. 3) While they are based on a relatively small set of examples, our experiments with multilingual models suggest that performance gains can be obtained by specifying the target language, translating prompts and by providing examples. However, the efficacy of these modifications depends on the model used and the language in question; they appear to harm performance for English (which is, presumably, the most-represented language in the model training regimens) while producing the largest benefit for the much rarer Galician.

Chapter 6

Publication IV: Stands to Reason: Investigating the Effect of Reasoning on Idiomaticity Detection

Dylan Phelps[1,2], Maggie Mi[2], Rodrigo Wilkens[3], Edward Gow-Smith[2], Thomas Pickard[2], Aline Villavicencio[2,3]

1 Healthy Lifespan Institute, The University of Sheffield

2 School of Computer Science, The University of Sheffield

3 Institute of Data Science and Artificial Intelligence, The University of Exeter

Abstract

The recent trend towards utilisation of reasoning models has improved the performance of Large Language Models (LLMs) across many tasks which involve logical steps. One linguistic task that could benefit from this framing is idiomaticity detection, as a potentially idiomatic expression must first be understood before it can be disambiguated and serves as a basis for reasoning. In this paper, we explore how reasoning capabilities in LLMs affect idiomaticity detection performance and examine the effect of model size. We evaluate, as open source representative models, the suite of DeepSeek-R1 distillation models ranging from 1.5B to 70B parameters across four idiomaticity detection datasets. We find the effect of reasoning to be smaller and more varied than expected. For smaller models, producing chain-of-thought (CoT) reasoning increases performance from Math-tuned intermediate models, but not to the levels of the base models,

whereas larger models (14B, 32B, and 70B) show modest improvements. Our in-depth analyses reveal that larger models demonstrate good understanding of idiomaticity, successfully producing accurate definitions of expressions, while smaller models often fail to output the actual meaning. For this reason, we also experiment with providing definitions in the prompts of smaller models, which we show can improve performance in some cases.

6.1 Introduction

Large Language Models (LLMs) have been shown to excel at many tasks across many disciplines (Ouyang et al., 2022; OpenAI et al., 2024a; Grattafiori et al., 2024), including on tasks involving idiomatic expressions, such as idiomaticity detection and multiword expression identification (Phelps et al., 2024; Smădu et al., 2024). More recently, reasoning models – LLMs which generate a chain of thought (CoT) 'reasoning' responses before reaching a final answer (Wei et al., 2022b) – have been shown to outperform traditional LLMs in a range of tasks (OpenAI, 2025a). Notably, DeepSeek-R1 and smaller 'distilled' models trained on data produced by it (DeepSeek-AI et al., 2025), have recently matched and exceeded other reasoning models such as OpenAI's GPT-o1 whilst being open source and offered at a lower price.

The motivation for such reasoning models is that training them to generate outputs in chain-of-thought format allows them "think" step-by-step, working out a final answer incrementally. This output format also allows for potentially higher explainability, since the CoT becomes part of the context used to make the final classification. However, some work has shown that CoT explanations can be unfaithful, misrepresenting the reason for a model's prediction (Turpin et al., 2023; Lyu et al., 2023).

In the field of computational idiom processing, reasoning models are of interest as one could imagine such step-by-step thinking to improve the handling of potentially idiomatic expressions (PIEs), where the meaning is ambiguous and determined by context. Using knowledge of a definition as an indicator of a level of understanding, the CoT also allows analysis of A) how well models can define PIEs in their literal or idiomatic senses, and B) how well models can use subsequent reasoning to work out whether a PIE is used idiomatically or literally in a given context.

Therefore, in this paper, we explore how the distilled DeepSeek-R1 models perform on a range of idiomaticity detection tasks. Additionally, we evaluate the reasoning outputs of each model to further explore how they represent and understand idiomaticity. Our research questions are:

1. Does the production of reasoning chains improve the ability of models to detect idiomaticity, and how does this vary across model scale?
2. Do the reasoning outputs (CoT) reflect understanding of the target idiomatic expressions?
3. Can the definitions produced by the larger models be used as an additional knowledge source, improving the performance of the smaller models?

To this end, we find that although the performance of the models improves as the models scale, the effect of adding reasoning generation is smaller and more varied. Particularly, for the smaller models we find that producing CoTs on average reduces the performance across our datasets, whereas for the larger models a small increase in performance can be seen in the reasoning variants. Our manual analysis of the dataset shows that the larger models also have a good understanding of idiomaticity and can reliably produce accurate definitions of the given expressions, while the smaller models often fail to do so. However, even the largest models are hindered by their ability to use context and reasoning to disambiguate PIEs.

Our experiments using definitions from the larger models as a knowledge source show that performance of the smallest models improves by an average of 0.069 macro F1 on FLUTE, but does not affect performance significantly for DICE. These results suggest that this methodology has potential as a knowledge distillation technique for certain tasks.

The paper is structured as follows: §6.2 and §6.3 present the methodology, §6.4 and §6.5 the results, and it finishes with conclusions.

6.2 Methodology

In this section, we introduce the datasets, models, and approach that we use in this work.

6.2.1 Datasets

To enable us to compare the performance of the newer models to those evaluated in Phelps et al., 2024 we evaluate on the same datasets. We also include the recently released DICE dataset (Mi et al., 2025). We provide brief descriptions of each dataset.

SemEval 2022 Task 2a (Tayyar Madabushi et al., 2022) is a binary classification task for detecting idiomaticity of noun compounds within context sentences, with examples in English, Portuguese, and Galician. The test set contains 150 PIEs split equally across the three languages, with a total of 2342 examples. To maximize the number of test examples, we follow Phelps et al., 2024 by combining the few-shot and zero-shot test sets. However, when evaluating we don't provide few-shot examples for any of the instances.

Figurative Language Understanding through Textual Explanations (FLUTE; Chakrabarty et al., 2022b) presents an English-language Natural Language Inference (NLI) task in which models should predict whether a premise, containing a figurative expression, follows from a hypothesis containing a correct or incorrect paraphrase. We evaluate only the idiom subset of the dataset, which contains a test set of 250 examples with 69 idioms represented.

MAGPIE (Haagsma et al., 2020) is a multi-class idiomaticity detection dataset where a large number of potentially idiomatic expressions in context must be classified as idiomatic, literal, or other with a split of 70/29/1. The entire dataset is in English, with 1134 expressions represented across 4840 examples.

Dataset for Idiomatic Contrastive Evaluation (DICE; Mi et al., 2025) aims to assess the ability of models to use context in idiomaticity detection. Existing datasets fail to assess the role of context in idiom interpretation, as literal meanings often stem from grammatical changes to the form of the idiomatic expression, which allows models to rely on surface cues as a reasoning shortcut instead of true comprehension. To avoid this, DICE contains 2066 sentences that are balanced for sense, in which the form of the expression is kept the same across both literal and figurative uses. Being a newer dataset, the test labels were not publically available when the DeepSeek models were trained, so there is no chance of contamination.

6.2.2 Models and Experiments

The models we evaluate are the suite of DeepSeek-R1 distillation (DeepSeek-AI et al., 2025) models ranging from 1.5-70B parameters, that have been fine-tuned on a reasoning dataset collected from the larger DeepSeek-R1 model (the **reasoning** models). Whilst processing the input, these models produce chain-of-thought reasoning.

Each of the distilled models is a fine-tuned version of another open source model (the **base** models), and so to compare the performance with and without CoT reasoning, we also run our evaluations on the base models. Namely, we evaluate the Qwen2.5 suite of models (1.5B, 7B, 14B, and 32B parameters) (Qwen, 2024; Qwen et al., 2025) and their reasoning-tuned DeepSeek-R1 versions, and Llama3.3-70B (Grattafiori et al., 2024) alongside its reasoning-tuned variant.

The DeepSeek-R1 1.5B and 7B parameter variants follow a different training pipeline. Rather than direct reasoning-tuning from base Qwen2.5 models. These variants first undergo intermediate training on approximately 1 trillion tokens of mathematical data, with prompting designed to elicit reasoning behaviors . This produces three model variants for each size: the original base model (Qwen2.5), the math-specialized intermediate model (Qwen2.5-Math), and the final reasoning-tuned model (DeepSeek-R1 Qwen). We evaluate all three variants to understand the training progression's impact on capabilities. While we anticipate that math-specialization may compromise general domain performance, we include these intermediate models in our evaluation to examine whether subsequent reasoning-tuning with broader domain data can recover the lost general capabilities.

All models are run using the vLLM library (Kwon et al., 2023) utilizing Q6_K_M quantizations of each model (Frantar et al., 2023). This allows all the models to be run on a single A100 80GB GPU.

For each dataset, we prompt the models with the prompts given in Section A. The reasoning models generate a CoT before the output, which we split off for analysis. Where the models do not output the required label in an easily parseable format, we use GPT-4o (OpenAI, 2024) to extract the returned label (using the prompt also given in Section A). We evaluate the models using macro F1 for all datasets.

6.3 Model Performance

We present the results from our experiments in Table 6.1. As expected, we see the larger models achieving the highest performance on the idiomaticity detection tasks. The best performing model is DeepSeek-R1 Qwen-32B, which slightly outperforms the larger DeepSeek-R1 Llama-70B on all of the datasets. The only model that surpasses comparable larger models is the base version of Qwen2.5-7b, which outperforms the larger non-reasoning models on FLUTE, and, surprisingly, gets the highest score on SemEval-2022, being the only model to score over 0.7.

	Model	Flute	SemEval	MAGPIE	DICE
Base	Qwen2.5-1.5B	0.849	0.458	0.430	0.366
	Qwen2.5-7B	0.921	**0.737**	0.786	0.710
	Qwen2.5-14B	0.924	0.586	0.823	0.800
	Qwen2.5-32B	0.914	0.612	0.888	**0.873**
	Llama-70B	0.921	0.658	0.778	0.816
Math	Qwen2.5-Math-1.5B	0.551	0.484	0.495	0.482
	Qwen2.5-Math-7B	0.691	0.404	0.482	0.507
Reasoning	DeepSeek-R1 Qwen-1.5B	0.577	0.533	0.516	0.499
	DeepSeek-R1 Qwen-7B	0.812	0.585	0.626	0.462
	DeepSeek-R1 Qwen-14B	0.929	0.573	0.863	0.863
	DeepSeek-R1 Qwen-32B	**0.948**	0.641	**0.890**	**0.866**
	DeepSeek-R1 Llama-70B	**0.947**	0.628	0.873	0.857

Table 6.1: Results of both the base and the reasoning models on the four datasets. Reported is the mean macro F1 averaged across 5 runs, using different random seeds. The best score(s) per dataset shown in bold.

In relation to the first question, if reasoning improves the ability of models to detect idiomaticity, the results observed paint a mostly positive picture. In Table 6.2 we show the difference between the base and reasoning variants of each model across the datasets, as well as the difference between the base and intermediate math models: Qwen2.5-Math-1b and Qwen2.5-Math-7b.

Math-tuning produces the large drops in performance we expected for both the Qwen2.5-1b and Qwen2.5-7b models. This is slightly less pronounced on the 1.5b model, though this primarily reflects floor effects as performance on SemEval, MAGPIE, and DICE was already near-random levels, leaving little room for further deterioration. Conversely, the reasoning tuning has a large positive effect when moving from the Math variants to the DeepSeek variants, with overall positive improvements of

	Model	FLUTE	SemEval				MAGPIE	DICE	**Mean Diff.**
			ALL	EN	PT	GL			
Base	1.5b	-0.272	0.075	0.075	0.068	0.082	0.086	0.132	0.006
Base	7b	-0.109	-0.153	-0.107	-0.138	-0.077	-0.160	-0.247	-0.167
Base	14b	0.005	-0.014	-0.015	0.051	0.011	0.040	0.064	0.024
Base	32b	0.034	0.039	0.039	0.081	-0.010	0.002	-0.007	0.017
Base	70b	0.026	-0.021	-0.066	-0.085	-0.019	0.095	0.042	0.035
Math	1.5b	0.026	0.049	0.050	0.050	0.057	0.021	0.017	0.028
Math	7b	0.121	0.181	0.212	0.231	0.206	0.115	-0.044	0.093

Table 6.2: Difference in performance of the reasoning models compared to 'Base' and 'Math' variants, in absolute difference in macro F1. A negative value means reasoning hurts performance, whilst a positive value means it improves performance. The "Mean Diff." column reports the average improvement from using a reasoning model across the four datasets.

0.028 and 0.093, an effect that can especially be seen through single task improvements for the 7b parameter model (0.121, 0.181, 0.115).

Larger models (14B, 32B, and 70B) demonstrate consistent benefits from CoT reasoning integration, with performance improvements reaching 0.095 for Llama-70B on MAGPIE. However, results vary in magnitude across tasks, and some performance decreases occur on the SemEval dataset—particularly for Portuguese and Galician languages, where the effect of reasoning is less clear.

6.4 Manual Error Analysis

As an initial step in exploring the understanding of the models, we manually inspect samples of the reasoning outputs and evaluate them. We split the evaluation into two facets: whether the model understands the relevant PIE (whether it generates a valid idiomatic definition, either wholly or in parts throughout the CoT), and whether the reasoning is valid (whether the model successfully disambiguates the PIE in context within its CoT). This categorization allows us to separate the effects of the model's understanding of the expression and context and its reasoning ability.

6.4.1 Labelling Setup

Three of the authors, acting as annotators, each annotate the same 30 responses (15 correct and 15 incorrect) for each reasoning model. We initially chose responses to the

MAGPIE dataset as this is the largest dataset, with the largest variance in expressions used. However, whilst labelling we found the dataset to be quite noisy (mislabelling, example sentences which do not actually contain the target expression, etc.) and so we additionally label examples from DICE. Overall, 300 examples were labelled 3 times (once by each annotator) for understanding and reasoning ability on a 5-point scale. Each point on the scale was awarded a score, which is averaged across annotators and examples, to allow us to get a representative score for the model. The full set of labels and scores can be seen in Section B.

For understanding, labels range from No definition, Focus on one word, and Inaccurate definition (all scored 0 for lacking accurate information) to Partial definition (scored 2) and Accurate definition (scored 3). Reasoning uses a similar scale: Nonsensical reasoning (0), OK reasoning that partially attends to expression and context (2), and Good reasoning with mostly correct connections (3). We label meta-reasoning, which occurs (rarely) when the model doesn't correctly identify the expression in the context, instead relying on meta-knowledge typical expression usage. We use 0-2-3 scoring instead of 1-5 to emphasize the gap between inaccurate and accurate responses, with some labels being assigned the same score as they identify different scenarios of similar quality.

