University of Sheffield

Adaptive Text Summarisation for Broadening Access to Technical Texts

Tomas Goldsack

Supervisors: Chenghua Lin & Carolina Scarton

A report submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Speech and Language Technologies and their Applications

in the

Department of Computer Science

July 22, 2025

Abstract

Technical and specialist texts play a crucial role in documenting and advancing various fields, institutions, and academic disciplines. However, these texts often assume a certain level of prior knowledge, relying on domain-specific concepts and language that can be difficult for non-experts to understand. Text Summarisation techniques aim to retrieve and express the key points of a text within a coherent, informative summary. When trained or prompted appropriately, these techniques can also make textual information more accessible by adapting content to suit different audiences. Therefore, when applied to technical documents, summarisation has the potential to help bridge the gap between experts and non-experts, expanding access to specialised knowledge. However, the complexity and unique characteristics of such texts present significant challenges for automatic processing, limiting the effectiveness of modern summarisation systems.

This thesis explores strategies to enhance the application of summarisation models to technical texts. Furthermore, significant emphases are placed on broadening accessibility to non-technical audiences, improving content understanding through detailed analyses, and exploring how relevant external knowledge can be leveraged to improve model outputs. It is structured around five individual publications that separately investigate several pertinent directions. Publication I: Domain-driven and Discourse-guided Scientific Summarisation analyses the structure of research paper abstracts and introduces a lightweight approach for abstract generation. Publication II: Making Science Simple: Corpora for the Lay Summarisation of Scientific Literature introduces, analyses, and benchmarks two high-quality datasets for the Lay Summarisation of research articles, enabling further research on the development of Summarisation models that can cater to a non-expert audience. Utilising one

of these datasets, Publication III: Enhancing Biomedical Lay Summarisation with External Knowledge Graphs investigates the potential benefits of incorporating external knowledge (in the form of knowledge graphs) within Lay Summarisation models. Similarly, Publication IV: Leveraging Large Language Models for Zero-shot Lay Summarisation in Biomedicine and Beyond proposes a prompting technique for Lay Summarisation with Large Language Models (LLMs) that leverages new-found zero-shot abilities and knowledge of real-world industrial practices. Finally, Publication V: From Facts to Insights: A Study on the Generation and Evaluation of Analytical Reports for Deciphering Earnings Calls explores a conversational multi-agent approach for the novel task of analytical report generation with LLMs.

Contents

0.1 Glossary of Key NLP Terms

Abstractive Summarisation An approach to text summarisation where the system rewrites content and ideas using novel phrasing.

Attention Network Any neural model that uses attention to decide which input parts matter most when producing each output element. Transformers are the best-known example.

Coherent (summary) One whose sentences flow logically and grammatically, forming a connected, easy-to-follow narrative.

Earnings Call A live or recorded presentation where a publicly traded company discusses its quarterly or annual financial results with analysts and investors; transcripts are a common testbed for summarisation systems.

Extractive Summarisation An approach to text summarisation lifts whole sentences from the source text and stitches them together.

Factuality The degree to which a generated summary stays consistent with the source article, with no invented details or *hallucinations*.

Fine-tuning Continuing to train a pre-trained model on a small, task-specific dataset so it specialises.

Graph Attention Network (GAT) A neural network that applies attention over the neighbours of each node in a graph, letting it learn which connections carry the most useful information.

Interpretable (approach) A model or method whose inner workings can be understood by humans—important for trust and debugging.

Knowledge Graph (KG) A network of facts where nodes are concepts (e.g. “ACE2 protein”) and edges express relationships (e.g. “is-part-of”).

Large Language Model (LLM) An enormous neural network trained on vast text collections that can answer questions or write passages from a prompt.

Lay Summarisation A specialised form of abstractive summarisation that rewrites a technical paper so that non-experts can understand it.

Lightweight (approach) A method designed to run quickly or on limited hardware (e.g. laptops or edge devices) without heavy computation.

MetaMap A tool from the U.S. National Library of Medicine that spots UMLS concepts in free text and links biomedical terms in each article to their plain-language definitions.

Pointer-Generator Network A summarisation model that can either *generate* new words from its vocabulary or *copy* ("point to") words directly from the source, giving it both creativity and factual faithfulness.

Prompting Giving the LLM carefully worded instructions or examples so it performs the desired task.

Recurrent Neural Network (RNN) A family of neural models (including LSTMs and GRUs) that processes sequences one step at a time, passing information forward through hidden states.

Rhetorical (Discourse) Structure The typical order of information in a paper (background → method → results, etc.).

Sequence-to-Sequence (Seq2Seq) Model A model that reads an input sequence (e.g. a paper) and produces an output sequence (e.g. a summary). Early versions used Recurrent Neural Networks; newer ones use Transformers.

Term Frequency–Inverse Document Frequency (TF–IDF) A classic weighting scheme that scores how important a word is to a document: frequent in the document (high TF) but rare across the corpus (high IDF) means high weight.

Text Summarisation Any computational technique that produces a much shorter version of a long document describing the key points.

Unified Medical Language System (UMLS) A large, curated biomedical vocabulary.

XML (Extensible Markup Language) A tree-structured text format for tagging and exchanging data; still common in scientific and financial document pipelines.

Zero-shot Using an LLM on a task it has never seen before, relying only on a cleverly crafted prompt.

Chapter 1

Introduction

A text is considered *technical* when a certain level of prior knowledge in a specific field, industry, or science is necessary to fully comprehend its content (Hamlin et al., 2015). These texts are widespread and contain critical knowledge that supports the preservation and advancement of key global institutions, including those in science, finance, medicine, and law. In today's information-driven era, technical articles are being published at a substantial rate, with science and engineering output averaging 5% from 2017-2020, reaching approximately 2.9 million publications in 2020 (National Science Board (NSB), 2021). For those seeking to locate specific information, this ever-expanding reservoir of specialised knowledge can present a significant challenge. The difficulty is even greater for individuals without the necessary technical background, as they often encounter complex jargon and unfamiliar formal structures that obscure the content's meaning (Howe and Wogalter, 1994; Kuehne and Olden, 2015; King et al., 2017). Such situations are not uncommon — for instance, it is easy to imagine a person without legal expertise who may need to understand the details of a housing contract, or someone diagnosed with an illness who might want to read about the latest research on a potential cure.

Automatic summarisation approaches specialise in expressing the salient information from a longer text and in a short digestible format, allowing a reader to grasp its key content without reading the entire text (El-Kassas et al., 2021a). Summaries and the approaches that produce them are typically classified as either *extractive* or *abstractive*, where extractive summaries are constructed using sentences

Technical Abstract
The virus SARS-CoV-2 can exploit biological vulnerabilities (e.g. host proteins) in susceptible hosts that predispose to the development of severe COVID-19. To identify host proteins that may contribute to the risk of severe COVID-19, we undertook proteome-wide genetic colocalisation tests, and polygenic (pan) and cis-Mendelian randomisation analyses leveraging publicly available protein and COVID-19 datasets...

Lay Summary
Individuals who become infected with the virus that causes COVID-19 can experience a wide variety of symptoms. These can range from no symptoms or minor symptoms to severe illness and death. Key demographic factors, such as age, gender and race, are known to affect how susceptible an individual is to infection. However, molecular factors, such as unique gene mutations and gene expression levels can also have a major impact on patient responses by affecting the levels of proteins in the body...

Figure 1.1: *The first few sentences of the abstract and lay summary of a scientific article, illustrating differences in the language and focus on explaining relevant background information. Figure from Goldsack et al. (2022a).*

from the source article, and abstractive summaries describe the key points of a text using alternative words and phrasing. As such, abstractive summarisation approaches in particular have the potential to help broaden access to technical texts to wider audiences by adopting alternative language and focusing on the most relevant content (Dudy et al., 2021). Figure 1.1 provides an example of how an abstractive summary can be adapted to the target audience, presenting the first few sentences of two different summaries of a biomedical research article. One of these summaries is intended for a technical audience (Technical Abstract), and the other, a non-technical audience (Lay Summary), illustrating key differences in terms of both the summary content and the language used. The research and development of summarisation approaches that are able to broaden accessibility to technical texts in such a way is the central motivation of this thesis.