6.4.2 Labelling Findings

Mislabelling in the dataset

The first thing we observe whilst manually labelling is that, of the examples where the model is judged to be incorrect, there are many labels which are either incorrect or ambiguous. In Table 6.3, we show the proportion of examples where the gold label and prediction do not match, for which our annotators agreed that the gold label was either incorrect or ambiguous, and in Table 6.4 show some selected examples from MAGPIE. There is some overlap between the examples, as he idiomatic subset of DICE (which accounts for most of these examples) is taken from MAGPIE. For the larger models, this effect is particularly pronounced, with a large proportion of their "incorrect" predictions actually being judged valid by human annotators. This indicates that accuracy for the larger models may be higher than suggested by the results in Table 6.1, which further reinforces the high level of understanding shown by the larger models. We emphasise that this does not imply a large proportion of

	% Judged Valid	
	DICE	MAGPIE
DeepSeek-R1 Qwen-1.5B	0%	0%
DeepSeek-R1 Qwen-7B	0%	6.7%
DeepSeek-R1 Qwen-14B	40%	60%
DeepSeek-R1 Qwen-32B	46.7%	33.3%
DeepSeek-R1 Llama-70B	40%	33.3%

Table 6.3: For each model, for DICE and MAPGIE, the percentage of sampled "incorrect" predictions (where the model prediction does not match the gold label), where the model prediction was judged to be valid by the annotators, due to a mislabelled or ambiguous example.

these datasets are incorrectly labelled or ambiguous, instead the accuracy of the large models means that when they are incorrect they have a high chance of picking out the few examples that are ambiguous.

Gold Label	Mislabel Type	Example
Idiomatic	Incorrect	When I turn **in the saddle** and point it out to him with vigorous gestures , his curiosity is aroused.
Idiomatic	Incorrect	Vauxcelles refers to some of the painters of the Indépendants of 1910 who were working **under the influence** of Cézanne as ‘ ignorant geometricians , who reduce scenery and the human body to dull cubes’ .
Idiomatic	Ambiguous	The best way is to calculate the **break-even** inflation rate . This is the inflation rate that makes the money yield on an index - linked gilt equal to the yield to maturity on a conventional gilt of the same maturity .

Table 6.4: Examples of incorrect and ambiguous labels found during manual analysis of the MAGPIE datasets. The relevant expression is highlighted in bold for each example.

Error Patterns by Model Size

Figure 6.1 shows the manual label scores for each model in the case where the model is correct and incorrect. We observe that mistakes made by different model sizes often fall into different categories of errors. The larger models appear to have an understanding of most of the expressions in the data, and therefore as part of their reasoning produce a definition of the idiom. The smaller models, however, often produce an incorrect

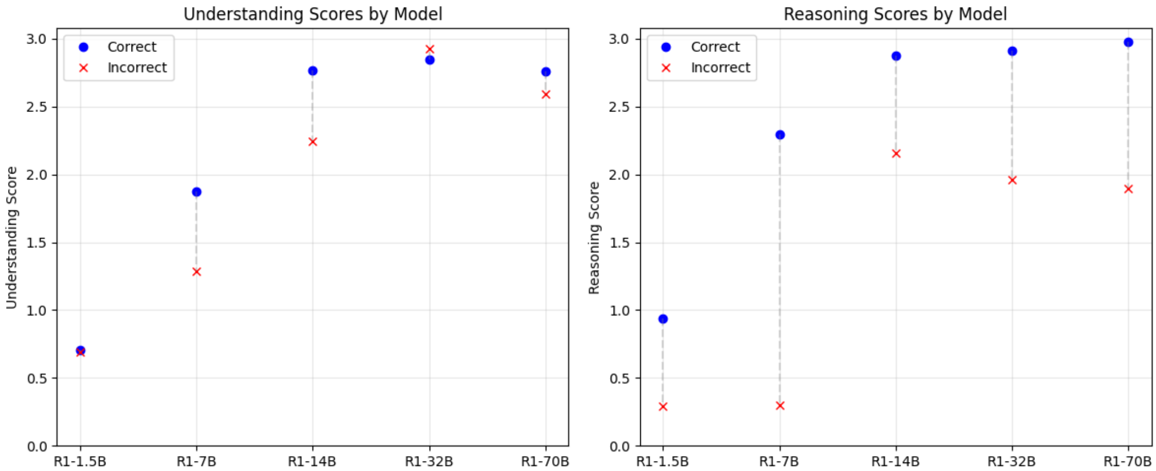


Figure 6.1: The average manually labelled reasoning and understanding scores (between 0 and 3) for each model, for 15 correct and 15 incorrect examples, averaged across DICE and MAGPIE.

definition or struggle to identify or recognize the expression all together. This then leaves a poor basis for the rest of the reasoning, leaving the label as almost random.

The 1.5B parameter model shows a low level of both understanding and reasoning ability, with most of its correct outputs being due to chance. The 7B parameter model has on average higher levels of both reasoning and understanding, which correlates with the results in Table 6.1. The main factor to whether the model is correct or incorrect appears to be the reasoning performance, as this varies from close to the larger models when correct, but closer to the 1.5B model when incorrect. The larger models understanding of the expressions is generally high when it is both correct and incorrect, with more variation being identified in the reasoning capabilities. However, the overall reasoning scores are still relatively high, which may indicate that they are incorrect on the more nuanced examples, something we noted whilst labelling.

6.5 Chain-of-Thought Length

We investigate the potential relationship between the Chain of Thought (CoT) size and model accuracy. The DeepSeek paper showed that as model training progressed and model performance improved, average CoT length went up (DeepSeek-AI et al., 2025). To explore the correlation between reasoning length and performance in idiomaticity detection, we split the CoTs into two categories for incorrect and correct predictions, and then tokenise them (using spaCy + en_core_web_sm) to give lengths. We show

Model	Flute	SemEval			MAGPIE	DICE
		EN	PT	GL		
DeepSeek 1.5B	-0.038	0.009	-0.037	-0.061	0.000	-0.011
DeepSeek 7B	**-0.151**	**-0.075**	-0.001	**-0.089**	**-0.057**	-0.006
DeepSeek 14B	-0.022	0.020	-0.036	0.000	**-0.135**	**-0.177**
DeepSeek 32B	**-0.246**	-0.003	-0.064	**0.126**	**-0.119**	**-0.147**
DeepSeek 70B	**-0.305**	-0.036	-0.002	0.028	**-0.132**	**-0.156**

Table 6.5: Correlation between CoT length and model correctness. Values in bold indicate a p-value < 0.05.

two histograms of CoT length for correct and incorrect predictions in Figure 6.2. We can see that the CoT length distribution appears similar across both cases. Next, we calculate the point-biserial correlation between CoT size and prediction correctness for each model size category (1.5B, 7B, 14B, 32B and 70B), which can be seen in Table 6.5. Additionally, we compute the pseudo-R^2 for a logistic regression model that predicts accuracy based on CoT length.

The results reveal that, across various datasets, there is no consistent pattern suggesting a significant relationship between CoT size and prediction accuracy. For instance, the most considerable observed correlation (-30.50%) in the EN-FLUTE corpus corresponds to a modest R^2 value of 0.11, meaning only 11% of the variance in prediction outcomes is explained by CoT size. While models of larger sizes (32B and 70B) show slight improvements in R^2 for certain corpora, such as FLUTE and DICE, the overall findings suggest no clear trend linking CoT length to predictive success.

6.6 Definition Generation as Distillation

The results from our manual labelling reveal that the 32B and 70B parameter models can accurately produce definitions of the expressions represented in the dataset. Given the lack of comprehensive resources for idiomatic expression definitions, this finding presents an opportunity to use model-generated definitions as a novel distillation technique.

As an initial exploration, we investigate whether providing the model generated definitions from the larger models to the smaller reasoning models can improve performance. In our manual analysis, we found that the 1.5B and 7B parameter reasoning

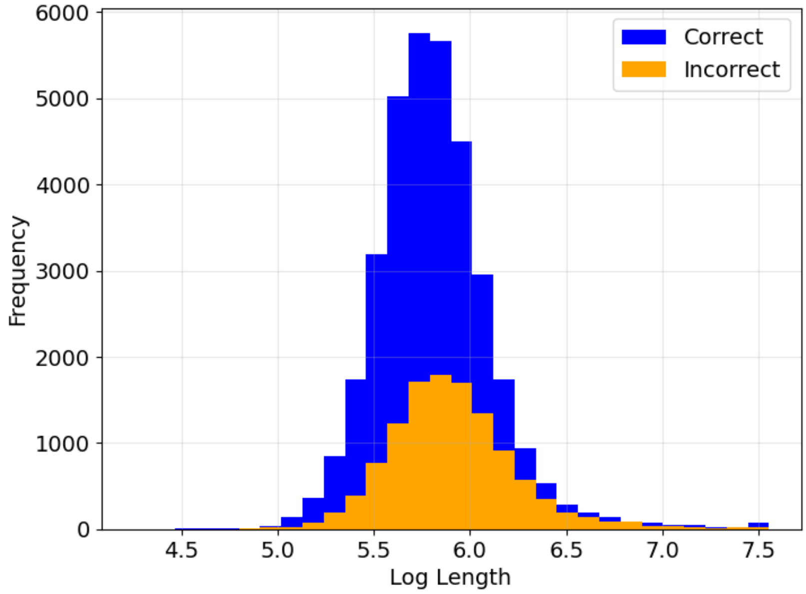


Figure 6.2: The distribution of reasoning output log lengths for correct (blue) vs. incorrect (orange) model predictions, aggregated across all models and datasets.

models were not able to accurately reproduce definitions, which may suggest that they do not contain the required knowledge to do so, partly explaining their poor performance on the detection tasks. By providing this information in the prompt, this knowledge is able to be used by the models whilst making predictions.

For this experiment, we use DeepSeek-R1 32B to generate definitions for all the expressions in the FLUTE and DICE datasets (examples shown in Table 6.6). We select DeepSeek-R1 32B as this shows the highest overall 'understanding' score in the manual analysis (see Figure 6.1). We then append the definition to the prompt used in our main experiments and rerun the evaluation for the 1.5B, 7B, and 14B parameter reasoning models. These models represent a range of 'understanding' scores, allowing us to analyse the effectiveness of the definitions across different performance and 'understanding' levels. We expect that the models with the lowest performance on the detection tasks and 'understanding', e.g. the 1.5B and 7B parameter models, will benefit the most from the additional information in the prompt.

6.6.1 Results

The results from the experiment (shown in Table 6.7) show that adding definitions to the prompt significantly increases performance for the 1.5B and 7B parameter models on the FLUTE dataset. However, for DICE, and for the 14B parameter model on

PIE	Generated Definition
all hell broke loose	"All hell broke loose" means that a situation suddenly became extremely chaotic, noisy, or violent.
off the hook	To be off the hook means to be free from a responsibility, obligation, or troublesome situation, often after being expected to handle it.
against the grain	To go against the grain is to act in a way that is contrary to the usual approach, often resulting in resistance.
play with fire	'Play with fire' means to engage in a risky or dangerous situation that could lead to negative consequences.
make a killing	To make a killing means to achieve great success or earn a lot of money.

Table 6.6: Examples of the definitions generated by DeepSeek-R1 32B for given PIEs from the DICE dataset.

Model	Original Prompt		Definition Prompt	
	FLUTE	DICE	FLUTE	DICE
DeepSeek-R1 Qwen-1.5B	0.577	0.499	**0.663**	0.487
DeepSeek-R1 Qwen-7B	0.812	0.462	**0.864**	0.461
DeepSeek-R1 Qwen-14B	0.929	0.863	0.908	0.842

Table 6.7: The results using both the original and definition appended prompt, when evaluating the smaller models on FLUTE and DICE. We report the mean macro F1 across 5 runs with different random seeds, and bold any significantly improved results for each model/dataset pair.

both datasets, no significant difference can be seen between adding the definition to the prompt or not.

We suggest that the difference in effectiveness on FLUTE and DICE comes from the fact that, as suggested by Mi et al. (2025), DICE requires models to attend to the surrounding context to make decisions. Adding the definition will not improve the models' ability to do this, and so no increase in performance is seen.

However, the lack of any decrease in performance is promising for the technique as it means that adding definitions to the prompt can safely be done without risk of reducing performance. While the 14B parameter model shows slightly larger absolute decreases on both datasets, these differences are not statistically significant.

6.7 Conclusion

In this study, we have investigated the effect that reasoning has on the idiomaticity detection ability of LLMs. On four idiomaticity datasets (Flute, SemEval, MAGPIE, DICE), we have evaluated the performance of the distilled DeepSeek-R1 models across a range of sizes, in addition to their corresponding non-reasoning base models, and intermediate Math variants for the small parameter count models. To further explore the results, we have performed both a manual and a quantitative analysis on the reasoning outputs of the models in order to gain some insight into the models' understanding of idiomaticity, as well as their ability to reason with the expression in context. As a result, we have three main findings:

1. The effect of reasoning on idiomaticity detection is relatively small, although it varies across model size and per dataset. For the smaller models, training on math specific reasoning data greatly reduces performance, but subsequent training on more general domain reasoning data can restore performance. Larger models show slight increases in performance when trained on general domain reasoning data.

2. Our manual analysis shows the larger models are consistently able to produce high quality definitions of the expressions, something which the smaller models struggle to do. However, for all models the quality of the reasoning is the most prominent factor which decides whether a model will be correct or not. This highlights that even the largest models struggle to use the expression and the context to disambiguate the meaning.

3. Our further experiments utilising the definitions produced by the larger models as knowledge sources for smaller models show promise as a knowledge distillation technique. Appending the definitions to the prompt improves the performance in some cases, whilst having no significant effect in others, implying this technique can be applied generally without risk of regression.

Our results show that while CoT reasoning offers small benefits for idiomaticity detection, the modest improvements suggest that current reasoning approaches may not fully address the underlying contextual disambiguation challenges inherent in idiomaticity detection. Even if the reasoning generated appears to indicate idiomatic understanding, further examination indicates otherwise. Future work should explore alternative reasoning frameworks specifically designed for figurative language processing, perhaps incorporating more structured approaches to context interpretation. Additionally, it could explore other ways to utilise the high-quality definitions generated by the models, either through in-context learning or as fine-tuning data.

A Prompts

For the FLUTE dataset, we use the following prompt:

"You will be given two sentences, a premise and a hypothesis. Respond with either 'entailment' if the hypothesis follows from the premise, or 'contradiction' otherwise."

For the other datasets, we use:

"Predict whether the MWE given in the sentence being used idiomatically or literally. Respond 'idiomatic' or 'literal' respectively."

For using GPT-4o to extract the label, we use:

"Given the following model output, predict what label the model was trying to assign out of {option_string}. Only output either the word {option_string}"

When adding the definition as part of the prompt, we append:

"Definition: {definition}"

B Error Analysis Labelling

Type	Label	Score
Understanding	No definition	0
	Focus on single word	0
	Inaccurate definition	0
	Partial definition	2
	Accurate definition	3
Reasoning	Nonsensical reasoning	0
	OK reasoning	2
	Good reasoning	3
	Meta-reasoning (inaccurate)	2
	Meta-reasoning (accurate)	3

Table 6.8: The labels used for the manual labelling of the models CoT, with the scores assigned to each label when calculating averages.

C Limitations

Whilst we do on average see a small improvement in performance for idiomaticity detection in the larger models, this result isn't wholly consistent, as the reasoning variant of Llama-3.3 70B performs worse by 0.037 macro F1 on the English subset of SemEval compared to the non-reasoning variant. The average performance increasing from reasoning is also lower for this model than Qwen2.5 14B and Qwen2.5 32B (0.013 compared to 0.039). Further work could investigate a reason for this inconsistency.