Automatic abstractive summarisation has largely been made possible by the rise of supervised neural sequence-to-sequence models, starting with Recurrent Neural Networks (Rush et al., 2015; Chopra et al., 2016; Nallapati et al., 2016) and progressing to pre-trained transformer-based language models that were introduced around the start of this PhD (Lewis et al., 2020; Beltagy et al., 2020; Raffel et al., 2020). For many years following the introduction of sequence-to-sequence models, the

Natural Language Processing (NLP) community largely framed the research of Text Summarisation as the generation of a few-sentence summary based on a relatively short input article, with news articles providing the primary source of large-scale parallel data required to train them (Sandhaus, 2008; Nallapati et al., 2016; Paulus et al., 2018). The application of Text Summarisation to technical texts subsequently emerged as a popular alternative, with large-scale parallel datasets being proposed for domains such as scientific (Cohan et al., 2018; Yasunaga et al., 2019; Cachola et al., 2020) and legal (Sharma et al., 2019a; Kornilova and Eidelman, 2019) texts, allowing for the training (or fine-tuning) of these supervised neural approaches. However, the vast majority of these datasets contained only technical summaries, restricting the applicability of models trained on them to expert audiences. Those few which contained non-expert summaries (Chandrasekaran et al., 2020; Guo et al., 2021; Zaman et al., 2020) consisted of at most only a few thousand article-summary pairs, making them relatively small in scale compared to popular datasets used within the broader field, which regularly contained >100,000 pairs. Furthermore, these datasets were typically limited in their scope, often catering to only a specific form of input article (e.g., systematic reviews). Therefore, in addition to the development of models capable of generating non-expert summaries, this thesis also aims to enable others to develop such approaches through the introduction of novel datasets.

Importantly, during the course of this PhD, the field of NLP was significantly reshaped by the development of Large Language Models (LLMs). These scaled-up transformer-based language models, trained on vast quantities of text data, have exhibited strong general capabilities across numerous downstream tasks, driven by newfound emergent behaviors (Kaplan et al., 2020; Brown et al., 2020; Wei et al., 2021). In the context of summarisation, advanced zero-shot capabilities have enabled strong performance without the need for parallel training data (Goyal et al., 2022). Beyond serving as a highly competitive alternative to fine-tuned models, this shift opened the door to the investigation of more summarisation-based tasks that were previously restricted due to data or complexity limitations (Ahmed and Devanbu, 2023; Chang et al., 2024). This is also reflected in this recent technical datasets, with recent benchmarks such as CLSum (Liu et al., 2024) in the legal domain and (Cai et al., 2025) in the scientific domain addressing novel low-resource summarisation tasks for which there is a lack of parallel training data. Therefore, this thesis explores various approaches widely used in the NLP community at the time, in addition to leveraging the capabilities of novel state-of-the-art models to explore novel low-resource tasks.

For both early and contemporary abstractive summarisation approaches, technical texts pose a number of challenges. Factors such as their significant length, domain-specific discourse, and standardised formal structures often require additional consideration (Cohan et al., 2018; Sharma et al., 2019b; Yasunaga et al., 2019). Moreover, given that such texts typically assume some level of expertise from their reader, fully comprehending such texts is likely to depend on external domain knowledge that is not explicitly stated, presenting an additional challenge when looking to generate summaries for non-expert audiences. For these reasons, a significant emphasis was also placed on the textual analysis of utilised data and on leveraging relevant external knowledge in the research and development of summarisation approaches for technical texts.

1.1 Research Aims and Objectives

Here the research objectives that were addressed within this thesis are described. Overall, the following three objectives provide an underlying motivation behind all of its chapters:

(1) Developing and enabling the development of novel summarisation approaches that allow the contents of technical texts to be accessed by a wider audience. As discussed in this introduction, abstractive summarisation can play an important role in increasing access to technical texts. Despite this, only a limited body of previous research addressed this problem, partly due to a scarcity of high-quality datasets. Therefore, the central aims of this thesis surround the development of summarisation approaches that can broaden access to technical texts, and enable future research on such tasks through the introduction of novel datasets.

(2) Improving our understanding of technical texts and their summaries through in-depth, varied linguistic analyses. In order to know what makes a good summary of a given text, it is essential to understand the intricacies and patterns within existing source and target data. This is especially true of technical texts and summaries, which often conform to a standardised (or at least commonly practiced) structure or rhetoric. Therefore, an emphasis was consistently placed on analysing the diverse properties of textual data, focusing an aspects such as readability, structural discourse, and vocabulary.

(3) Exploring how the use of relevant external knowledge can be used to improve the summarisation of technical texts. A defining characteristic of all technical texts is that readers must possess some level of expertise to comprehend their contents. However, given that these texts are intended for such an audience, they are unlikely to explicitly contain all the background information required to produce a summary for a non-expert audience. Therefore, the potential benefits of utilising knowledge from external sources were often explored within the design and implementation of the methodologies proposed in this thesis.

1.2 Thesis Overview

This section provides a description of the structure and contributions of the thesis. Each of the subsequent five chapters in the thesis corresponds to a publication made over the course of the PhD. The seventh and final chapter provides a conclusion for the thesis. A description of each chapter is provided below.

Chapter 2 provides a detailed analysis of the *rhetorical structures* of abstracts of different scientific domains, and subsequently introduces a lightweight and interpretable abstract generation approach. This chapter represents an initial exploration into the summarisation of scientific articles where, in particular, objective (2) is addressed given the emphasis on analysing this form of technical data. It is based on the publication *"Domain-Driven and Discourse-Guided Scientific Summarisation"* (Goldsack et al., 2023b).

The following research questions are associated with this chapter:

- Does scientific domain of an article has a strong influence over the rhetorical structure of its abstract?
- Can scientific domain be leveraged in the context of abstract generation to improve both the content and structure of the generated output?

Chapter 3 introduces, analyses, and benchmarks two novel datasets for Lay Summarisation in the Biomedical domain, a task that involves the generation of non-expert summaries for scientific research articles. This chapter primarily addresses objective (1), as future work on training models able to generate non-technical summaries is enabled through these datasets. Additionally, objective (2) is also addressed, given the extensive characterisation of the dataset summaries. It is based on the publication *"Making Science Simple: Corpora for the Lay Summarisation of Scientific Literature"* (Goldsack et al., 2022b).

The following research questions are associated with this chapter:

- How do expert-authored lay summaries differ from technical abstracts in terms of readability, rhetorical structure, vocabulary overlap, and abstractiveness?
- How effective are current summarisation methods at producing clear, accurate lay summaries of biomedical research?

Chapter 4 describes the potential benefits of using external knowledge sources within Lay Summarisation approaches and provides an in-depth comparison of three different methods for incorporating knowledge graphs within encoder-decoder Transformer-based models. This chapter partially addresses all three of the research objectives of this thesis, given that it (1) proposed novel approaches for Lay Summarisation that incorporate knowledge graphs; (2) deepens understanding of source articles in the creation and analysis of knowledge graphs; and (3) leverages external knowledge sources to enrich knowledge graph content. It is based on the publication *"Enhancing*

Biomedical Lay Summarisation with External Knowledge Graphs" (Goldsack et al., 2023c).

The following research questions are associated with this chapter:

- Can current approaches for lay summarisation be improved by augmenting source articles with external domain knowledge in the form of knowledge graphs?

Chapter 5 examines the use of LLMs for the task of Lay Summarisation in a zero-shot setting, comparing a simple 1-stage prompt with a novel multi-stage prompting approach. This chapter addresses objectives (1) and (3), as it introduces a novel approach utilising LLMs to generate non-expert summaries that is based on external knowledge of real-world practices. It is based on the publication *"Leveraging Large Language Models for Zero-shot Lay Summarisation in Biomedicine and Beyond"* (Goldsack et al., 2025).

The following research questions are associated with this chapter:

- Can zero-shot lay summarisation with LLMs be improved by introducing a prompting framework that emulates real-world practices?

Chapter 6 investigates the potential benefits of adopting a multi-agent approach for the novel task of generating analytical reports of financial *Earnings Call* meetings (ECs), and discusses how LLMs might be used to help evaluate such reports. This chapter takes the opportunity to delve into an entirely unexplored task that was made possible by the capabilities of modern LLMs, with the generation of analytical reports showing significant promise for the application of novel agent-based approaches. Here, all research objectives were again addressed. Specifically, in the generation of an analytical report, the explored approach summarises EC content for potential investors whilst also leveraging external data (objectives 1 and 3). Finally, an in-depth analysis of generated reports is also provided and compared against human-authored reports (objective 2). It is based on the publication *"From Facts to Insights: A Study on the Generation and Evaluation of Analytical Reports for Deciphering Earnings Calls."* (Goldsack et al., 2024b).

The following research questions are associated with this chapter:

- How do generated analytical reports differ from human-authored analytical reports?
- Can a multi-agent approach be used to generate more insightful analytical reports?
- How effective are LLM-based evaluation methods in assessing the quality of analytical reports?

Chapter 7 concludes this thesis, providing a summary of the findings of this thesis, an assessment of the completion of its research questions, and a discussion of possible future directions.

Importantly, each inner chapter (2-6) is self-contained, including a description of the related work specific to the original publication. Additionally, the content of the original publication is unchanged, with only aesthetic edits made to accommodate divergences in formatting. Additional content related to each chapter can be found in the Appendices, which are organised by publication in accordance with the thesis structure.[1]

1.3 Contributions

1.3.1 First-author Published Work

The inner chapters within this thesis correspond to the five first-authored publications. They are list below alongside corresponding author contribution statements.

1. ***Goldsack, T.***, *Zhang, Z., Lin, C., Scarton, C. (2023). Domain-Driven and Discourse-Guided Scientific Summarisation. In: Kamps, J., et al. Advances in Information Retrieval. ECIR 2023. Lecture Notes in Computer Science, vol 13980. Springer, Cham.* The methodology design and implementation, subsequent analyses and writing of this publication were handled by T. Goldsack. This encompasses the primary contribution of this publication, the novel

[1]In addition to the publication-specific Appendices (A - E) which are equal to those included in the original publications, and additional Appendix (F) is included that contains greyscale-friendly versions of several Figures.

unsupervised methodology for scientific document summarisation. All authors were involved in project conceptualisation and results discussions, led by T. Goldsack. C. Lin and C. Scarton were responsible for project supervision, and Z. Zhang further contributed through the running of baseline approaches.

2. ***Goldsack, T.***, *Zhang, Z., Lin, C., and Scarton, C. 2022. Making Science Simple: Corpora for the Lay Summarisation of Scientific Literature. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP).*

   The data collection and complication, subsequent analyses, benchmark method design and implementation, and writing of this publication were handled by T. Goldsack. This encompasses the primary contribution of this publication, the release, analysis, and benchmarking of two novel datasets for Scientific Lay Summarisation. All authors were involved in project conceptualisation and results discussions, led by T. Goldsack. C. Lin and C. Scarton were responsible for project supervision.

3. ***Goldsack, T.***, *Zhang, Z., Tang, C., Scarton, C., and Lin, C. 2023. Enhancing Biomedical Lay Summarisation with External Knowledge Graphs. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP).*

   The methodology design and implementation, subsequent analyses and writing of this publication were handled by T. Goldsack. This encompasses the primary contribution of this publication, the implementation, comparison, and analysis of three knowledge-graph augmented methodologies for Lay Summarisation. All authors were involved in project conceptualisation and results discussions, led by T. Goldsack. C. Lin and C. Scarton were responsible for project supervision.

4. ***Goldsack, T.***, *Scarton, C., and Lin, C. 2024. Leveraging Large Language Models for Zero-shot Lay Summarisation in Biomedicine and Beyond. Currently under review in ACL Rolling Review (ARR).*

The methodology design and implementation, writing of this publication were handled by T. Goldsack. This encompasses the primary contribution of this publication, the novel two-stage LLM framework for Lay Summarisation. All authors were involved in project conceptualisation and results discussions, led by T. Goldsack. C. Lin and C. Scarton were responsible for project supervision.

5. ***Goldsack, T.***, *Wang, Y., Lin, C, and Chen, C. 2025. From Facts to Insights: A Study on the Generation and Evaluation of Analytical Reports for Deciphering Earnings Calls. 2025 International Conference on Computational Linguistics (COLING).*

   The methodology design and implementation, writing of this publication were handled by T. Goldsack. This encompasses the primary contribution of this publication, the novel multi-agent approach to analytical report generation. All authors were involved in project conceptualisation and results discussions, led by T. Goldsack. C. Lin and C. Scarton were responsible for project supervision.

1.3.2 BioLaySumm Shared Task

Although not directly included in this thesis, a significant contribution was also made in the form of the BioLaySumm shared task - a public competition that was organized, leveraging the datasets introduced in the publication covered in Chapter 3. The overview papers for each of the first two editions of this (ongoing) task are as follows:

6. ***Goldsack, T.***, *Luo, Z., Xie, Q., Scarton, C., Shardlow, M., Ananiadou, S., and Lin, C. 2023. Overview of the BioLaySumm 2023 Shared Task on Lay Summarization of Biomedical Research Articles. In The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks.*

   The data collection, submission system setup, participant interaction, and cverview paper writing were handled by T. Goldsack. All authors were involved in project conceptualisation and results discussions, led by T. Goldsack, and participant paper reviewing, led by M. Shardlow. C. Lin and S. Ananiadou were responsible for project supervision.

7. ***Goldsack, T.**, Scarton, C., Shardlow, M., and Lin, C. 2024. Overview of the BioLaySumm 2024 Shared Task on the Lay Summarization of Biomedical Research Articles. In The 23rd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks.*

   The data collection, submission system setup, and participant interaction were handled by T. Goldsack. Overview publication writing was split between T. Goldsack for subtask 1, and Z. Luo and Q. Xie for subtask 2. All authors were involved in project conceptualisation and results discussions, led by T. Goldsack, and participant paper reviewing, led by M. Shardlow. C. Lin and S. Ananiadou were responsible for project supervision.

1.3.3 Other Published Work

In addition to those mentioned above, contributions to the following related publications were also made during the PhD:

8. *Tang, C., Wang, S., **Goldsack, T.**, and Lin, C. 2023. Improving Biomedical Abstractive Summarisation with Knowledge Aggregation from Citation Papers. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP).*

   The methodology design and implementation, subsequent analyses and writing of this publication were handled by C. Tang, encompassing the primary contribution of this publication. All authors were involved in project conceptualisation and results discussions, led by C. Tang. C. Lin was responsible for project supervision.

9. *Zhang, Z., **Goldsack, T.**, Scarton, C. and Lin, C. 2024. ATLAS: Improving Lay Summarisation with Attribute-based Control. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL, Volume 2: Short Papers).*

   The methodology design and implementation and subsequent analyses were handled by Z. Zhang, encompassing the primary contribution of this publication.

Z. Zhang and T. Goldsack shared paper writing responsibilities. All authors were involved in project conceptualisation and results discussions, led by C. Tang. C. Lin and C.Scarton were responsible for project supervision.

1.4 Acknowledgments

This work was supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [grant number EP/S023062/1]. This PhD was completed as part of the programme ran by Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications, hosted within the Department of Computer Science at the University of Sheffield. This programme additionally included cohort-based research training, most notably in the form of an Integrated Postgraduate Diploma (PGDip) in SLT ledership.

This project received partial funding from the University of Sheffield Advanced Manufacturing Research Center (AMRC). They have described applications for the research conducted in this thesis surrounding stakeholders without the required technical expertise (e.g., clients, corporate staff, etc.) receiving adapted non-expert summaries of their technical projects.

During final year of this PhD, an 2-month onsite research placement was undertaken under the supervision of Dr. Chung-Chi Chen at the National Institute of Advanced Industrial Science and Technology (AIST) in Tokyo, Japan. This placement was part of a research collaboration that ultimately resulted the final publication presented in this thesis: "From Facts to Insights: A Study on the Generation and Evaluation of Analytical Reports for Deciphering Earnings Calls".