On our human analysis of the model CoTs, we break the evaluation into "understanding", which we take as the ability of a model to define the relevant PIE in its idiomatic sense, and "reasoning", which we take as the ability of a model to use the context and the PIE to determine whether it's literal or idiomatic. This provides a useful analysis, but a more fine-grained evaluation using other features would perhaps be more useful, for example looking at whether a model is logically consistent, whether it's hallucinating, etc.

Chapter 7

Publication V: Beyond surface form: A pipeline for semantic analysis in Alzheimer's Disease detection from spontaneous speech

Dylan Phelps[1,2], Rodrigo Wilkens[3], Edward Gow-Smith[2], Lilian Hubner[4], Bárbara Malcorra[4], César Rennó-Costa[5], Marco Idiart[6], Maria-Cruz Villa-Uriol[1,2], Aline Villavicencio[2,3,5]

1 Healthy Lifespan Institute, The University of Sheffield, Sheffield, United Kingdom
2 School of Computer Science, The University of Sheffield, Sheffield, United Kingdom
3 Institute of Data Science and Artificial Intelligence, The University of Exeter, Exeter, United Kingdom
4 Faculty of Linguistics, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, Brazil
5 Bioinformatics Multidisciplinary Environment, Digital Metropolis Institute, Federal University of Rio Grande do Norte, Natal, Brazil
6 Institute of Physics, Federal University of Rio Grande do Sul, Porto Alegre, Brazil

Abstract

Alzheimer's Disease (AD) is a progressive neurodegenerative condition that adversely affects cognitive abilities, including language comprehension and production. These language-related changes can be automatically identified through the analysis of outputs from linguistic assessment tasks, such as picture description. Language models show promise as a basis for screening tools for AD, but their limited interpretability poses a challenge in distinguishing true linguistic markers of cognitive decline from surface-level textual patterns. To address this issue, we examine how surface form variation affects classification performance, with the goal of assessing the ability of language models to represent underlying semantic indicators. We introduce a novel approach where texts surface forms are transformed by altering syntax and vocabulary while preserving semantic content. The transformations significantly modify the structure and lexical content, as indicated by low BLEU and chrF scores, yet retain the underlying semantics, as reflected in high semantic similarity scores. This allows us to isolate the effect of semantic information on AD detection, finding models perform similarly to if they were using the original text, with small ± 0.1 deviations in macro F1. We also investigate whether language from picture descriptions retains enough detail to reconstruct the original image using generative models. We found that visual elements are poorly preserved, and image-based transformations add substantial noise, reducing text alignment and classification accuracy. Our methodology provides a novel way of looking at what features influence model predictions, and allows the removal of possible spurious correlations. We find that just using semantic information, language model based classifiers can still detect AD. This work shows that difficult to detect semantic impairment can be identified, addressing an overlooked feature of linguistic deterioration, and opening pathways for better early detection systems.

7.1 Introduction

Globally, both the proportion and the number of people 65 years or older are growing rapidly, with 16% of the population, or 1.6 billion people, expected to be over 65 by 2050 (of Economic and Social Affairs, 2022). Age is the highest risk factor for Alzheimer's Disease (AD), with 1 in 14 over 65 being affected by the condition (Wittenberg et al., 2020). With this aging population and the increasing number of people living with AD, there is a need for methods of diagnosis that are more efficient and that capture the changes early.

AD is a clinical condition characterized by progressive cognitive decline. Subtle changes in speech and language often provide early indicators of impairment (Forbes-McKay and Venneri, 2005) and can provide information on the nature of deterioration

(Ahmed et al., 2013). The effects of AD on memory interact with language abilities, with both the access to words (the lexicon) and their meaning (semantics) being affected by the disease (Rodrigues et al., 2024; Coffey et al., 2024). Additionally, some effect on syntax has been observed (Saffran et al., 1989; Thompson et al., 1997), however, it has been argued that this effect is due to underlying semantic impairment (Reilly et al., 2011).

The current battery of tests used during diagnosis includes extensive speech and language assessments, which have proven valuable for identifying the deterioration of linguistic and other cognitive constructs (Hernández-Domínguez et al., 2018; Sanborn et al., 2022). One tool generally included in batteries of linguistic diagnosis tests is the 'picture description task'. Within this task, a participant should describe or tell a story depicted in a fairly simple image or set of images, and their responses can be recorded and further analysed for symptoms of AD (Forbes-McKay and Venneri, 2005; Hübner et al., 2019; Luz et al., 2020). In this context the semi-automatic processing of the speech or transcription has been investigated for diagnosis with feature based methods extracting lexical and syntactic information from the surface form (Petti et al., 2020), and many recent works building upon language models such as BERT (Devlin et al., 2019; Balagopalan et al., 2020) and GPT-2 (Radford et al., 2019; Li et al., 2022), and further using Large Language Models (LLMs), such as GPT4 (B T and Chen, 2024).

However, while feature based methods explicitly incorporate lexical and syntactic features such as pronoun-noun ratio (Bittner et al., 2022), lexical frequency (Almor, 1999), and repetition (Ben Ammar and Ben Ayed, 2018a) (Petti et al., 2020), language model–based methods are more difficult to interpret due to the opaque nature of the distributed representations within language models (Rogers et al., 2020). As they can use various features of the text, it remains uncertain whether they depend primarily on superficial characteristics or can truly grasp the more nuanced linguistic structures beneath, and how robust they are to surface form changes.

Therefore, in this work we seek to produce transformed versions of the speech of AD patients, that capture just the underlying meaning whilst changing the surface form, and thus syntactic and lexical features. Semantic degradation has been shown to be a key factor in AD-related language impairment (Reilly et al., 2011), it is important to explore specifically the role of semantic information in automatic AD detection, and separate it from the surface level indicators.

7.1.1 Research questions

We propose a pipeline to disentangle semantic features from surface form features in AD patients' speech by transforming the original text. Specifically, we use generative language models to standardize the surface characteristics of the texts, therefore, producing text that still has a high level of semantic similarity with the original, allowing for classification models to be trained that focus only on the semantic information. In this study, these transformations consist of restructuring the content into a storyboard format that sequentially organizes key information into text-only scenes and producing summaries of different lengths.

Additionally, as an extreme case study where all surface elements of the text are removed, we propose using models that can create images from text descriptions (text-to-image) and, vice versa (image-to-text) to extract only the semantic component from the original texts.

We evaluate our methodologies by comparing the original data to our created data using two similarity metrics for the surface form (BLEU (Papineni et al., 2002), chrF (Popović, 2015)) and a semantic similarity metric (embedding cosine similarity (Reimers and Gurevych, 2019)). We also use the transformed data to train a BERT classifier (Devlin et al., 2019) for automatic AD detection, to compare the performance on a downstream task.

In this work, we aim to address three research questions:

1. Can we use LLMs to standardise the AD transcripts surface form, removing lexico-syntactic indicators, whilst retaining the semantic information?

2. How do language model based AD classifiers perform when focusing solely on the semantic information?

3. What does this tell us about the importance of semantic information in automatic AD detection?

By answering these questions, we hope to explore what information language models use in the detection of AD and propose methods for extracting semantic information so that the semantic impairment of patients with AD can be explored further.

Overall, our results show that our text-to-text methods can alter the surface form sufficiently, leading to texts with low surface form similarity (BLEU < 0.1, chrF < 0.5)

and high semantic similarity (> 0.6). We also show that this transformed text can still be used to automatically classify between AD patients and the control group with small ±0.1 macro F1 differences between the best performing transformations and the original data. Conversely, we show that current text-to-image and image-to-text generation algorithms introduce to much noise, which hinders the accurate preservation of semantic information. This results in low semantic similarity and greater performance variability.

The small changes in AD detection performance for our text-to-text pipeline show that language models are robust to changes in the surface form and therefore are capable of making classifications based on the underlying semantic information. Future work employing this technique can further explore specifically the role of semantics in AD-related language degradation. Additionally, the method opens up possibilities for anonymisation, as surface level identifying information could be removed, and synthetic data generation.

7.2 Background

This section has been adapted for use as part of the general background for the thesis and can be found in Section 2.3.

7.3 Materials and methods

7.3.1 Datasets

We use two datasets in our analysis to ensure our results transfer across different groups of AD patients and control groups. We use one dataset in English and one in Portuguese, allowing us to compare the performance of the proposed method in two different languages. A summary of the size of the datasets and the average length of each entry is show in Table 7.1

Dog Story dataset

"The Dog Story", a subtest of the Battery for Language Assessment in Ageing (Hübner et al., 2019), is a dataset of transcribed responses, in Brazilian Portuguese, where participants were asked to tell a story based on seven scenes. The dataset has transcripts

Dataset	Number of Texts			Transcript Length	
	AD	Control	Total	AD	Control
Dog Story	23	116	139	105(±58)	127(±54)
ADReSS	78	78	156	90(±51)	105(±50)

Table 7.1: Count of the number of texts for each group in each dataset and the mean(±std) transcript length, by number of words, in the AD and control groups.

from 139 participants. Of those participants, 23 had been diagnosed with AD, whilst 116 were from the control group. As a result, the dataset is very imbalanced, which presents challenges when using it to train and evaluate a classification model. The measures taken to handle this imbalance are described in the next section.

Alzheimer's Dementia Recognition through Spontaneous Speech

In addition to evaluating the Dog Story task, we also evaluate the Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) dataset (Luz et al., 2020). ADReSS is a set of transcripts of patients' and controls' spontaneous speech in response to the Cookie Theft picture task (Goodglass and Kaplan, 1983). The dataset is balanced for diagnosis, age, and gender, with a total of 156 participants, 78 diagnosed with AD and 78 controls. This allows results to be compared with the larger body of work on AD recognition through language, due to the widespread use of ADReSS.

Ethics Statement

Both datasets used in the study, Dog Story and ADReSS, are obtained via third parties. Data was fully deidentified and included speech recordings, transcripts and basic demographic data. The Dog Story dataset was collected between 2014 and 2023 to analyse speech in the elderly for individuals with cognitive decline. Ethics approval was granted by the Pontificia Universidade Catolica do Rio Grande do Sul (Ref: 53696221.4.1001.5336). DementiaBank reviewed and approved our data access request to use the ADReSS dataset, which is a subset of the Pitt Corpus (Becker et al., 1994), collected between 1983 and 1988 by the University of Pittsburgh Alzheimer's Research Program.

For both datasets, all enrolled participants provided informed written consent. The University of Sheffield provided ethical approval and confirmed that this work meets the conditions consented to by participants in both datasets (Ref: 067658).

7.3.2 Transformation pipeline

We use LLMs to transform the transcribed speech of participants responding to the picture description task, to alter the surface form, whilst maintaining the semantic information. To do this, we propose a pipeline of methodologies, with each stage producing text that abstracts the meaning from the original descriptions. The pipeline can be seen in Figure 7.1, and an example of the output from the pipeline applied to a random example from the Dog Story dataset can be found in Appendix B.

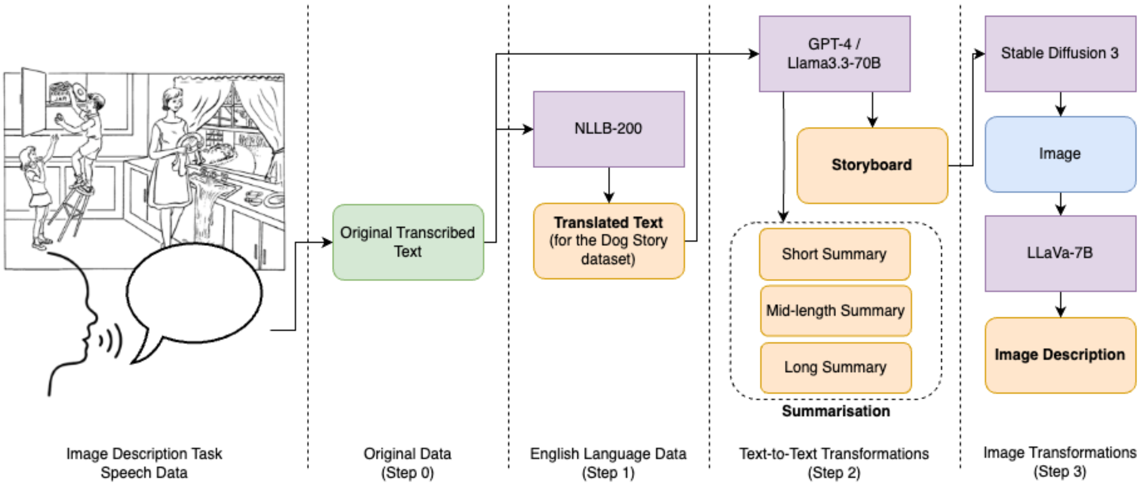


Figure 7.1: Our pipeline to transform the text. We present 3 main steps: translation, text-to-text transformation, and image and caption generation. The models used in the pipeline are shown in purple, the pipeline outputs used in the experiment are shown in orange.

Step 1 in the pipeline translates non-English data into English using a generative language model. This applies only to the Dog Story dataset (originally in Portuguese). We use NLLB-200 (Team et al., 2022) for this translation, enabling all datasets to utilize the same state-of-the-art English language models in subsequent steps.

Steps 2a and 2b both perform text-to-text transformations using generative language models. For the Dog Story dataset, we employ GPT-4o (OpenAI et al., 2024b), the current SOTA model. Due to data agreement restrictions for ADReSS, we cannot use 3rd party models, therefore, we use a 6-bit quantized version of Llama-3.3 70B (Grattafiori et al., 2024), which is optimized for our hardware (1 80GB H100 GPU). All the prompts used across steps 2 and 3 can be seen in Appendix A.

In step 2a, we transform the text into structured "Storyboards", while in step 2b, we generate summaries at three different lengths. For step 2b, we produce 3 different summary lengths to investigate how the similarity and downstream performance change as the model is forced to remove more information to create shorter texts. Both methodologies will keep the key semantic information described in the original text while altering the surface form. Additionally, the transformation in step 2a should improve the text's formatting for image generation in step 3.

The final transformation (step 3 in Figure 7.1) converts the storyboard descriptions into images using a text-to-image model, then regenerates text using an image-to-text model. This process leverages the fact that images represent semantic content without syntactic information, allowing us to extract meaning independently from the original surface forms. After finding similar performance between third-party and local models for image and caption generation, we use StableDiffusion-3XL (Esser et al., 2024) for image generation and LLaVa-8B (Liu et al., 2023a) for caption generation across both datasets.

7.4 Metrics

To assess to what extent the proposed method successfully changes surface forms while maintaining the relevant semantic information, we adopt similarity measures targeting both surface forms and semantic information. Additionally, we evaluate the extrinsic impact of these transformations in a classification task of AD vs. Control groups.

7.4.1 Similarity scoring

We use a number of similarity score metrics to compare the outputs from the different types of transformation directly to each other and the original texts. These metrics verify to what extent the surface form of the generated texts is substantially different from the originals while maintaining the underlying information linked to AD relevant for classification.

For surface form similarity, we use the similarity metrics BLEU(Papineni et al., 2002) and chrF(Popović, 2015). BLEU uses the proportion of matching n-grams (sequences of n adjacent words) to calculate a measure of lexical overlap between pairs of texts, whereas chrF performs a similar calculation but instead at the character n-gram level (sequence of n characters). We use both metrics, as whilst BLEU can capture word choice and phrase structure, it can also miss partial matches such as from morphological inflection or from different word forms (e.g. "quick" vs "quickly" scores 0 BLEU, but > 0 chrF). Both metrics produce scores between 0 and 1, with higher scores implying a closer match. chrF will typically be higher for any pair of sentences as character-level matching is less strict, with scores < 0.4 indicating a poor match, whilst a BLEU score of < 0.2 indicates similarly poor quality. Therefore, if both metrics are low, we can conclude that there is very little overlap between the surface forms resulting from the different transformations.

For semantic similarity, we use the cosine similarity between SentenceBERT (Reimers and Gurevych, 2019) embeddings. SentenceBERT has been shown to effectively encode semantic information into dense vector representations, positioning semantically similar texts closely within the embedding space. Texts with comparable meanings have vectors separated by small angles, resulting in high cosine similarity values. Cosine values range between 0 and 1, with > 0.6 generally indicating a high level of similarity and thus semantic information being conserved across the transformation, while values < 0.4 indicate low similarity and low semantic conservation. In our experiments, we use all-mpnet-base-v2, which is a variant of MPNet (Song et al., 2020) fine-tuned for high-quality sentence representations, to create the embeddings for each of our texts.

7.4.2 Classification

After verifying that the transformations substantially altered the surface form of the original texts, we now aim to determine whether we can automatically classify AD from the information after each of the transformations. To this end, our classification is based on a standard methodology using BERT, which has been shown to have a high performance on the ADReSS dataset by Balagopalan et al. (2020).

For each step of the pipeline for the Dog Story dataset, we perform 5-fold cross-validation and train our BERT model on the remaining 80% of the dataset for each fold. The ADReSS dataset contains ready-defined train and test splits that we have adhered to in our evaluation. When training and evaluating the data in English, we use the bert-base-uncased model (Devlin et al., 2019), and for Portuguese, we use BERTimbau Base (Souza et al., 2020).

Each classification model is trained for 10 epochs with early stopping to prevent overfitting. To calculate the overall score for each step, we average the accuracy (on each class individually) and macro F1 achieved on each of the folds. Due to the large class imbalance in the Dog Story dataset, we use a weighted loss function when training the classifier on this dataset, using the '`calculate_class_weight`' function from scikit-learn (Pedregosa et al., 2011). Additionally, we use macro F1 as the main metric for our analysis.

To allow for comprehensive comparison, we perform 5 classification runs with each transformation. Significance tests are performed using a two-sample t-test using the average and standard deviation across the 5 runs.

7.5 Results

We discuss the results obtained both in terms of the intrinsic similarities between the original texts and their transformations, and also of the extrinsic performance of AD classification. The goal is to determine if the relevant signals for the latter can still be accessed even with transformations, and we use the former to measure substantial changes.

7.5.1 Similarity results

Results with all three similarity measures are compatible with our expected results for the transformations on both datasets. We calculate the mean similarity between corresponding pairs of sentences in the original and transformed texts. Table 7.2 presents the mean similarity scores between each transformation and the original text (or the translated English text for the Dog Story dataset). Pairwise similarity metrics between each transformation are shown and discussed in Appendix C.

Step	Transformation	Dog Story			ADReSS		
		chrF	BLEU	Cosine	chrF	BLEU	Cosine
2	Short Summary	0.39	0.03	0.69	0.40	0.03	0.60
2	Medium Summary	0.47	0.06	0.74	0.43	0.03	0.59
2	Long Summary	0.35	0.05	0.76	0.17	0.02	0.67
2	Storyboard	0.26	0.05	0.68	0.27	0.03	0.66
3	Image Description	0.25	0.01	0.39	0.23	0.01	0.45

Table 7.2: The similarity scores compared to original data for the surface form (BLEU and chrf) and semantic (Cosine) metrics. The results shown are average values across the examples in each dataset.

The Dog Story dataset

Firstly, for the Dog Story dataset, all the transformations in the pipeline produce texts with low surface form similarity, as can be seen by the low BLEU scores. The score for each transformation when paired with the translated text is < 0.06, which shows an extremely low level of lexical overlap between the two texts.

For semantic similarity, the mean cosine similarity scores are relatively high (≥ 0.68) between the translated text and all the step 2 text-to-text transformations (Table 7.2 Cosine). In contrast to the BLEU scores, this shows that much of the semantic information from the original text has been carried through into the transformed data. However, when looking at the similarity compared to the step 3 image captions, a pronounced drop to a cosine similarity score of 0.39 can be seen, suggesting that in transforming between modalities (from text to image to text) some of the relevant information may be lost. This may be due to the text lacking sufficient detail to reconstruct the salient elements of the original image, or possibly because current multi-modal models aren't sophisticated enough to preserve the relevant information.

ADReSS dataset

The similarity scores on the ADReSS dataset follow a similar trend, with low surface form (BLEU ≤ 0.03; chrF ≤ 0.43 and high semantic similarity (≥ 0.59) between the transformed data and the original text. However, in general, we see slightly lower semantic similarity scores, which may be a side effect of using the (smaller) local models on this dataset compared with the GPT-4o for the Dog Story data. Again, we see lower similarity with the image caption, 0.45, suggesting that the loss of information through the image generation is not unique to any one dataset.

7.5.2 Classification results

Dog Story dataset

Table 7.3 shows both the accuracy per class and the macro F1 scores achieved when a BERT classifier was trained at each step of the transformed Dog Story dataset. All metrics are the mean obtained across 5 runs, each performed with a different seed; each run adopts a 5-fold cross-validation, with the metrics averaged across the folds.

We achieve an F1 score of 0.602 when training the BERTimbau Base model on the

Step	Transformation	macro F1	Accuracy per class	
			Positive (AD)	Negative (C)
0	Original Portuguese	0.602	0.270	0.914
1	Translated English	0.667	0.383	0.922
2	Short Summaries	0.562	0.226	0.890
2	Medium Summaries	0.569	0.217	0.907
2	Long Summaries	**0.677**	**0.391**	0.929
2	Storyboard	**0.677**	0.365	0.943
3	Image Description	0.518	0.096	**0.953**

Table 7.3: The mean metrics across the 5 runs. Each data transformation is labelled with its corresponding step in Figure 7.1. The best model for each metric is shown in bold. AD group as Positive (AD) and Control as Negative (C).

original Portuguese data. This performance increases to 0.667 when translating the data to English and using BERT (bert-base-uncased) model for classification.

Our summarized texts show a range of performance, with the short and medium-length summaries achieving lower performance than the original data in Portuguese and English, with F1 scores of 0.562 and 0.569, respectively. Our other step 2 transformations, namely the long summaries and storyboards, achieve higher performance than the translated English data, both achieving 0.677 F1. However, this is not a statistically significant increase ($p = 0.18$ and $p = 0.09$). The imbalance in the data is visible in the > 0.890 negative accuracy for all the classifiers and the relatively lower positive class accuracy.

Finally, we see a substantial decrease in performance when training on the step 3 data, with the model only achieving a 0.518 F1 score. Looking further, we can see that the mean positive class accuracy is only around 0.1, which indicates that the model nearly always predicts the majority negative class.

ADReSS dataset

Table 7.4 shows the mean metrics across 5 runs when a classifier is trained on the data from each of the transformations.

On the ADReSS dataset, similar trends to the Dog Story dataset are obtained with the transformations as the Long Summaries and Storyboards achieve similar macro F1 scores of 0.707 and 0.704, higher than any of the other transformations. Moreover, the Storyboards maintain the same balance of performance for the two classes as the Original data. All the other transformations result in more accuracy for the Negative

Step	Transformation	Macro F1	Accuracy per class	
			Positive (AD)	Negative (C)
0	Original data	**0.795**	**0.792**	**0.800**
2	Short Summaries	0.661	0.608	0.717
2	Medium Summaries	0.677	0.600	0.758
2	Long Summaries	0.707	0.683	0.733
2	Storyboard	0.704	0.708	0.700
3	Image Description	0.538	0.475	0.617

Table 7.4: The mean metrics across the 5 runs. Each data transformation is labelled with its corresponding step in Figure 7.1. The best model for each metric is shown in bold. AD group as Positive (AD) and Control as Negative (C).

class. Finally, although none of the transformations in isolation can reach the 0.795 macro F1 score achieved when using the original data to train the classifier, they are still much higher than the random accuracy of 0.5.

Additionally, we see the same pattern: the image description transformation is far behind the others, with a macro F1 score of 0.538, close to the random baseline score.

7.6 Discussion

The results obtained confirm that while our transformations reduce the surface form (or lexico-syntactic) similarities, they seem to maintain the more latent (or semantic) similarities. While the former are consistently low compared to the original text, with a BLEU < 0.06 for all transformations for both datasets, the latter remain high, and the transformations that score highest on classification, Long Summaries and Storyboards also have the highest semantic similarity. By calculating the Pearson Correlation Coefficient, we find that the semantic similarity is highly correlated with classification performance, with correlation of 0.721 ($p = 0.106$) on Dog Story and 0.919 ($p = 0.01$) on ADReSS.

Further, the image description transformation has the lowest semantic similarity and lowest classification performance on both datasets. Overall, these seem to indicate that while text-to-text transformations are able to keep the relevant semantic indicators linked to AD, these are not maintained in text-to-image-to-text transformations. The poorer performance could be caused by relatively weaker multimodal models, but could also be due to text descriptions containing insufficient information required to produce an image (for example, the clothes worn, everything in the environment, or colours, etc.). This requires the image generation model to invent such details, which are then incorporated into the generated caption. This spurious information reduces the similarity to the original description, thus, reducing utility.

7.6.1 Further syntactic and lexical measures

We now analyse additional metrics for quantifying the surface form differences between the texts produced by each transformation and the original data. We do this to understand the impact of the transformations on the data and whether the transformations also substantially change the distributional profile of the surface forms of the original texts.

Type-Token Ratio and Lexical Frequency

There are a number of syntactic and lexical measures that have been shown in the literature to have different distributions in texts produced by patients with cognitive decline due to AD and study participants without AD. Due to the increased chance of repetition, it has been shown that the speech from AD patients has a lower type-token ratio (ttr) (Fang et al., 2017; Ben Ammar and Ben Ayed, 2018a; Mueller et al., 2016) and use of words with a higher lexical frequency (lf) (Almor et al., 1999). These two measures have also been shown to be significant markers of AD when used as features for machine learning systems (Ben Ammar and Ben Ayed, 2018b), and have been used as evaluation metrics for text generated by language models as a response to picture description tasks (Li et al., 2022). To evaluate to what extent these measures differ in the original texts from both groups (AD and Control) and if our transformations alter them, we calculate the average of these values in the original text and the transformations and measure if there are significant differences between them. Again, we use a two-sample t-test to compare the means and standard deviations of the control and AD groups. We focus this analysis only on the ADReSS dataset, for comparative reasons with related work.

Comparing the original data for the two groups in the ADReSS dataset, neither of these measures is significantly different between the groups (Table 7.5, Original). Moreover, our transformations do not seem to introduce any marked differences between the two groups (from Short Summary to Image Description), which a model could be using as a clue for distinguishing between them in the classification task. Although the ‘Long Summaries’ show some difference, with lower average ttr and higher average lf in the AD group than in the control, the effect is not statistically significant ($p = 0.103$ and $p = 0.092$).

Transformation	ttr_C	ttr_{AD}	ttr_p	lf_C	lf_{AD}	lf_p
Original	0.607	0.603	0.768	1.679	1.706	0.546
Short Summaries	0.821	0.828	0.478	1.225	1.215	0.496
Medium Summaries	0.818	0.833	0.199	1.231	1.209	0.175
Long Summaries	0.541	0.510	0.103	2.006	2.390	0.092
Storyboard	0.597	0.609	0.465	1.770	1.700	0.380
Image Description	0.597	0.609	0.465	1.770	1.700	0.380

Table 7.5: The mean measures for the control ($_C$) and AD ($_{AD}$) on the ADReSS dataset and transformations. P-values ($_p$), calculated using a paired t-test, are given for the difference between the two groups.

Part-of-Speech based metrics

Other measures that have also been shown to be useful indicators of AD when analysing language characteristics involve changes in distributions of certain parts-of-speech, specifically, with increased pronoun-to-noun (pnr) ratio (Bittner et al., 2022), adverb ratio (RB), and participle ratio (VB) (Fraser et al., 2016) having been observed. The average value for each of these measures in the AD and control group for each of the translations, and the p-value for the statistical significance of the differences, are shown in Table 7.6. Although there is a statistically significant difference in all the measures in the original text (indicated by the p-values lower than 0.001, for Original), for two of the measures, pronoun-to-noun ratio and participle usage, this effect is not statistically significant in all of our transformations. The difference remains statistically significant for increased adverb usage in 3 of our 5 transformations. However, the absolute differences are lower.

These analyses provide additional confirmation that our transformations significantly change the surface form characteristics of the texts. This, paired with the high semantic similarity and the classification performance, getting close to or surpassing the original text for some of our transformations, shows that language models are still able to access the relevant latent indicators linked to AD even after the transformations applied to a patient’s response.

Trans.	pnr_C	pnr_{AD}	pnr_p	RB_C	RB_{AD}	RB_p	VB_C	VB_{AD}	VB_p
Original	**0.265**	**0.441**	**<0.001**	**0.033**	**0.054**	**<0.001**	**0.073**	**0.054**	**<0.001**
Short Summaries	0.108	0.107	0.964	**0.010**	**0.019**	**0.007**	0.109	0.106	0.648
Medium Summaries	0.122	0.157	0.155	0.012	0.016	0.256	0.103	0.101	0.853
Long Summaries	0.140	0.148	0.529	0.030	0.032	0.497	0.075	0.072	0.291
Storyboard	0.100	0.106	0.604	**0.019**	**0.024**	**0.042**	0.052	0.056	0.245
Image Description	0.100	0.106	0.604	**0.019**	**0.024**	**0.042**	0.052	0.056	0.245

Table 7.6: The mean measures for the control ($_C$) and AD ($_{AD}$) on the ADReSS dataset and transformations. P-values ($_p$) are given for the difference between the two groups. Any statistically significant differences ($p < 0.05$) are shown in bold.

7.6.2 Effect of translation

When translating from a source to a target language, substantial changes in the surface form would be expected due to differences in conventionality between the two languages, with possible loss of information, which could affect performance for AD classification. On the other hand, the translation into a more resourced language, like English, allows access to potentially bigger models trained on larger quantities of data. Therefore, the impact of these changes needs to be determined.

From our results for the Dog Story dataset, translating the original texts into English and using an English-only classification model increases the performance compared to processing the data in Portuguese (in Table 7.7 Original Portuguese vs Translated English). As we cannot directly compare the syntax across languages, it is hard to understand whether this improvement comes from the translation fundamentally changing the text, or it is a result of English language models performing better due to English being a more richly resourced language.

To test this, we explore translating the text back into Portuguese from the translated English data using the NLLB-200 model. We then train a classifier based on BERTimbau as we did with the original data. The results are in Table 7.7.

Data Transformation	Macro F1	Positive Accuracy	Negative Accuracy
Original Portuguese	0.602	0.270	0.914
Translated English	**0.667**	**0.383**	0.922
Back Translated Portuguese	0.636	0.313	**0.927**

Table 7.7: The mean metrics for the translations across 5 runs of 5-fold cross-validation. The best model for each metric is shown in bold.