Chapter 2

Publication I: Domain-driven and Discourse-guided Scientific Summarisation

Abstract

Scientific articles tend to follow a standardised discourse that enables a reader to quickly identify and extract useful or important information. We hypothesise that such structural conventions are strongly influenced by the scientific domain (e.g., Computer Science, Chemistry, etc.) and explore this through a novel extractive algorithm that utilises domain-specific discourse information for the task of abstract generation. In addition to being both simple and lightweight, the proposed algorithm constructs summaries in a structured and interpretable manner. In spite of these factors, we show that our approach outperforms strong baselines on the arXiv scientific summarisation dataset in both automatic and human evaluations, confirming that a scientific article's domain strongly influences its discourse structure and can be leveraged to effectively improve its summarisation.

2.1 Introduction

Scientific abstracts are used by researchers in determining whether a given article is relevant to their own work. Therefore, a well-written scientific abstract should concisely describe the essential content of an article from the author's perspective (Johnson, 1995), whilst following some standardised discourse structure (Graetz, 1982; Swales, 1990). Several structural classification schemes exist that attempt to model the sentence-level discourse of scientific articles within a particular scientific domain (e.g., Computer Science, Chemistry, etc.), with a focus on categorising sentences according to factors such as rhetorical status (Teufel, 1999; Liakata, 2010; Cohan et al., 2019). These schemes have proven utility in the automatic summarisation of scientific articles (Teufel and Moens, 2002; Contractor et al., 2012; Liakata et al., 2013).

Neural network-based approaches have gained increasing popularity for abstractive summarisation in recent years (El-Kassas et al., 2021b). Although these models are capable of producing coherent summaries, they are prone to generating hallucinations (i.e., text that is unfaithful to the source document) which can result in factual inconsistencies in their output (Maynez et al., 2020; Cheng et al., 2021). When it comes to scientific content, this is especially problematic as it may result in the dissemination of false or misleading information relating to important research findings (Peng et al., 2021). On the other hand, extractive summarisation systems typically ensure the basic syntactic and semantic correctness of their output and thus remain widely used (El-Kassas et al., 2021b).

Recent extractive works have attempted to leverage coarse *section-level* discourse structure of scientific articles, assuming that this provides a strong signal in determining the most informative content of the input. Dong et al. (Dong et al., 2021) build on the unsupervised graph-based approach of Zheng and Lapata (Zheng and Lapata, 2019), introducing a hierarchical document representation that accounts for both intra- and inter-section connections and exploits positional cues to determine sentence importance. Similarly, Zhu et al. (Zhu et al., 2021b) exploit discourse structural information, proposing a supervised approach that extracts article sec-

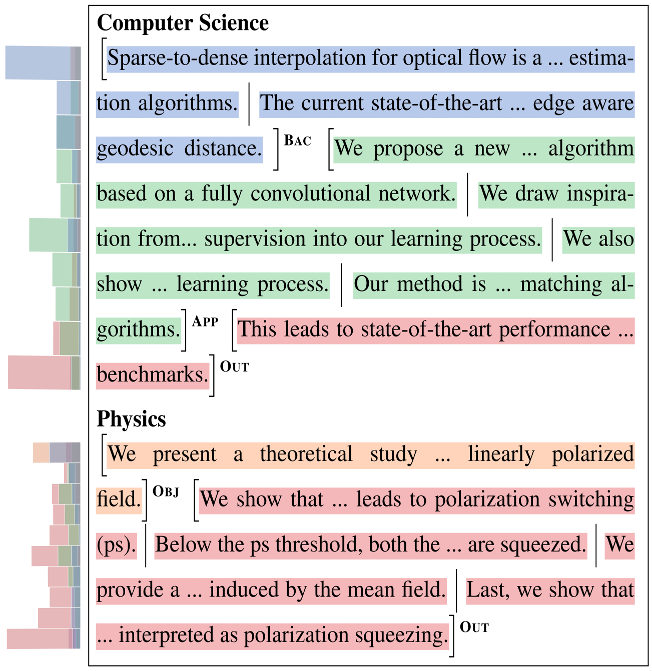


Figure 2.1: *Examples of the rhetorical abstract structures typical of different scientific domains (Computer Science and Physics). Left side distributions are covered further in Figure 2.4. Highlighted text donates rhetorical labels:* **Back**ground, **Obj**ective, *f99***App**roach, *and* **Out**come.

tions based on predicted salience before ranking their sentences via a hierarchical graph-based summariser. While the aforementioned works have explored modelling the discourse of articles to some extent, few have addressed how it can be used to directly influence the structure and content of their generated output. Furthermore, to our knowledge, no prior work has studied how the *scientific domain* of an article impacts the discourse structure of its abstract.

In this work, we tackle the problem of generating scientific abstracts, proposing a novel summarisation method that explicitly models and exploits *domain-specific* discourse, a valuable source of information that has not been explored in the prior literature. We hypothesise that the scientific domain of an article has a strong influence over the rhetorical structure of its abstract (exemplified in Figure 2.1), and thus can be leveraged in the context of abstract generation to improve both the content and structure of the generated output. As such, we conduct the first study on the influence of scientific domain over the discourse structure of abstracts, using it to inform our summarisation method, **DodoRank** (**Do**main and **D**iscourse-**o**riented

Dataset	**Domain**	**# Train / Val / Test**
ART corpus† (2010)	Chemistry	169 / 28 / 28
CSAbstruct† * (2019)	CS	1668 / 295 / 226
AZ Abstract† (2010)	Biomedical	750 / 150 / 100
AZ Article† (2013)	Biomedical	37 / 7 / 6
arXiv* (2018)	Multi-domain	203K / 6.4K / 6.4K

Table 2.1: *Scientific datasets used in this work. Here, † denotes that sentences are manually annotated with rhetorical labels and * denotes the official data splits.*

Ranking model). DodoRank consists of two primary components: (i) *a discourse extraction module* that, for a given dataset, determines which sections contain the most salient content for abstract generation and computes domain-specific, sentence-level rhetorical abstract structures for governing the generation process; and (ii) an *unsupervised extractive summariser*, which produces a scientific abstract based on the extracted domain-specific discourse information. Specifically, discourse information is used to *both* reduce the input to the most salient sections and impose a rhetorical structure upon our generated output that conforms to the conventions of the scientific domain. The sentences of salient sections are ranked and extracted in an unsupervised fashion, using sentence *centrality* as a domain-independent measure of importance (Erkan and Radev, 2004b; Zheng and Lapata, 2019; Dong et al., 2021).

Consequently, DodoRank constitutes a lightweight and interpretable approach to summarisation that, despite its simplicity, gives better or comparable performance to strong supervised baselines on the multi-domain arXiv dataset (Cohan et al., 2018), whilst also reducing the size of the input by an average of 66.24%. We further illustrate the effectiveness our approach via human evaluation, achieving superior performance to the state-of-the-art centrality-based model. Finally, we provide a domain-specific breakdown of results on arXiv, conclusively demonstrating that consideration of a scientific article's domain is beneficial for its summarisation.

2.2 Related Work

Scientific Discourse. The formalised nature of scientific writing has led to the creation of several classification schemes which categorise sentences according to their role within the larger discourse structure. By capturing the different types of information scientific articles contain at a fine-grained level, these schemes have the ability to support both the manual study and automatic analysis of scientific literature and its inherent structural discourse (Teufel and Moens, 2002; Guo et al., 2010; Goldsack et al., 2022b). The two most prevalent of these schemes are Argumentative Zoning (Teufel, 1999) and Core Scientific Concepts (CoreSC) (Liakata, 2010), both of which categorise sentences according to their rhetorical status and are provided alongside manually annotated corpora with scientific articles from the Computational Linguistics and Chemistry domains, respectively. Subsequent works have since introduced corpora annotated with similar schemes, typically focusing on a single scientific domain (Teufel et al., 2009; Guo et al., 2010, 2013; Cohan et al., 2019). Several of these datasets are used within this work (see Table 2.1).

Extractive Summarisation of Scientific Articles. Extractive summarisation approaches aim to build a summary using text spans extracted from the source document. These approaches remain attractive as they prevent factual hallucinations in the generated output, resulting in more reliable and usable summaries.