Whilst the results are not as high as those achieved with the English data, they are significantly better than those achieved with the original Portuguese data. This suggests that there are performance gains from using the English language models. Moreover, the translation seems to fundamentally change the text, but in a way that improves access to the latent information that is relevant for AD.

A manual analysis of the data confirms that both the translations into English and back-translations to Portuguese are mostly good quality, with some small grammatical mistakes made by the model. The back translations appear to be more fluent, removing some of the markedness of the original texts, especially those of AD patients. Although we would expect this change to negatively affect the AD classification performance as the irregularities should be useful indicators, this may remove some of the confounding clues from the surface form and make detection of the latent semantic differences easier. Further work is needed to verify this, however.

7.7 Conclusion

In this work, we examined how we can use large language models to modify the transcripts of patients with Alzheimer's Disease, such that the surface (lexico-syntactic) form is changed, whilst keeping the meaning (semantics) the same. Using a range of metrics, we show that our pipeline of text transformations successfully modifies the lexico-syntactic features of the texts, as measured with low BLEU and chrF scores, whilst maintaining high semantic similarity, as measured with cosine similarity from SentenceBERT. Across two datasets in English and Portuguese, we train BERT classifiers with both these modified texts and the originals, finding only small, both positive and negative, performance differences on Alzheimer's detection. These results contribute to our understanding of how Alzheimer's affects language production, suggesting that semantic degradation is the most fundamental effect of the progression of the disease, with lexico-syntactic degradation being a byproduct, supporting the conclusions of some prior work.

We also look at an approach which generates images from LLM-generated storyboards, and then captions them back into text, as a further way to create summaries. This approach, however, leads to a reduction in performance, requiring further investigation.

Our work presents future opportunities to specifically analyse the effects of AD on the semantic component of linguistic production. As semantic degradation is often an early marker of AD Vonk et al. (2020), this could include use within temporal pipelines, allowing researchers to monitor linguistic deterioration over time. Further more, the transformation of AD-related texts into versions that maintain semantic information could be used for anonymisation and synthetic data generation.

A Generation prompts

Step	Transformation	Prompt
2	Short Summary	Summarise the following text into a one sentence summary. Just output the summary and no other information.
2	Medium Summary	Summarise the following text into a concise summary. Just output the summary and no other information.
2	Long Summary	Summarise the following text into a long summary containing as much information as possible. Just output the summary and no other information.
2	Storyboard	Transform any text you are given into key story scenes. Focus on the most important story moments. Break down complex actions into separate scenes if needed. Just output the storyboard and no other information.
3	Image Caption	Describe in detail what is happening in the image.

Table 7.8: Prompts used with the (multimodal) Large Language Models for generation at each step in our pipeline

B Example pipeline output

An example of the output produced by our pipeline for a randomly selected example from the Dog Story dataset can be seen in Table 7.9.

C Full similarity results

In addition to performing the similarity test between the original texts and the transformations, we also calculate the same metrics between all pairs of texts. These metrics are presented as confusion matrices in Figure 7.2 for the Dog Story dataset, and Figure 7.3 for ADReSS.

Overall, a similarly low level of syntactic similarity is seen between the transformed texts, as was seen when comparing each transformation to the original data. This shows that our transformations produce diverse texts and strengthens our claims that the classification models are robust to changes in the surface form.

For semantic similarity, again, the pairwise similarities are high (except comparisons with the image caption) and in the range of scores between the transformed texts and the originals. This shows that the general meaning of the texts is being preserved across all the transformations, with some small variations in the information kept by the models. This may explain some of the variation in classifier performance observed.

Figure 7.2: Complete pairwise similarity scores for all the transformations of the Dog Story dataset, presented in confusion matrices.

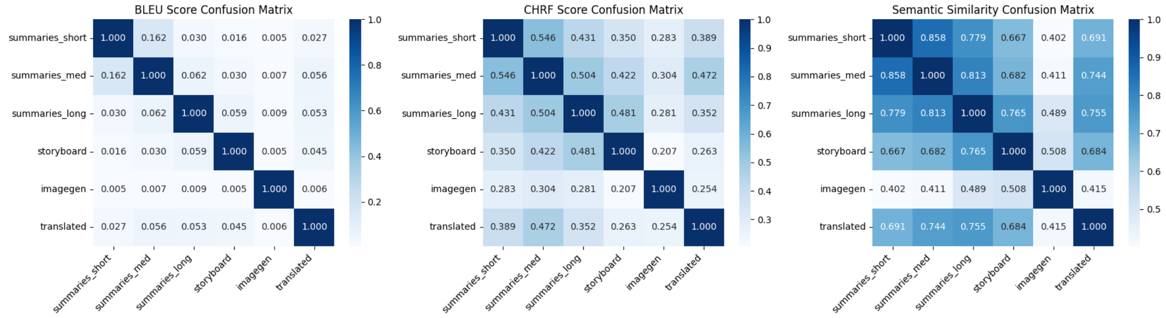
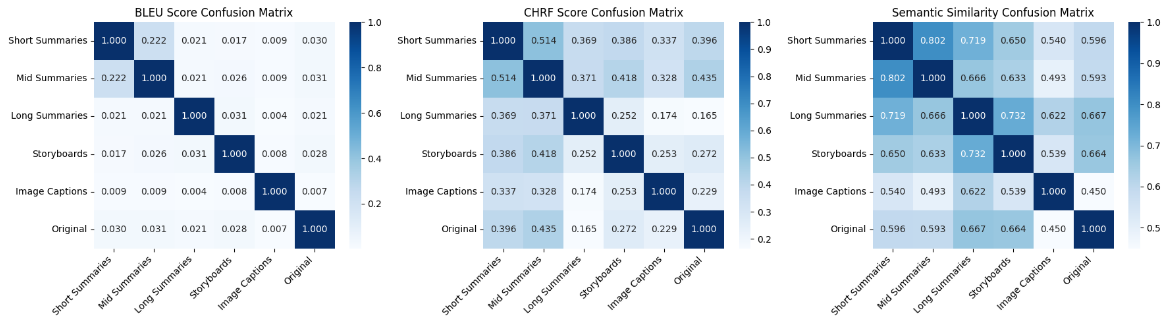


Figure 7.3: Complete pairwise similarity scores for all the transformations of ADReSS, presented in confusion matrices.



Step	Transformation	Prompt
0	Original	O menino ia passando na rua e viu um cachorrinho. Gostou do cachorrinho, aí chamou o cachorrinho, e o cachorrinho veio do lado dele. Aí, o cachorrinho acompanhou ele, né? Aí, chegando em casa, ele escondeu o cachorrinho dentro do roupeiro pra sua mãe não ver. Mas sua mãe achou o cachorrinho dentro do roupeiro e brigou com o menino. O menino pediu, por favor, pra ela, e aí ela deixou. Tanto é que fez a casinha e tudo, ajudou ele a fazer a casinha pra colocar o cachorro.
1	Translated	The boy was walking down the street and saw a puppy. He liked the puppy, so he called the puppy, and the puppy came to his side. Then the puppy followed him, didn't it? Then, coming home, he hid the puppy inside the wardrobe so his mother wouldn't see it. But his mother found the puppy inside the wardrobe and got into a fight with the boy. The boy asked her for it, and then she left. So he did the house and everything, helped him make the house to put the puppy.
2	Short Summary	A boy found a puppy he liked on the street, hid it from his mother who eventually discovered it, leading to a disagreement after which he helped make a place for the puppy at home.
2	Medium Summary	The boy found a puppy on the street and brought it home, hiding it in a wardrobe to keep it from his mother. However, she discovered the puppy, leading to an argument. Eventually, the boy convinced her to let him keep the puppy, and they created a space for it at home.
2	Long Summary	A boy spotted a puppy while walking and immediately took a liking to it. When he called the puppy, it obediently came to him and began to follow him. Wanting to keep the puppy a secret, the boy hid it inside a wardrobe upon reaching home. However, his mother soon discovered the puppy and confronted the boy, leading to a disagreement. Despite the initial conflict, the boy pleaded with his mother, who eventually relented. The boy took on the responsibility of making a suitable home for the puppy, diligently preparing a space for it.
2	Storyboard	1. **The Encounter:** The boy walks down the street and spots a playful puppy. He feels an instant fondness for the little creature. 2. **The Bonding:** The boy calls out gently to the puppy, and it eagerly trots to his side, forming an immediate companionship. 3. **The Secret:** As they reach home, the boy decides to keep the puppy a secret and carefully hides it inside his wardrobe, hoping his mother won't notice. 4. **The Discovery:** The boy's mother unknowingly opens the wardrobe and is surprised to find the hidden puppy, leading to a confrontation with her son. 5. **The Conflict:** The boy pleads with his mother to keep the puppy, expressing his desire and attachment to it. 6. **The Resolution:** Touched by her son's plea, the mother decides to leave the room, allowing him to figure out how to care for the puppy. 7. **The New Beginning:** With determination, the boy takes on responsibilities, cleaning the house and creating a makeshift home to care for his new puppy, symbolizing his commitment and love.
3	Image Caption	The image displays a red door with a white envelope tucked into the door jamb. This suggests a scenario where someone has left a message or letter for the person inside. The envelope, peeking out from the door, adds a sense of anticipation or communication, perhaps left for someone who isn't home. The area is otherwise devoid of people or additional actions, focusing solely on the simple, but potentially significant act of leaving a note.

Table 7.9: Example outputs from each stage of our pipeline for a randomly chosen example from the Dog Story dataset.

Chapter 8

Conclusion

This thesis investigated two forms of distributional semantic outliers: idiomatic expressions and semantic changes in Alzheimer's disease. The work contributes state-of-the-art techniques for idiomaticity representation, provides the first systematic evaluations of LLMs' ability to process idiomaticity, and introduces novel semantic isolation methods for automatic AD detection systems.

8.1 Summary

Publication I uses the BERTRAM framework to generate new single-token embeddings for a number of noun-compound idiomatic expressions. As a submission to SemEval-2022 Task 2, a SentenceBERT model using these embeddings placed 1st and 2nd in the two settings it was entered into. It also achieved the overall highest scores on the idiomatic split, suggesting these embeddings are SOTA for idiom representation. The representations can also be seen as sample efficient as only a few randomly scraped examples of each expression are required, with even better results from as low as 10 examples being achieved with a small amount of manual labelling.

Publication II takes these embeddings, and another sample efficient method, PET, and evaluates their performance on an idiomaticity detection task. Overall, the sample efficient methods provide modest benefits to idiomaticity detection in English, with small increases when using PET and larger ones from BERTRAM. For Portuguese and Galician, the results are more mixed, with a significant reduction in performance being found for most cases in Portuguese. Initial error analyses using monolingual prompts and language models are unable to explain this discrepancy.

Publications III and IV produce the first studies on how well multi-billion parameter language models represent idiomaticity by benchmarking them against a number of idiomaticity detection datasets in 3 languages. Both papers find that the largest models show very high levels of idiomatic understanding, but that smaller models specifically fine-tuned for the task still perform better. In addition, Publication IV uses chain-of-thought reasoning outputs to probe the model's understanding of idioms and their ability to reason about them. Again, these analyses find that the largest models can accurately reproduce and reason with definitions for the majority of expressions, whilst smaller models struggle.

Publication V presents work on isolating semantic information in responses to a common linguistic diagnostic test for AD, the picture description task. By performing a number of transformations, summarisation, storyboard creation and translation, on the original data we show that surface form features often used for diagnosis are removed, and low surface similarity is achieved whilst semantic similarity remains high. Using this transformed data for automatic classification produces similar performance to using the original data.

8.2 Research Questions

This section discusses how the findings from each of the papers in the thesis contributes to our understanding of each of the research questions presented in 1.2.

RQ1: How can low resource techniques be used to improve the representation of idiomatic expressions, and can better representations be used to improve idiomaticity detection?

Publications I and II sought to directly address RQ1 by applying existing sample efficient techniques, used in other domains, to the tasks of idiomaticity representation and detection. Using only a few examples, the single-token idiomatic expression embeddings generated using BERTRAM in Publication I achieved the highest overall performance on the idiomatic subset and placed 1st and 2nd in SemEval-2022 Task 2b. This high performance suggests that these embeddings reach state-of-the-art performance, demonstrating the power of low resource techniques.

On idiomaticity detection, the results in Publication II show that the low resource embeddings can improve performance on the English split of the dataset considerably from the baseline. Another low resource technique, PET, also shows improvement over the baseline. Overall, the results show, that at least on English, low-resource techniques can be effective for idiomaticity representation and detection, however they alone cannot reach SOTA detection performance.

Results from both papers on Portuguese and Galician raise questions about the efficacy of these methods on languages other than English, which explorations in both papers are unable to fully explain.

RQ2: How can very large generative language models with billions of parameters be leveraged for idiomaticity detection, and do they demonstrate enhanced understanding of idiomatic expressions compared to smaller models?

Using idiomaticity detection as a benchmark task, Publication III demonstrates that both large software-as-a-service models and smaller local models achieve high out-of-the-box performance. Most larger models reach performance saturation on the FLUTE dataset, while results on MAGPIE approach similar levels. The exception is SemEval-2022, where generative LLMs fail to match fine-tuned smaller models, this can largely be attributed to challenges with the Portuguese and Galician examples. Few-shot prompting and prompt translation help address these multilingual limitations.

Publication IV builds on these findings by showing that chain-of-thought prompting delivers substantial performance gains, eliminating the gap with fine-tuned models and pushing macro-F1 scores towards 0.9 on both FLUTE and MAGPIE. However, significant performance gaps persist on SemEval, particularly on non-English subsets, demonstrating that reasoning approaches are not a silver bullet. The paper also evaluates DICE, a dataset its authors claim remains challenging for LLMs, yet performance again approaches saturation levels.

The manual analysis in Publication IV reveals that larger models consistently reproduce accurate definitions of idiomatic expressions and reason effectively about them. Notably, the manual analysis reveals that many apparent model errors could be attributed to mislabelled or ambiguous examples in the datasets themselves, rather than poor understanding from the models.

These findings collectively demonstrate that out-of-the-box large language models possess sufficient understanding of idiomatic expressions to effectively saturate English idiomaticity detection tasks, even on datasets specifically designed to challenge current

LLMs.

RQ3: Can language models detect Alzheimer's Dementia from speech when relying solely on semantic information rather than surface-level linguistic features?

The work in Publication V address the final research question, RQ3. These results show that LLMs can be used to transform text data in a way that removes surface features and allows the isolation of just the semantic information. Analysis in the paper shows that many surface features commonly used to detect AD are removed, and surface similarity metrics are low, whereas semantic similarity metrics remain high. This opens up the possibility of evaluating automatic AD detection systems just on how they detect semantic shifts in patients with AD.

Despite the removal of the surface form features, evaluation with commonly used frameworks for AD detection with language models produce similar results to when the original data is used. The results vary slightly between the Dog Story dataset, where performance is improved with transformation, and the ADReSS dataset, where performance is slightly reduced. However, overall, it can be seen that semantic information alone can be used as a marker for AD.