Prior to the rise of deep neural models, the use of traditional supervised algorithms proved popular, commonly used in combination with statistical text features such as sentence length and location, TFIDF scores, and frequency of citations (Teufel and Moens, 2002; Contractor et al., 2012; Collins et al., 2017; Liakata et al., 2013). Rhetorical classification schemes, such as those previously described, have also been shown to have value as features to these algorithms (Teufel and Moens, 2002; Contractor et al., 2012). We make use of rhetorical classes in a style similar to that of Liakata et al. (Liakata et al., 2013) who use the distribution of CoreSC classes within Chemistry articles to create a rhetorical plan for their generated output. In contrast to this work, we derive rhetorical structures directly from abstracts themselves and deploy them on *multiple scientific domains*. These earlier works also fail to address the section-based structure of scientific documents, which has since been shown to have influence over the distribution of summary-worthy content (Dong et al., 2021; Zhu et al., 2021b).

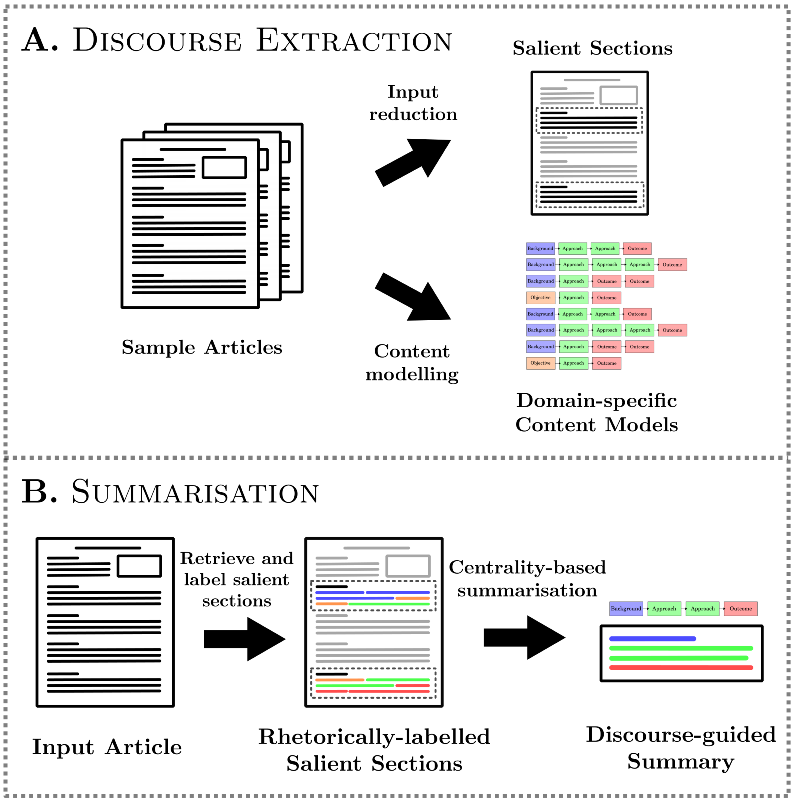
More recently, Xiao and Carenini (Xiao and Carenini, 2019) incorporate section information within a bi-directional RNN document encoder, before outputting confidence scores for each sentence. As covered in §2.1, both Dong et al. (Dong et al., 2021) and Zhu et al. (Zhu et al., 2021b) model coarse-level structural discourse information within hierarchical document graphs. Additionally, Dong et al. (Dong et al., 2021) determine the importance of a sentence calculating its centrality within a group of sentences (Zheng and Lapata, 2019; Erkan and Radev, 2004b). In this work, we also make use of centrality to compute sentence importance. However, where Dong et al. (Dong et al., 2021) group sentences based on the origin section and place emphasis on the position of sentences within sections, we group sentences based on rhetorical status and place emphasis on structuring output in a way that conforms with the conventions of the specific scientific domain.

2.3 Method

Our summarisation framework consists of two key components: A) a discourse extraction component and B) a summarisation component, as illustrated in Figure 2.2. The discourse extraction component operates offline during an initial 'learning phase'. It is responsible for (i) determining the most salient article sections for abstract generation (§2.3.1), and (ii) capturing domain-specific, sentence-level rhetorical abstract structures (§2.3.2). The summarisation component then employs the extracted discourse information in an online setting for guided abstract generation based on the centrality ranking of sentences (§2.3.3).

2.3.1 Salient Section Determination

A scientific abstract is a concise summary of a research paper, typically highlighting a few important components such as the research motivation, problem, approach, and key results and/or findings. We hypothesise that some of the article sections



ht

Figure 2.2: DodoRank *model overview.*

serve a more similar communicative goal to the abstract than others, and hence contribute more significantly to its content.

To deduce the most salient article sections for a given dataset, we propose to assess their individual contribution to the content of the abstract. We conduct our analysis and discourse information extraction based on a *sample set*, derived from the arXiv dataset, that contains articles from all eight scientific domains.[1] We balance sampling in terms of the scientific domain to ensure that the discourse information extracted is representative of all domains in the dataset. Further details on the sample set are given in §2.4.

For each abstract within the sample, we calculate the similarity between each sentence it contains and every sentence in the main body of the respective article. Due to the scientific nature of the text, similarity is measured using cosine similarity between SciBERT embeddings (Beltagy et al., 2019). Subsequently, the sentence with the greatest similarity to each abstract sentence (referred to as the *oracle*

[1]This sample set is also used within §2.3.2, as indicated in Figure 2.2.

Conflated heading	Matching terms
Introduction	"introduction"
Conclusion	"conclu", "summary"
Discussion	"discussion"
Result/analysis	"result", "analys", "ablat"
Background/motivation	"background", "motivation"
Method	"implement", "method"
Model	"architec", "system", "model"
Future work	"direction", "future"
References	"referenc"
Acknowledgements	"acknowledg"
Related work	"related"

Table 2.2: *Matching terms used for each conflated heading.*

sentence) and the heading of the section in which it is located are retrieved.

In verbatim form, the retrieved section headings are noisy, with much variation in phrasing and formatting when referring to semantically similar concepts (e.g., the concluding section of an article can be titled "conclusion", "summary", "concluding remarks", etc., all of which are semantically identical). Therefore, following Ermakova et al. (Ermakova et al., 2018), verbatim headings are conflated into a standardised form by matching selected words and sub-words against them using regular expressions. Matching terms for all conflated headings are derived empirically based on the sample set and given in Table 2.2.

The most important sections are regarded as those which the oracle sentences most frequently originate from, and we use only sentences from these sections as input to our summarisation component. Specifically, we select the minimum amount of sections that cumulatively contribute to at least 50% of all oracle sentences (see Table 2.5), to ensure sufficient coverage of salient content. Our analysis based on the arXiv dataset shows that the Introduction and Conclusion sections are the most salient contributors for the abstracts across *all* tested domains. Please refer to §2.5.1 for detailed analysis.

2.3.2 Rhetorical Content Modelling

To govern the generation process of our summariser, we aim to extract a rhetorical structure which is representative of a typical abstract for each specific scientific

Our label	**Verbatim label**			
	ART Corpus	**CSAbstruct**	**AZ Article**	**AZ Abstract**
Background	Background	Background	Background Connection Difference	Background Related Work
Objective	Motivation Goal Hypothesis Object	Objective	Problem Future Work	Objective Future Work
Approach	Experiment Model Method	Method	Method	Method
Outcome	Observation Result Conclusion	Result	Result Conclusion	Result Conclusion
Other	-	Other	-	-

Table 2.3: *Mapping of the rhetorical labels of other datasets to our classification scheme.*

domain. We refer to these structures as *content models*. To this end, we adopt a sentence-level classification scheme similar to that of CSAbstruct (Cohan et al., 2019), a dataset consisting of rhetorically labelled Computer Science abstracts. Specifically, our scheme contains the rhetorical labels: {Background, Objective, Approach, Outcome, Other}, where each label represents a high-level rhetorical role that may be assigned to any sentence in a given article, regardless of scientific domain.

Applying this classification scheme requires us to obtain the rhetorical labels for the unlabelled scientific abstracts. Therefore, we train a SciBERT-based sequential classifier (Cohan et al., 2019; Devlin et al., 2019a; Beltagy et al., 2019) on a combination of four datasets, all of which contain scientific articles and/or abstracts manually annotated with similar sentence-level rhetorical schemes by their creators. Specifically, we convert the labels of CSAbstruct (Cohan et al., 2019), the ART corpus (Liakata, 2010), AZ Abstract (Guo et al., 2010), and AZ Article (Guo et al., 2013) to our given scheme via a simple label mapping procedure, illustrated in Table 2.3. In combining these datasets and exposing the classifier to instances of our rhetorical classes from different scientific domains (see Table 2.1), we aim to make it more robust to unseen domains. We validate the reliability of this mapping by evaluating the trained classifier on the test set of CSAbstruct (§2.5.1), by far the largest of the contributing datasets.

Following label prediction on the abstracts of the sample set, we employ a frequency-based approach to extract the domain-specific content models. The core idea of this approach is to find the most common pattern of rhetorical labels observed in the abstracts of a given domain. A content model M corresponds to a sequence of K rhetorical sentence labels, where K is a hyperparameter to our model and determined based on overall model performance on the validation split (e.g., for a value of $K = 3$, an example M could be [Background, Approach, Outcome]).

To extract our domain-specific content models, one intuitive solution would be to sequentially compute the rhetorical label that occurs most frequently within its sample set abstracts for each sentence position from 0 to K. However, abstracts vary in length, and therefore a simple frequency-based method without taking this into account will be sub-optimal in capturing the rhetorical label distributions of the sample set. To tackle this challenge, we propose to normalise a sample abstract's rhetorical label sequence $\mathbf{s}$ by performing position-aware up/down sampling, thus producing a normalised label sequence $\mathbf{s}'$ of length K that approximates the original rhetorical label distribution. This normalisation process (i.e., construction of $\mathbf{s}'$) is formally defined below:

$$i_{\mathbf{s}} = \lfloor \frac{i_{\mathbf{s}'}}{K-1}(|\mathbf{s}| - 1) \rfloor, \tag{2.1}$$

where $|\mathbf{s}|$ is the number of sentence labels (equivalent to the number of sentences) in the sample abstract. Eq. (2.1) essentially samples an index $i_{\mathbf{s}}$ in $\mathbf{s}$ from which we retrieve a label value for sentence position $i_{\mathbf{s}'}$ ($0 \leq i_{\mathbf{s}'} < K$) of the normalised sequence $\mathbf{s}'$. More specifically, if $|\mathbf{s}|$ is larger than the content model length K, Eq. (2.1) is used to perform position-aware down-sampling of $\mathbf{s}$, retrieving a subset of its labels to form a condensed version of its label distribution; likewise, if $|\mathbf{s}|$ is smaller than K, it will be used to perform position-aware up-sampling of $\mathbf{s}$ to form an expanded version of its label distribution. Figure 2.3 exemplifies how Eq. (2.1)

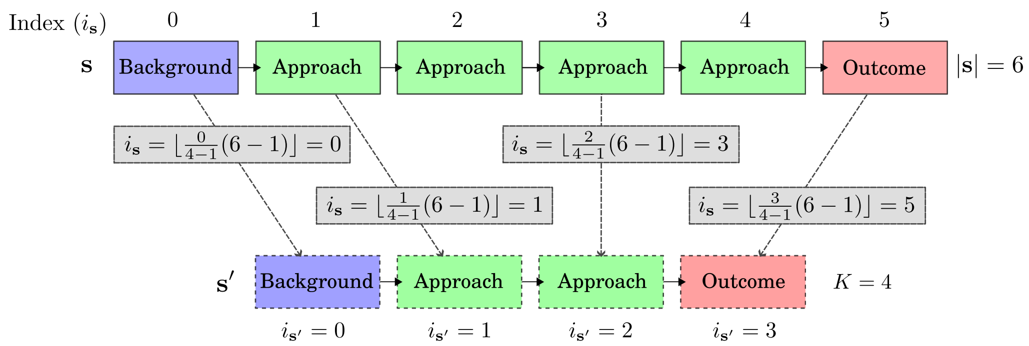


Figure 2.3: *Example showing how a sample abstract* $\mathbf{s}$ *is compressed to* $\mathbf{s}'$ *using* (2.1).

samples the labels from $\mathbf{s}$ to form a representation $\mathbf{s}'$ with K elements.[2] After deriving an $\mathbf{s}'$ for each $\mathbf{s}$ within a given domain, we construct a content model M by calculating the most frequent label observed at each normalised sentence position. By following this position-aware approach rather than truncating/padding sample abstracts to K labels, we ensure that our content models better reflect the true rhetorical distributions of the sample abstracts.

2.3.3 Centrality-based Summariser

As per §2.3.1, the summariser receives the previously identified salient sections as input (i.e., Introduction and Conclusion). Prior to summarisation, the sentences of these sections are assigned a rhetorical label using the classifier described in §2.3.2. To generate an extractive summary guided by a content model, we first group input sentences by their assigned rhetorical class. For each class label within the content model, we extract the candidate sentence with the greatest centrality from the corresponding group. In order to avoid redundancy, a sentence is no longer considered a candidate sentence once it is extracted, and thus can not be extracted for subsequent occurrences of the same rhetorical label.

The centrality of a sentence s_i is equal to the average cosine similarity between

[2] Note that we also experimented with both rounding up and rounding to the nearest integer value for Eq. (2.1), but found that rounding down gave the best performance.

Domain	**Frequency**		
	Train	**Valid**	**Test**
Physics	169,827	5,666	5,715
Mathematics	20,141	360	322
Computer Science	9,041	280	258
Quantitative Biology	1,842	59	71
Statistics	1,399	47	46
Quantitative Finance	701	24	28
EESS	80	0	0
Economics	6	0	0
Total	203,037	6,436	6,440

Table 2.4: *The frequencies of different scientific domains within the different data splits of arXiv (EESS = Electrical Engineering and Systems Science).*

s_i and every other sentence within its group:

$$\text{Centrality}(s_i) = \frac{1}{N}\sum_{j=1}^{N} e_{i,j}, \tag{2.2}$$

$$e_{i,j} = \text{Cosine}(v_i, v_j) \tag{2.3}$$

Here v_i and v_j are the SciBERT embeddings of sentences s_i and s_j, and N is the number of sentences of the same rhetorical class.

2.4 Experimental Setup

Datasets. For the task of abstract generation, we experiment on the test split of the popular arXiv scientific summarisation dataset (Cohan et al., 2018), for which document-summary pairs consist of full research articles (taken from the arXiv online repository) and their author-written abstracts. We assume a scenario where the domain of the input article is known by the user (as such information is typically available or easily predictable). We retrieve the scientific domain of each article within arXiv using the unique article IDs.[3] As shown in Table 2.4, arXiv contains articles from eight scientific domains, with Physics being the most

[3]The domain names retrieved are equal to highest-level categories as defined in the arXiv category taxonomy: https://arxiv.org/category_taxonomy

frequent. For discourse extraction (i.e., the derivation of salient sections and domain-specific content models), we sample 5,000 instances from the train split of arXiv ($\approx$ 840 samples from each domain, or all training instances when less than this is available), allowing for meaningful statistics to be calculated for each domains whilst remaining relatively computationally inexpensive. Also note that we exclude EESS and Economics domains for content modelling and the main experiment due to their limited size and no valid/test sets.

Implementation details. DodoRank contains only one hyperparameter, the length of the content model K. We found that a value of $K = 6$ gave the best performance on the validation set of arXiv, meaning our output summaries exclusively consist of 6 sentences.[4] Moreover, this happens to be the median number of sentences for abstracts within arXiv (Ju et al., 2021).

Baselines. We compare our model with traditional baselines Oracle, Lead, LexRank (Erkan and Radev, 2004b), and LSA (Steinberger and Jezek, 2004a). We include supervised baselines Discourse-Aware (Cohan et al., 2018), Seq2Seq-Loc&Glob (Xiao and Carenini, 2019), Match-Sum (Zhong et al., 2020), Topic-GraphSum (Cui et al., 2020) and SSN-DM (Cui and Hu, 2021). For unsupervised models, we include the results of PacSum (Zheng and Lapata, 2019) and HipoRank (Dong et al., 2021), the latter of which achieves state-of-the-art performance for an unsupervised model.