8.2.1 Overall Findings

Beyond addressing the individual research questions, the results shown in this thesis have revealed broader insights into how modern LLMs handle distributional semantic outliers. A summary of contributions and overall findings can be seen in Figure 8.1, showing the main finding for each research question as well as the main finding taken from research questions 2 and 3.

Both idiomatic expressions and language in Alzheimer's disease represent challenging cases of semantic outliers, though for different reasons. Idioms violate the principle of compositionality, whereas the semantic changes in AD manifest as subtle shifts in meaning comprehension and production. These two phenomena thus represent complementary challenges: idioms require models to move beyond compositional semantics, while AD language requires models to represent the subtle out-of-distribution changes that serve as markers of AD.

For idiomaticity, the thesis employed a progression of approaches with increasing

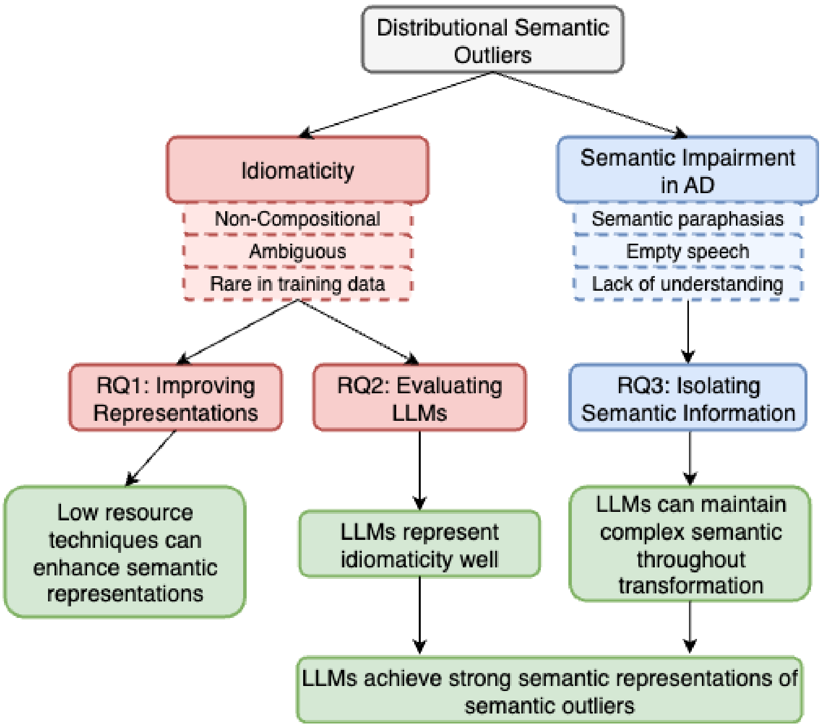


Figure 8.1: Updated version Figure 1.1 giving a summary of the findings for each research question as well as the overall findings for RQ2 and RQ3

model scale. Initial work with encoder models began with sample-efficient methods to inject knowledge through specialised embeddings and pattern-exploit training (PET). The work then progressed to evaluating the performance of LLMs, and revealed that multi-billion parameter models achieve near-saturation performance on idiomaticity detection tasks out-of-the-box, with chain-of-thought reasoning further closing the gap with specialised fine-tuned systems.

For AD detection, the thesis used LLMs to implement multiple types of transformation: summarization, storyboard creation, and translation. These transformations were used to strip away surface-level features while preserving semantic content. This demonstrated that even in the cases of out-of-distribution changes, LLMs can still accurately capture semantics, as shown by the similar performance achieved when using the transformed data for the downstream detection task.

The central finding across both research strands is that modern multi-billion parameter language models possess robust semantic representations that extend to distributional outliers. These models demonstrate strong performance on idiomaticity detection tasks, accurately reproduce and reason about idiomatic meanings, and can perform semantic-preserving transformations of AD patient narratives. This suggests that contemporary LLMs have internalized sufficiently rich semantic knowledge to handle non-compositional meaning and detect subtle semantic degradation, capabilities that smaller or more specialized models often struggle to achieve consistently.

Limitations

However, important limitations emerged across the work. The efficacy of all the techniques explored for idiomaticity, including even the largest models, varied significantly across languages, with Portuguese and Galician showing inconsistent results that remain partially unexplained. For AD detection, while semantic information alone proved diagnostically useful, the transformation approach introduced its own complexities, with performance impacts varying across datasets.

Additionally, a limitation of the work as a whole is the focus on just two cases of semantic outliers, future work should focus on expanding the analyses to other categories of semantic outliers. Two additional categories explored as part of this PhD, but not presented in this thesis, include other types of figurative language (metaphor, simile, sarcasm, etc.), which display similar non-compositionality to idiomatic expressions, and clinical expressions, which require additional domain knowledge to understand.

8.3 Impact of Work

Publications I and II represent an improvement on representations of idiomaticity, achieving SOTA which caused them to place 1st and 2nd in the two SemEval-2022 Task 2b subtasks. Although much of the field has moved on from explicit word embedding techniques, both papers have been cited in some of the more recent work in this area (He et al., 2024, 2025; Wada et al., 2023).

Publications III and IV provide a foundation for research on how large generative models handle idiomatic expressions. To our knowledge, Publication III represents the first published investigation into large language models' comprehension of any type of figurative language, while Publication IV builds upon this work by examining reasoning models, an analysis we believe is also a first of its kind. Publication III was awarded the Best Paper Award at the *Joint Workshop on Multiword Expressions and Universal Dependencies*.

Publication V opens a new research direction for work in AD detection using purely semantic information from language based diagnostic tests. As discussed in section

2.3, semantic impairment has been shown by clinicians to be an early symptom of cognitive decline in AD, and so dedicated work on detecting this can have wide-ranging effects. As such, we imagine the work, and the results within, as a baseline for future explorations into the utility of semantic information as a marker for automatic AD detection.

8.4 Future Research Directions

Hybrid Idiom Processing While the field has largely shifted toward large language models from the smaller embedding models examined in Publications I and II, smaller encoder models retain significant utility in efficient and/or constrained systems, and as such future work should still look at improving contextual embeddings.

A promising direction involves implementing hybrid processing (Titone and Connine, 1999) approaches for idiomatic expression, rather than the single-token model adopted in our work. The hybrid processing approach suggests that native speakers process potentially idiomatic expressions as both single tokens and compositionally simultaneously, whereas current language models must exclusively use one representation, and has been widely adopted within psycholinguistics. Utilising a hybrid approach may make detection of PIEs easier and provide a basis for richer representations.

Extension to other languages Much of the representational improvements and high performance gains from LLMs discussed in this thesis have been only demonstrated on English language data. While collections of idiomatic expressions exist for other languages, these have not been developed into structured datasets or tasks suitable for training and evaluating LLMs. Extending this work to multilingual contexts would broaden the applicability of these advances, and potentially provide insights into cross-language linguistic patterns.

New datasets for modern LLMs Idiomaticity detection is currently the main task used to test models understanding of idiomatic expressions. However, as the work in Publications III and IV show, LLMs can achieve near human performance out-of-the-box in these tasks and showing good idiomatic understanding, effectively saturating existing English idiomaticity detection benchmarks.

Therefore, to accurately evaluate and compare representation and understanding of idiomaticity within new LLMs, the community requires new tasks and datasets. Work like DICE takes a step towards this, but ultimately, existing models are already performing highly here.

Future work should therefore pursue two complementary directions:

1. Development of new types of tasks for idiomaticity representation evaluation. This work has already begun with several potential tasks, including question answering (Zeng and Bhat, 2023; He et al., 2025) and multimodal idiomaticity detection (Pickard et al., 2025).

2. Creation of datasets featuring novel expressions that require much higher levels of contextual disambiguation, rather than relying on memorized knowledge. Humans demonstrate the ability to interpret new idiomatic expressions, through contextual reasoning, whether LLMs also exhibit this ability is yet to be studied, and would provide more evidence on how they utilise context.

Exploration of new classification architectures for semantic information explicitly The work presented in this thesis used a BERT prediction model that is a standard across many tasks, and previously validated for AD detection on the ADReSS dataset. However, the isolation of semantic features creates opportunities to explore other language model architectures that could better leverage the semantic information. Future work could explore more complex BERT based systems or larger generative models that may capture more nuanced semantic relationships relevant to cognitive decline.

Integrating specific semantic processing into pipelines Our work demonstrates that semantic information alone can effectively detect AD, opening possibilities for more modular detection pipelines. These could integrate our semantic transformation approach with existing syntactic, acoustic, and other linguistic feature-based methods to create interpretable detection systems.

The use of a modular architecture may improve classification performance as the deterioration can be captured in different aspects of language. Additionally, they will enhance interpretability by isolating the contributions of distinct linguistic features. Interpretability is a critical requirement for clinical applications where patient trust in machine learning systems remains low (Hallowell et al., 2022; Thornton et al., 2024).

Being able to explain which specific linguistic factors influence model decisions is a step toward building clinical acceptance.

Moreover, modular pipelines could incorporate temporal analysis to track semantic degradation patterns over time, potentially enabling earlier detection and monitoring of disease progression.

Investigation of understanding of semantic outliers by those with cognitive impairment Although the exploration of idiomatic language and linguistic degradation in AD were explored separately in this thesis, the two research directions could be integrated through an investigation of how those with AD process idiomatic expressions. We could hypothesise that the semantic deterioration caused by AD may lead to lower production and comprehension of more semantically complex items such as idioms and other semantic outliers.

This sort of study could take many forms, for instance, it could look at responses to a picture description task given by AD patients and the control group and compare the number, type, and suitability of produced multi-word expressions.

Bibliography

Marah Abdin, Jyoti Aneja, Sebastien Bubeck, Caio César Teodoro Mendes, Weizhu Chen, Allie Del Giorno, Ronen Eldan, Sivakanth Gopi, Suriya Gunasekar, Mojan Javaheripi, Piero Kauffmann, Yin Tat Lee, Yuanzhi Li, Anh Nguyen, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Michael Santacroce, and 7 others. 2023. Phi-2: The surprising power of small language models.

Otavio Acosta, Aline Villavicencio, and Viviane Moreira. 2011. Identification and Treatment of Multiword Expressions Applied to Information Retrieval. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 101–109, Portland, Oregon, USA. Association for Computational Linguistics.

Felix Agbavor and Hualou Liang. 2022. Predicting dementia from spontaneous speech using large language models. *PLOS Digital Health*, 1(12):e0000168.

Samrah Ahmed, Anne-Marie F. Haigh, Celeste A. de Jager, and Peter Garrard. 2013. Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease. *Brain*, 136(12):3727–3737.

Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. Large Language Models for Mathematical Reasoning: Progresses and Challenges. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 225–237, St. Julian's, Malta. Association for Computational Linguistics.

A. Almor. 1999. Noun-phrase anaphors and focus: the informational load hypothesis. *Psychological Review*, 106(4):748–765.

A. Almor, D. Kempler, M. C. MacDonald, E. S. Andersen, and L. K. Tyler. 1999. Why do Alzheimer patients have difficulty with pronouns? Working memory, semantics, and reference in comprehension and production in Alzheimer's disease. *Brain and Language*, 67(3):202–227.

Alzheimer's Society. 2025. Facts for the media about dementia.

Balamurali B T and Jer-Ming Chen. 2024. Performance Assessment of ChatGPT versus Bard in Detecting Alzheimer's Dementia. *Diagnostics (Basel, Switzerland)*, 14(8):817.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Aparna Balagopalan, Benjamin Eyre, Jessica Robin, Frank Rudzicz, and Jekaterina Novikova. 2021. Comparing Pre-trained and Feature-Based Models for Prediction of Alzheimer's Disease Based on Speech. *Frontiers in Aging Neuroscience*, 13:635945.

Aparna Balagopalan, Benjamin Eyre, Frank Rudzicz, and Jekaterina Novikova. 2020. To BERT or not to BERT: Comparing Speech and Language-Based Approaches for Alzheimer's Disease Detection. In *Interspeech 2020*, pages 2167–2171. ISCA.

Kathryn A. Bayles, Alfred W. Kaszniak, and Cheryl K. Tomoeda. 1987. *Communication and cognition in normal aging and dementia*. Communication and cognition in normal aging and dementia. College-Hill Press/Little, Brown & Co, Boston, MA, US. Pages: xvi, 400.

James T. Becker, Francois Boller, Oscar Lopez, and Judith Saxton. 1994. The Natural History of Alzheimer's Disease: Description of Study Cohort and Accuracy of Diagnosis. *Arch Neurol.*, 1994(51):585–594.

Randa Ben Ammar and Yassine Ben Ayed. 2018a. Speech processing for early alzheimer disease diagnosis: Machine learning based approach. In *2018 IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA)*, pages 1–8.

Randa Ben Ammar and Yassine Ben Ayed. 2018b. Speech Processing for Early Alzheimer Disease Diagnosis: Machine Learning Based Approach. In *2018*

IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA), pages 1–8. ISSN: 2161-5330.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3(null):1137–1155.

Irina Bigoulaeva, Rachneet Singh Sachdeva, Harish Tayyar Madabushi, Aline Villavicencio, and Iryna Gurevych. 2022. Effective Cross-Task Transfer Learning for Explainable Natural Language Inference with T5. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 54–60, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Dagmar Bittner, Claudia Frankenberg, and Johannes Schröder. 2022. Changes in Pronoun Use a Decade before Clinical Diagnosis of Alzheimer's Dementia—Linguistic Contexts Suggest Problems in Perspective-Taking. *Brain Sciences*, 12(1):121.

Joanne Boisson, Luis Espinosa-Anke, and Jose Camacho-Collados. 2023. Construction Artifacts in Metaphor Identification Datasets. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6581–6590, Singapore. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146. Place: Cambridge, MA Publisher: MIT Press.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

John A. Bullinaria and Joseph P. Levy. 2007. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39(3):510–526.

Stéphanie Caillies and Kirsten Butcher. 2007. Processing of Idiomatic Expressions: Evidence for a New Hybrid View. *Metaphor and Symbol*, 22(1):79–108.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-SNLI: Natural Language Inference with Natural Language Explanations. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Marine Carpuat and Mona Diab. 2010. Task-based Evaluation of Multiword Expressions: a Pilot Study in Statistical Machine Translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 242–245, Los Angeles, California. Association for Computational Linguistics.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Tuhin Chakrabarty, Yejin Choi, and Vered Shwartz. 2022a. It's not Rocket Science: Interpreting Figurative Language in Narratives. *Transactions of the Association for Computational Linguistics*, 10:589–606. Place: Cambridge, MA Publisher: MIT Press.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022b. FLUTE: Figurative Language Understanding through Textual Explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. 2025. Unleashing the potential of prompt engineering for large language models. *Patterns*, page 101260.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2024. Navigate through Enigmatic Labyrinth A Survey of Chain of Thought Reasoning: Advances, Frontiers and Future. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1173–1203, Bangkok, Thailand. Association for Computational Linguistics.

Zheng Chu, Ziqing Yang, Yiming Cui, Zhigang Chen, and Ming Liu. 2022. HIT at SemEval-2022 Task 2: Pre-trained Language Model for Idioms Detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 221–227, Seattle, United States. Association for Computational Linguistics.

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, and 16 others. 2024. Scaling instruction-finetuned language models. *J. Mach. Learn. Res.*, 25(1):70:3381–70:3433.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*.