Evaluation. We perform automatic evaluation using standard ROUGE metrics (Lin, 2004).[5] Specifically, we give the average F1-scores for ROUGE-1, ROUGE-2 and ROUGE-L. Following Dong et al. (Dong et al., 2021), we also perform human evaluation on a subset of abstracts within the arXiv test set, allowing for a more comprehensive comparison of model performance.

[4]Increasing or decreasing K (which directly influences the number of sentences in the summaries produced by DodoRank) invariably led to a worse average performance, as measured by the metrics described in this Section.

[5]All ROUGE calculations are performed using the `rouge-score` Python package.

Conflated heading	**Contribution (%)**
Introduction	**35.98**
Conclusion	14.94
Results/Analysis	8.06
Discussion	5.13
Model	4.15
Method	2.68
Background/motivation	0.86
Compression ratio	**33.76**

Table 2.5: *The average section contribution to abstract content within the arXiv sample set (% of oracle sentences originating from each section). Underlined text denotes selected input sections.*

Oracle type	**R-1**	**R-2**	**R-L**
Full text	47.45	21.53	41.99
Extracted (I + C)	46.35	20.77	40.67

Table 2.6: *Sample set oracle ROUGE scores using the full text and extracted sections (I = Introduction, C = Conclusion).*

2.5 Experimental Results

2.5.1 Structural Discourse Analyses

Section contribution to abstract content. Table 2.5 gives the contribution of different sections to the content of the abstracts within the *sample set*, highlighting the selected sections. As stated in §2.3.1, this is measured as the percentage of oracle sentences that originate from a given section. We find that over 50% of abstract content is derived from the Introduction and Conclusion sections. The provided compression ratio indicates that these sections constitute 33.76% of an article on average. To validate the input reduction process, we also calculate the average ROUGE F-scores obtained by two oracle-based summaries (cf. §2.3.1) when compared to the reference abstracts: one containing oracle sentences extracted from the full text, and the other containing oracle sentences extracted from only the salient sections (Table 2.6).

The results given in Table 2.6 support this, showing only a small disparity between the oracle ROUGE scores obtained using only these sections and those obtained using the full text (average difference < 1). As further validation, we

Training data	F-score	Acc.
CSAbstruct	0.794	0.819
Combined datasets	0.811	0.824

Table 2.7: *Classifier performance for 5-way rhetorical label classification on the CSAbstruct test set. Results are average scores of 3 runs with different random seeds.*

	aX[P]	aX[M]	aX[CS]	aX[QB]	aX[S]	CSA
aX[P]	-	0.215	0.250	0.098	0.217	0.307
aX[M]	0.215	-	0.140	0.213	0.137	0.204
aX[CS]	0.250	0.140	-	0.201	0.094	0.138
aX[QB]	0.098	0.213	0.201	-	0.169	0.246
aX[S]	0.217	0.137	0.094	0.169	-	0.160
CSA	0.307	0.204	0.138	0.246	0.160	-

Table 2.8: *The Jensen-Shannon divergence between the distributions given in Figure 2.4, averaged across all rhetorical labels (CSA = CSAbstruct, aX = arXiv, P = Physics, M = Mathematics, CS = Computer Science, QB = Quantitative Biology, S = Statistics).*

carry out significance testing on oracle ROUGE scores (t-test), the results of which indicate that differences between performance using the full text and the extracted sections are is statistically insignificant for all variants ($p = 0.26$, 0.48, and 0.19 for ROUGE-1, 2, and L, respectively).

Rhetorical label prediction. Table 2.7 provides statistics on the performance of the SciBERT-based classifier for rhetorical label prediction on the Computer Science abstracts of the CSAbstruct test set. Although the classifier exhibits strong performance when using only the training data of CSAbstruct, the additional out-of-domain data provided by our combined datasets further improves performance in terms of both accuracy and F-score, attesting to the domain-independence of our classification scheme (described in §2.3.2). This suggests that we can rely on predicted rhetorical labels for subsequent experiments.

Rhetorical structure of abstracts. Figure 2.4 provides a visualisation of how the frequency of each rhetorical class changes according to the sentence position within the abstracts of different scientific domains. For each sub-graph, by observing the pattern of most frequent labels (tallest bars) across all positions, we can identify

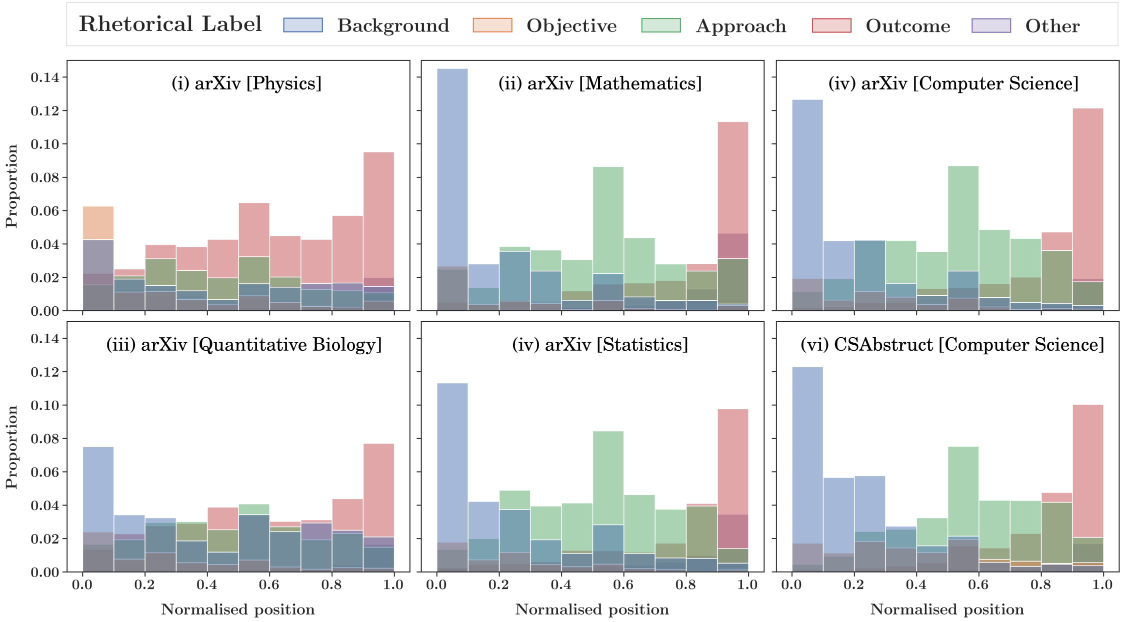


Figure 2.4: *Visualisation of the rhetorical distributions of abstracts for different domains.*

the dominant rhetorical structure for a given domain. It is evident that these structures differ significantly depending on the scientific domain. To quantify this difference, we compute the average Jensen Shannon divergence (JSD) between the distributions (Table 2.8).