Joseph R. Coffey, Margarita Zeitlin, Jean Crawford, and Jesse Snedeker. 2024. It's All in the Interaction: Early Acquired Words Are Both Frequent and Highly Imageable. *Open Mind: Discoveries in Cognitive Science*, 8:309–332.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, ICML '08, pages 160–167, New York, NY, USA. Association for Computing Machinery.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computa-*

tional Linguistics, pages 8440–8451, Online. Association for Computational Linguistics.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword Expression Processing: A Survey. *Computational Linguistics*, 43(4):837–892.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2008. The vnc-tokens dataset. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 19–22. Citeseer.

Silvio Cordeiro, Aline Villavicencio, Marco Idiart, and Carlos Ramisch. 2019. Unsupervised Compositionality Prediction of Nominal Compounds. *Computational Linguistics*, 45(1):1–57. Place: Cambridge, MA Publisher: MIT Press.

Brian C. Cronk, Susan D. Lima, and Wendy A. Schweigert. 1993. Idioms in sentences: Effects of frequency, literalness, and familiarity. *Journal of Psycholinguistic Research*, 22(1):59–82.

Nicholas Cummins, Yilin Pan, Zhao Ren, Julian Fritsch, Venkata Srikanth Nallanthighal, Heidi Christensen, Daniel Blackburn, Björn W. Schuller, Mathew Magimai-Doss, Helmer Strik, and Aki Härmä. 2020. A Comparison of Acoustic and Linguistics Methodologies for Alzheimer's Dementia Recognition. In *Interspeech 2020*, pages 2182–2186. ISCA.

Raj Dabre, Bianka Buschbeck, Miriam Exel, and Hideki Tanaka. 2023. A Study on the Effectiveness of Large Language Models for Translation with Markup. In *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 148–159, Macau SAR, China. Asia-Pacific Association for Machine Translation.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A Survey of Multilingual Neural Machine Translation. *ACM Comput. Surv.*, 53(5):99:1–99:38.

Verna Dankers, Christopher Lucas, and Ivan Titov. 2022. Can Transformer be Too Compositional? Analysing Idiom Processing in Neural Machine Translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3608–3626, Dublin, Ireland. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The commitmentbank: Investigating projection in naturally occurring discourse.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint*. ArXiv:2501.12948 [cs].

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. LLM.int8(): 8-bit matrix multiplication for transformers at scale. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, pages 30318–30332, Red Hook, NY, USA. Curran Associates Inc.

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.

J. S. Ehrlich, L. K. Obler, and L. Clark. 1997. Ideational and semantic contributions to narrative production in adults with dementia of the Alzheimer's type. *Journal of Communication Disorders*, 30(2):79–98; quiz 98–99.

Jeffrey L. Elman. 1990. Finding Structure in Time. *Cognitive Science*, 14(2):179–211.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *ICML'24*, pages 12606–12633, Vienna, Austria. JMLR.org.

Ahmad Faiz, Sotaro Kaneda, Ruhan Wang, Rita Chukwunyere Osi, Prateek Sharma, Fan Chen, and Lei Jiang. 2024. LLMCarbon: Modeling the end-to-end carbon footprint of large language models. In *The Twelfth International Conference on Learning Representations*.

Samin Fakharian and Paul Cook. 2021. Contextualized Embeddings Encode Monolingual and Cross-lingual Knowledge of Idiomaticity. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, pages 23–32, Online. Association for Computational Linguistics.

Chen Fang, Panuwat Janwattanapong, Harold Martin, Mercedes Cabrerizo, Armando Barreto, David Loewenstein, Ranjan Duara, and Malek Adjouadi. 2017. Computerized neuropsychological assessment in mild cognitive impairment based on natural language processing-oriented feature extraction. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 543–546.

J. R. Firth. 1957. *Papers in Linguistics 1934-1951.* Oxford University Press, London.

Marshal F. Folstein, Susan E. Folstein, and Paul R. McHugh. 1975. "Mini-mental state": A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12(3):189–198.

K. E. Forbes-McKay and A. Venneri. 2005. Detecting subtle spontaneous language decline in early Alzheimer's disease with a picture description task. *Neurological Sciences: Official Journal of the Italian Neurological Society and of the Italian Society of Clinical Neurophysiology*, 26(4):243–254.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2023. GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers. ICLR 2023. ArXiv:2210.17323 [cs].

Kathleen C. Fraser, Jed A. Meltzer, Frank Rudzicz, and Peter Garrard. 2016. Linguistic Features Identify Alzheimer's Disease in Narrative Speech. *Journal of Alzheimer's Disease*, 49(2):407–422. Publisher: SAGE Publications.

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021a. Assessing the Representations of Idiomaticity in Vector Models with a Noun Compound Dataset Labeled at Type and Token Levels. 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 2730–2741, Online. Association for Computational Linguistics.

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021b. Probing for idiomaticity in vector space models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3551–3564, Online. Association for Computational Linguistics.

Harold Goodglass and Edith Kaplan. 1983. *The Assessment of Aphasia and Related Disorders*, 2 edition. Lea & Febiger, Philadelphia, PA.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 Herd of Models. *arXiv preprint*. ArXiv:2407.21783 [cs].

Yuling Gu, Yao Fu, Valentina Pyatkin, Ian Magnusson, Bhavana Dalvi Mishra, and Peter Clark. 2022. Just-DREAM-about-it: Figurative Language Understanding with DREAM-FLUTE. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 84–93, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. MAGPIE: A Large Corpus of Potentially Idiomatic Expressions. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 279–287, Marseille, France. European Language Resources Association.

Hessel Haagsma, Malvina Nissim, and Johan Bos. 2019. Casting a Wide Net: Robust Extraction of Potentially Idiomatic Expressions.

Nina Hallowell, Shirlene Badger, Aurelia Sauerbrei, Christoffer Nellåker, and Angeliki Kerasidou. 2022. "I don't think people are ready to trust these algorithms at face value": trust and the use of machine learning algorithms in the diagnosis of rare disease. *BMC Medical Ethics*, 23(1):112.

Zellig S. Harris. 1954. Distributional Structure. *WORD*, 10(2-3):146–162.

Reyhaneh Hashempour and Aline Villavicencio. 2020. Leveraging Contextual Embeddings and Idiom Principle for Detecting Idiomaticity in Potentially Idiomatic Expressions. In *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*, pages 72–80, Online. Association for Computational Linguistics.

Wei He, Marco Idiart, Carolina Scarton, and Aline Villavicencio. 2024. Enhancing Idiomatic Representation in Multiple Languages via an Adaptive Contrastive Triplet Loss. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12473–12485, Bangkok, Thailand. Association for Computational Linguistics.

Wei He, Tiago Kramer Vieira, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2025. Investigating Idiomaticity in Word Representations. *Computational Linguistics*, pages 1–51.

Jonathan Heitz, Gerold Schneider, and Nicolas Langer. 2025. Linguistic Features Extracted by GPT-4 Improve Alzheimer‘s Disease Detection based on Spontaneous Speech. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1850–1864, Abu Dhabi, UAE. Association for Computational Linguistics.

Laura Hernández-Domínguez, Sylvie Ratté, Gerardo Sierra-Martínez, and Andrés Roche-Bergua. 2018. Computer-based evaluation of alzheimer's disease and mild cognitive impairment patients during a picture description task. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 10:260–268.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, and 3 others. 2022. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, pages 30016–30030, Red Hook, NY, USA. Curran Associates Inc.

Lilian Hübner, Fernanda Loureiro, Anderson Smidarle, Bruna Tessaro, Ellen C. Siqueira, Gislaine Jerônimo, Thais Dias de Quadros, Vitor Garcia, and Renata

Kochhann. 2019. *Bateria de Avaliação da Linguagem no Envelhecimento (BALE)*, pages 188–218.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv preprint.* ArXiv:2310.06825 [cs].

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models. *arXiv preprint.* ArXiv:2001.08361 [cs].

Gitit Kavé and Ayelet Dassa. 2018. Severity of Alzheimer's disease and language features in picture descriptions. *Aphasiology*, 32(1):27–40. Place: United Kingdom Publisher: Taylor & Francis.

D. Kempler, S. Curtiss, and C. Jackson. 1987. Syntactic preservation in Alzheimer's disease. *Journal of Speech and Hearing Research*, 30(3):343–350.

Filip Klubička, Vasudevan Nedumpozhimana, and John Kelleher. 2023. Idioms, Probing and Dangerous Things: Towards Structural Probing for Idiomaticity in Vector Space. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 45–57, Dubrovnik, Croatia. Association for Computational Linguistics.

Ioannis Korkontzelos, Torsten Zesch, Fabio Massimo Zanzotto, and Chris Biemann. 2013. SemEval-2013 Task 5: Evaluating Phrasal Semantics. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 39–47, Atlanta, Georgia, USA. Association for Computational Linguistics.

Roman Koshkin, Katsuhito Sudoh, and Satoshi Nakamura. 2024. TransLLaMa: LLM-based Simultaneous Translation System. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 461–476, Miami, Florida, USA. Association for Computational Linguistics.

Murathan Kurfalı and Robert Östling. 2020. Disambiguation of Potentially Idiomatic Expressions with Contextual Embeddings. In *Proceedings of the Joint Workshop on*

Multiword Expressions and Electronic Lexicons, pages 85–94, online. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Jinqi Lai, Wensheng Gan, Jiayang Wu, Zhenlian Qi, and Philip S. Yu. 2024. Large language models in law: A survey. *AI Open*, 5:181–196.

Thomas K. Landauer and Susan T. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240. Place: US Publisher: American Psychological Association.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Alessandro Lenci. 2018. Distributional Models of Word Meaning. In *Annual Review of Linguistics*, volume 4, pages 151–171. ISSN: 2333-9683, 2333-9691 Issue: 1 Journal Abbreviation: Annu. Rev. Linguist.

Alessandro Lenci and Magnus Sahlgren. 2023. *Distributional Semantics*. Studies in Natural Language Processing. Cambridge University Press, Cambridge.

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.

Bofang Li, Tao Liu, Zhe Zhao, Buzhou Tang, Aleksandr Drozd, Anna Rogers, and Xiaoyong Du. 2017. Investigating Different Syntactic Context Types and Context Representations for Learning Word Embeddings. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2421–2431, Copenhagen, Denmark. Association for Computational Linguistics.

Changye Li, David Knopman, Weizhe Xu, Trevor Cohen, and Serguei Pakhomov. 2022. GPT-D: Inducing Dementia-related Linguistic Anomalies by Deliberate Degradation of Artificial Neural Language Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1866–1877, Dublin, Ireland. Association for Computational Linguistics.

Jia Li, Ge Li, Yongmin Li, and Zhi Jin. 2025. Structured Chain-of-Thought Prompting for Code Generation. *ACM Trans. Softw. Eng. Methodol.*, 34(2):37:1–37:23.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. Textbooks Are All You Need II: phi-1.5 technical report. *arXiv preprint.* ArXiv:2309.05463 [cs].

G. Libben. 1998. Semantic transparency in the processing of compounds: consequences for representation, processing, and impairment. *Brain and Language*, 61(1):30–44.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual Instruction Tuning. *Advances in Neural Information Processing Systems*, 36:34892–34916.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023b. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Comput. Surv.*, 55(9):195:1–195:35.

Yang Liu and Mirella Lapata. 2019. Text Summarization with Pretrained Encoders. 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 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint.* ArXiv:1907.11692 [cs].

Pan Lu, Liang Qiu, Wenhao Yu, Sean Welleck, and Kai-Wei Chang. 2023. A Survey of Deep Learning for Mathematical Reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14605–14631, Toronto, Canada. Association for Computational Linguistics.

Sheng Lu, Irina Bigoulaeva, Rachneet Sachdeva, Harish Tayyar Madabushi, and Iryna Gurevych. 2024. Are Emergent Abilities in Large Language Models just In-Context Learning? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5098–5139, Bangkok, Thailand. Association for Computational Linguistics.

Kevin Lund and Curt Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments & Computers*, 28(2):203–208. Place: US Publisher: Psychonomic Society.

Saturnino Luz, Fasih Haider, Sofia De La Fuente, Davida Fromm, and Brian MacWhinney. 2020. Alzheimer's Dementia Recognition Through Spontaneous Speech: The ADReSS Challenge. *Interspeech 2020*, pages 2172–2176. Conference Name: Interspeech 2020 Publisher: ISCA.

Kelly Lyons, Kemper , Susan, Labarge , Emily, Ferraro , F. Richard, Balota , David, , and Martha Storandt. 1994. Oral language and alzheimer's disease: A reduction in syntactic complexity. *Aging, Neuropsychology, and Cognition*, 1(4):271–281.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. In *The 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL 2023)*.

Bárbara Luzia Covatti Malcorra, Alberto Osa García, Karine Marcotte, Hanna de Paz, Lucas Porcello Schilling, Irênio Gomes da Silva Filho, Ricardo Soder, Alexandre da Rosa Franco, Fernanda Loureiro, and Lilian Cristine Hübner. 2024. Exploring Spoken Discourse and Its Neural Correlates in Women With Alzheimer's Disease With Low Levels of Education and Socioeconomic Status. *American Journal of Speech-Language Pathology*, 33(2):893–911.

Bárbara Luzia Covatti Malcorra, Natália Bezerra Mota, Janaina Weissheimer, Lucas Porcello Schilling, Maximiliano Agustin Wilson, and Lilian Cristine Hübner. 2021. Low Speech Connectedness in Alzheimer's Disease is Associated with Poorer Semantic Memory Performance. *Journal of Alzheimer's disease: JAD*, 82(3):905–912.

Oren Melamud, Jacob Goldberger, and Ido Dagan. 2016. context2vec: Learning Generic Context Embedding with Bidirectional LSTM. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 51–61, Berlin, Germany. Association for Computational Linguistics.

Maggie Mi, Aline Villavicencio, and Nafise Sadat Moosavi. 2025. Rolling the DICE on Idiomaticity: How LLMs Fail to Grasp Context. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7314–7332, Vienna, Austria. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Jeff Mitchell and Mirella Lapata. 2010. Composition in Distributional Models of Semantics. *Cognitive Science*, 34(8):1388–1429.

Andriy Mnih and Geoffrey E Hinton. 2008. A Scalable Hierarchical Distributed Language Model. In *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc.

Rosamund Moon. 1998. *Fixed Expressions and Idioms in English: A Corpus-based Approach*. Clarendon Press.

Kimberly Diggle Mueller, Rebecca L. Koscik, Lyn S. Turkstra, Sarah K. Riedeman, Asenath LaRue, Lindsay R. Clark, Bruce Hermann, Mark A. Sager, and Sterling C. Johnson. 2016. Connected Language in Late Middle-Aged Adults at Risk for Alzheimer's Disease. *Journal of Alzheimer's disease : JAD*, 54(4):1539–1550.

Navnita Nandakumar, Timothy Baldwin, and Bahar Salehi. 2019. How Well Do Embedding Models Capture Non-compositionality? A View from Multiword Expressions. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 27–34, Minneapolis, USA. Association for Computational Linguistics.

Maryam Nasiri, Saeideh Moayedfar, Mehdi Purmohammad, and Leila Ghasisin. 2022. Investigating sentence processing and working memory in patients with mild Alzheimer and elderly people. *PloS One*, 17(11):e0266552.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, Renqian Luo, Scott Mayer McKinney, Robert Osazuwa Ness, Hoifung Poon, Tao Qin, Naoto Usuyama, Chris White, and Eric Horvitz. 2023. Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine. *arXiv preprint*. ArXiv:2311.16452 [cs].

Geoffrey Nunberg, Ivan A. Sag, and Thomas Wasow. 1994. Idioms. In Stephen Everson, editor, *Language: Companions to Ancient Thought, Vol. 3*, pages 491–538.

United Nations Department of Economic and Population Division Social Affairs. 2022. World population prospects 2022: Summary of results.

OpenAI. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI. 2025a. Learning to reason with LLMs.

OpenAI. 2025b. OpenAI o3 and o4-mini System Card.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024a. GPT-4 Technical Report. *arXiv preprint*. ArXiv:2303.08774 [cs].

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, A. J. Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, and 400 others. 2024b. GPT-4o System Card. *arXiv preprint*. ArXiv:2410.21276 [cs].

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

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.

Barbara H. Partee. 1984. Compositionality. In *Varieties of Formal Semantics*, pages 281–311. Foris, Dordrecht.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. 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 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Joshua C. Peterson, Dawn Chen, and Thomas L. Griffiths. 2020. Parallelograms revisited: Exploring the limitations of vector space models for simple analogies. *Cognition*, 205:104440.

Ulla Petti, Simon Baker, and Anna Korhonen. 2020. A systematic literature review of automatic Alzheimer's disease detection from speech and language. *Journal of the American Medical Informatics Association: JAMIA*, 27(11):1784–1797.

Dylan Phelps. 2022. drsphelps at SemEval-2022 Task 2: Learning idiom representations using BERTRAM. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 158–164, Seattle, United States. Association for Computational Linguistics.

Dylan Phelps, Xuan-Rui Fan, Edward Gow-Smith, Harish Tayyar Madabushi, Carolina Scarton, and Aline Villavicencio. 2022. Sample Efficient Approaches for Idiomaticity Detection. In *Proceedings of the 18th Workshop on Multiword Expressions*

@LREC2022, pages 105–111, Marseille, France. European Language Resources Association.

Dylan Phelps, Thomas M. R. Pickard, Maggie Mi, Edward Gow-Smith, and Aline Villavicencio. 2024. Sign of the Times: Evaluating the use of Large Language Models for Idiomaticity Detection. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 178–187, Torino, Italia. ELRA and ICCL.

Dylan Phelps, Rodrigo Wilkens, Edward Gow-Smith, Thomas Pickard, Maggie Mi, and Aline Villavicencio. 2025. Stands to reason: Investigating the effect of reasoning on idiomaticity detection. *Preprint*, arXiv:2508.13365.

Thomas Pickard, Aline Villavicencio, Maggie Mi, Wei He, Dylan Phelps, and Marco Idiart. 2025. SemEval-2025 Task 1: AdMIRe - Advancing Multimodal Idiomaticity Representation. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2597–2609, Vienna, Austria. Association for Computational Linguistics.

Yuval Pinter, Robert Guthrie, and Jacob Eisenstein. 2017. Mimicking Word Embeddings using Subword RNNs. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 102–112, Copenhagen, Denmark. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

XiPeng Qiu, TianXiang Sun, YiGe Xu, YunFan Shao, Ning Dai, and XuanJing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10):1872–1897.

Qwen. 2024. Qwen2.5: A Party of Foundation Models! Section: blog.

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, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. Qwen2.5 Technical Report. *arXiv preprint*. ArXiv:2412.15115 [cs].

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training.

Alec Radford, Jeff Wu, R. Child, D. Luan, Dario Amodei, and I. Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Preprint*, arXiv:2305.18290.

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 learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1):140:5485–140:5551.

Livy Real, Erick Fonseca, and Hugo Gonçalo Oliveira. 2020. The ASSIN 2 Shared Task: A Quick Overview. In *Computational Processing of the Portuguese Language: 14th International Conference, PROPOR 2020, Evora, Portugal, March 2–4, 2020, Proceedings*, pages 406–412, Berlin, Heidelberg. Springer-Verlag.

Jamie Reilly, Joshua Troche, and Murray Grossman. 2011. Language processing in dementia. *The handbook of Alzheimer's disease and other dementias*, pages 336–368.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Erica dos Santos Rodrigues, Lilian Cristine Hübner, and Ana Paula Rodrigues Bastos. 2024. Por que observar critérios psicolinguísticos? : Reflexões para a construção de tarefas linguísticas em pesquisas e práticas em Saúde e Educação. *Letrônica*, 17(1):e46985–e46985. Number: 1.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics*, 8:842–866. Place: Cambridge, MA Publisher: MIT Press.

David E. Rumelhart and Adele A. Abrahamson. 1973. A model for analogical reasoning. *Cognitive Psychology*, 5(1):1–28. Place: Netherlands Publisher: Elsevier Science.

Romy Räling, Sandra Hanne, Astrid Schröder, Carla Keßler, and Isabell Wartenburger. 2017. Judging the Animacy of Words: The Influence of Typicality and Age of Acquisition in a Semantic Decision Task. *Quarterly Journal of Experimental Psychology (2006)*, 70(10):2094–2104.

Arkadiy Saakyan, Tuhin Chakrabarty, Debanjan Ghosh, and Smaranda Muresan. 2022. A Report on the FigLang 2022 Shared Task on Understanding Figurative Language. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 178–183, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

E. M. Saffran, R. S. Berndt, and M. F. Schwartz. 1989. The quantitative analysis of agrammatic production: procedure and data. *Brain and Language*, 37(3):440–479.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. In *Computational Linguistics and Intelligent Text Processing*, pages 1–15, Berlin, Heidelberg. Springer.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2025. A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications. *arXiv preprint*. ArXiv:2402.07927 [cs].

Seyed Ahmad Sajjadi, Karalyn Patterson, Michal Tomek, and Peter J. Nestor. 2012. Abnormalities of connected speech in semantic dementia vs Alzheimer's disease. *Aphasiology*, 26(6):847–866. Publisher: Routledge _eprint: https://doi.org/10.1080/02687038.2012.654933.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2015. A Word Embedding Approach to Predicting the Compositionality of Multiword Expressions. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 977–983, Denver, Colorado. Association for Computational Linguistics.

Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. 2023. In-Context Impersonation Reveals Large Language Models' Strengths and Biases. *Advances in Neural Information Processing Systems*, 36:72044–72057.

Victoria Sanborn, Rachel Ostrand, Jeffrey Ciesla, and John Gunstad. 2022. Automated assessment of speech production and prediction of mci in older adults. *Applied Neuropsychology: Adult*, 29(5):1250–1257.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint*. ArXiv:1910.01108 [cs].

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are emergent abilities of large language models a mirage? In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, pages 55565–55581, Red Hook, NY, USA. Curran Associates Inc.

Timo Schick and Hinrich Schütze. 2019. Attentive Mimicking: Better Word Embeddings by Attending to Informative Contexts. 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 489–494, Minneapolis, Minnesota. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2020. BERTRAM: Improved Word Embeddings Have Big Impact on Contextualized Model Performance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3996–4007, Online. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021a. Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021b. It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, and 1 others. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*.

Vered Shwartz and Ido Dagan. 2019. Still a Pain in the Neck: Evaluating Text Representations on Lexical Composition. *Transactions of the Association for Computational Linguistics*, 7:403–419.

John Sinclair. 1991. *Corpus, Concordance, Collocation.* Oxford University Press. Google-Books-ID: Vcx5AAAAIAAJ.

Răzvan-Alexandru Smădu, David-Gabriel Ion, Dumitru-Clementin Cercel, Florin Pop, and Mihaela-Claudia Cercel. 2024. Investigating Large Language Models for Complex Word Identification in Multilingual and Multidomain Setups. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16764–16800, Miami, Florida, USA. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MPNet: masked and permuted pre-training for language understanding. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, pages 16857–16867, Red Hook, NY, USA. Curran Associates Inc.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In *Intelligent Systems*, Lecture Notes in Computer Science, pages 403–417, Cham. Springer International Publishing.

Irena Spasic and Goran Nenadic. 2020. Clinical Text Data in Machine Learning: Systematic Review. *JMIR Medical Informatics*, 8(3):e17984. Company: JMIR Medical Informatics Distributor: JMIR Medical Informatics Institution: JMIR Medical Informatics Label: JMIR Medical Informatics Publisher: JMIR Publications Inc., Toronto, Canada.

Mark Stevenson and Yorick Wilks. 2005. Word-Sense Disambiguation. In Ruslan Mitkov, editor, *The Oxford Handbook of Computational Linguistics*, 1st edition, pages 249–265. Oxford University Press.

Kevin Stowe, Prasetya Utama, and Iryna Gurevych. 2022. IMPLI: Investigating NLI Models' Performance on Figurative Language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5375–5388, Dublin, Ireland. Association for Computational Linguistics.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and Policy Considerations for Deep Learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Harish Tayyar Madabushi, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2022. SemEval-2022 Task 2: Multilingual Idiomaticity Detection and Sentence Embedding. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 107–121, Seattle, United States. Association for Computational Linguistics.

Harish Tayyar Madabushi, Edward Gow-Smith, Carolina Scarton, and Aline Villavicencio. 2021. AStitchInLanguageModels: Dataset and Methods for the Exploration of Idiomaticity in Pre-Trained Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3464–3477, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1331 others.

2024. Gemini: A Family of Highly Capable Multimodal Models. *arXiv preprint*. ArXiv:2312.11805 [cs].

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. *arXiv preprint*. ArXiv:2207.04672 [cs].

C. K. Thompson, K. J. Ballard, M. E. Tait, S. Weintraub, and M. Mesulam. 1997. Patterns of language decline in non-fluent primary progressive aphasia. *Aphasiology*, 11(4-5):297–321. Publisher: Routledge _eprint: https://doi.org/10.1080/02687039708248473.

Nell Thornton, Ahmed Binesmael, Tim Horton, and Tom Hardie. 2024. AI in health care: what do the public and NHS staff think? - The Health Foundation. Technical report, The Health Foundation.

Debra A. Titone and Cynthia M. Connine. 1999. On the compositional and noncompositional nature of idiomatic expressions. *Journal of Pragmatics*, 31(12):1655–1674.

T. N. Tombaugh and N. J. McIntyre. 1992. The mini-mental state examination: a comprehensive review. *Journal of the American Geriatrics Society*, 40(9):922–935.

C. K. Tomoeda, K. A. Bayles, M. W. Trosset, T. Azuma, and A. McGeagh. 1996. Cross-sectional analysis of Alzheimer disease effects on oral discourse in a picture description task. *Alzheimer Disease and Associated Disorders*, 10(4):204–215.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint*. ArXiv:2302.13971 [cs].

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, pages 74952–74965. Curran Associates, Inc.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Shubham Vatsal and Harsh Dubey. 2024. A Survey of Prompt Engineering Methods in Large Language Models for Different NLP Tasks. *arXiv preprint*. ArXiv:2407.12994 [cs].

David Vilares, Marcos Garcia, and Carlos Gómez-Rodríguez. 2021. Bertinho: Galician BERT Representations. *Procesamiento del Lenguaje Natural*, pages 13–26. ArXiv:2103.13799 [cs].

Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. 2024. Position: will we run out of data? limits of LLM scaling based on human-generated data. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *ICML'24*, pages 49523–49544, Vienna, Austria. JMLR.org.

Jet M. J. Vonk, Vincent Bouteloup, Jean-François Mangin, Bruno Dubois, Frédéric Blanc, Audrey Gabelle, Mathieu Ceccaldi, Cédric Annweiler, Pierre Krolak-Salmon, Catherine Belin, Thérèse Rivasseau-Jonveaux, Adrien Julian, François Sellal, Eloi Magnin, Marie Chupin, Marie-Odile Habert, Geneviève Chêne, Carole Dufouil, and MEMENTO cohort Study Group. 2020. Semantic loss marks early Alzheimer's disease-related neurodegeneration in older adults without dementia. *Alzheimer's & Dementia (Amsterdam, Netherlands)*, 12(1):e12066.

Takashi Wada, Yuji Matsumoto, Timothy Baldwin, and Jey Han Lau. 2023. Unsupervised Paraphrasing of Multiword Expressions. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4732–4746, Toronto, Canada. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. SuperGLUE: a stickier benchmark for general-purpose language understanding systems. In *Proceedings of the*

33rd International Conference on Neural Information Processing Systems, 294, pages 3266–3280. Curran Associates Inc., Red Hook, NY, USA.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024. MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent Abilities of Large Language Models. *arXiv preprint*. ArXiv:2206.07682 [cs].

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, pages 24824–24837, Red Hook, NY, USA. Curran Associates Inc.

Myron F. Weiner, Katherine E. Neubecker, Mary E. Bret, and Linda S. Hynan. 2008. Language in Alzheimer's Disease. *The Journal of clinical psychiatry*, 69(8):1223–1227.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Lowri Williams, Christian Bannister, Michael Arribas-Ayllon, Alun Preece, and Irena Spasić. 2015. The role of idioms in sentiment analysis. *Expert Systems with Applications*, 42(21):7375–7385.

Raphael Wittenberg, Bo Hu, Carol Jagger, Andrew Kingston, Martin Knapp, Adelina Comas-Herrera, Derek King, Amritpal Rehill, and Sube Banerjee. 2020. Projections of care for older people with dementia in England: 2015 to 2040. *Age and Ageing*, 49(2):264–269.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yonghui Wu, M. Schuster, Z. Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, M. Krikun, Yuan Cao, Qin Gao, Klaus Macherey, J. Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Y. Kato, Taku Kudo, H. Kazawa, and 12 others. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *ArXiv*.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. A Paradigm Shift in Machine Translation: Boosting Translation Performance of Large Language Models. *arXiv preprint*. ArXiv:2309.11674 [cs] version: 2.

Hao Yang, Ruihan Wang, Changyu Wang, Hui Gao, Hanlin Cai, Fengying Zhang, Jialin Liu, and Siru Liu. 2023. GPT-4 and Neurologists in Screening for Mild Cognitive Impairment in the Elderly: A Comparative Analysis Study. Repository: Neurology.

Lang Yu and Allyson Ettinger. 2020. Assessing Phrasal Representation and Composition in Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4896–4907, Online. Association for Computational Linguistics.

Ziheng Zeng and Suma Bhat. 2021. Idiomatic Expression Identification using Semantic Compatibility. *Transactions of the Association for Computational Linguistics*, 9:1546–1562. Place: Cambridge, MA Publisher: MIT Press.

Ziheng Zeng and Suma Bhat. 2023. Unified Representation for Non-compositional and Compositional Expressions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11696–11710, Singapore. Association for Computational Linguistics.

Haopeng Zhang, Philip S. Yu, and Jiawei Zhang. 2025. A Systematic Survey of Text Summarization: From Statistical Methods to Large Language Models. *ACM Comput. Surv.* Just Accepted.

Jinman Zhao, Sidharth Mudgal, and Yingyu Liang. 2018. Generalizing Word Embeddings using Bag of Subwords. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 601–606, Brussels, Belgium. Association for Computational Linguistics.

Tony Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.