We show the five domains most prevalent within the arXiv train set: Physics (83.6%), Mathematics (9.8%), Computer Science (4.5%), Quantitative Biology (0.9%), and Statistics (0.7%). As an additional point of comparison with our predicted label distributions, we also include the distribution of the manually labelled CSAbstruct. We find that the most similar rhetorical structures are shared by the Statistics and Computer Science instances within arXiv (0.094 JSD). Both domains, in addition to Mathematics which also attains a low JSD score with each, follow the rhetorical flow of Background→Approach→Outcome. Similar patterns can also be observed in the extracted content models for all of these domains. Additionally, we find that predicted distributions for arXiv's Computer Science instances are similar to that CSAbstruct (0.138 JSD), further supporting the reliability of the predicted labels. Interestingly, the abstracts of both the Physics and Quantitative Biology domain instances within arXiv exhibit a rhetorical structure that differs significantly from

Model	**R-1**	**R-2**	**R-L**
Oracle∗	53.88	23.05	34.90
Lead∗	33.66	8.94	22.19
LexRank (2004)∗	33.85	10.73	28.99
LSA (2004)∗	29.91	7.42	25.67
Supervised Models			
Discourse-aware (2018)†	35.80	11.05	31.80
Seq2Seq-Loc&Glob (2019)†	43.62	17.36	29.14
Match-Sum (2020)†	40.59	12.98	32.64
Topic-GraphSum (2020)†	44.03	18.52	32.41
SSN-DM+discourse (2021)†	44.90	19.06	32.77
Unsupervised Models			
PacSum (2019)∗	38.57	10.93	34.33
HipoRank (2021)∗	39.34	12.56	34.89
DodoRank (ours)	40.11	14.20	35.31
DodoRank$_{no_ss}$ (ours)	36.66	10.87	31.12
DodoRank$_{no_cm}$ (ours)	36.51	12.15	31.99

Table 2.9: *Test set results on arXiv (ROUGE F1). Results with † and ∗ are taken from (Cui and Hu, 2021) and (Dong et al., 2021), respectively. Results with ‡ are reproduced.*

the other presented domains, placing a much greater emphasis on the Outcome class. Furthermore, their abstract structures are judged to be very similar by way of JSD score (0.098).

2.5.2 Abstract Generation

Automatic evaluation. Table 2.9 presents the performance of DodoRank and selected baselines on the arXiv test split. We include the results of two ablated versions of DodoRank, DodoRank$_{no_ss}$ and DodoRank$_{no_cm}$. Here, DodoRank$_{no_ss}$ denotes omission of section selection (full article text is used) and DodoRank$_{no_cm}$, the omission of the sentence-level content models (i.e., K most central sentences are selected, regardless of rhetorical class).

DodoRank achieves strong performance, exceeding that of supervised models Discourse-aware (in all metrics) and Match-Sum (in R-2 and R-L, whilst achieving comparable performance in R-1), despite employing an unsupervised extractive summariser using centrality. DodoRank also outperforms both closely-related centrality-based unsupervised baselines in all metrics. Furthermore, the use of both

Domain	**Model**	**R-1**	**R-2**	**R-L**
Physics	DodoRank	40.34	14.47	35.49
	DodoRank$_{no_ss}$	37.07	11.24	32.48
	DodoRank$_{no_cm}$	36.40	12.19	31.87
Mathematics	DodoRank	34.65	10.21	30.53
	DodoRank$_{no_ss}$	30.94	7.12	27.03
	DodoRank$_{no_cm}$	33.76	10.17	29.77
Computer Science	DodoRank	41.13	14.01	36.27
	DodoRank$_{no_ss}$	34.51	8.38	30.27
	DodoRank$_{no_cm}$	39.79	13.46	35.13
Quantitive Biology	DodoRank	38.69	10.79	33.90
	DodoRank$_{no_ss}$	34.86	7.99	30.59
	DodoRank$_{no_cm}$	37.99	10.54	33.41
Statistics	DodoRank	39.06	10.93	34.24
	DodoRank$_{no_ss}$	35.19	7.28	31.88
	DodoRank$_{no_cm}$	38.64	11.20	33.89
Quantitive Finance	DodoRank	39.84	13.18	35.59
	DodoRank$_{no_ss}$	36.72	10.06	31.88
	DodoRank$_{no_cm}$	38.34	13.25	33.81

Table 2.10: *Domain-specific breakdown of results (ROUGE F1) on the arXiv test set.*

discourse-based sub-components results in a large improvement compared to when only one is used. This attests to the utility of both sub-components and indicates that: 1) the rhetorical structure imposed by the domain-specific content models results in the extraction of sentences more similar to those in the reference abstracts, and 2) the most central sentences within the Introduction and Conclusion sections better reflect the abstract content than the most central sentences of the document as a whole.

Table 2.10 provides a domain-specific breakdown of ROUGE scores on the arXiv test set for standard and ablated versions of DodoRank. Again, we observe a universal improvement in performance across all domains when both discourse-based analyses are included, providing further indication that our analyses capture the nuances of each scientific domain and, in doing so, improves the summarisation of their articles.

Human evaluation. Results for human evaluation are given in Table 2.11. We chose to replicate the procedure of Dong et al. (Dong et al., 2021) to facilitate a

Model	Content-coverage	Importance
HipoRank	12.17	37.21
DodoRank	**22.03**	**59.04**

Table 2.11: *Human evaluation results on 20 sampled reference summaries with 307 system summary sentences from arXiv. Results for both criteria are statistically significant with Mann-Whitney U tests.*

direct comparison between our model and the state-of-the-art for centrality-based extractive summarisation model HipoRank. For this evaluation, two human judges[6] are repeatedly presented with a reference abstract and a sentence extracted by either DodoRank or HipoRank model in a random anonymised order. Each sentence is evaluated on whether it contains content from the reference abstract (*content coverage*) and whether it contains content that would be important for a goal-oriented reader, regardless of whether it is covered in the reference abstract (*importance*). Each reference summary-sentences pair is annotated by two annotators (i.e., a score 1 will be assigned if a sentence is deemed content relevant or important; and 0 otherwise). Scores are aggregated and then averaged across the sentences from the tested models. The average annotator agreement of 73.31%, attesting to their reliability.

DodoRank significantly outperforms the HipoRank model for both content coverage and importance, in line with automatic evaluation results. Overall, these results provide further indication that summarising only the salient sections of an article and arranging output according to a domain-specific rhetorical structure improves the overall quality of the summaries produced, particularly in the selection of important content.

2.6 Conclusions

In this paper, we proposed an extractive, discourse-guided approach to the generation of scientific abstracts which adapts to the scientific domain. For a given domain, the model extracts rich discourse information which is used to both reduce the input and guide the output of a simple centrality-based summariser. Our approach exhibits

[6]Judges are native English speakers holding a bachelor's degree in scientific disciplines.

impressive performance on the multi-domain arXiv dataset, exceeding that of strong baselines, both supervised and unsupervised. This demonstrates that the scientific domain of an article can effectively be leveraged in a summarisation context and supports our original hypothesis, that the domain of an article has a strong influence over the structural discourse of its abstract.

2.7 Updated Limitations

Additional limitations of this work are outlined below for the purpose of this thesis:

- In addition to the datasets used in this work to the develop classification scheme and train the sentence classifier, the PubMed dataset (Dernoncourt and Lee, 2017) could have been utilised, which cotains abstracts with similar rhetorical sentence annotations. Despite being smaller in size than PubMed, the datasets selected in this work were prioritised due to having human-annotated rhetorical labels, whereas PubMed contains automatically-derived labels.

- The adoption of average sentence centrality as the measure of sentence importance in the extractive summarisation system is based on the assumption that sentence embeddings are clustered in the embedding space and the efficacy of this measure when used scientific data in related work (Zheng and Lapata, 2019). Other measures, such as the average dot product of sentence embeddings, are also viable but were not tested.

- In Table 2.10, the performance of DodoRank of articles from the Mathematics domain is significantly worse than that of other domains. Upon inspection of the failure cases, this appears to be largely due to to the fact that sentences in articles from this domain contain multiple mathematical formulae that are parsed incorrectly. As a result of this, elements of mathematical formulae are often included in the generated abstracts when they are generally absent from reference abstracts, resulting in lower ROUGE scores.

- In light of developments in the field of summarisation since the publication of this work, including an LLM based approach as a baseline would now be appropriate.

Chapter 3

Publication II: Making Science Simple: Corpora for the Lay Summarisation of Scientific Literature

Abstract

Lay summarisation aims to jointly summarise and simplify a given text, thus making its content more comprehensible to non-experts. Automatic approaches for lay summarisation can provide significant value in broadening access to scientific literature, enabling a greater degree of both interdisciplinary knowledge sharing and public understanding when it comes to research findings. However, current corpora for this task are limited in their size and scope, hindering the development of broadly applicable data-driven approaches. Aiming to rectify these issues, we present two novel lay summarisation datasets, PLOS (large-scale) and eLife (medium-scale), each of which contains biomedical journal articles alongside expert-written lay summaries. We provide a thorough characterisation of our lay summaries, highlighting differing levels of readability and abstractiveness between datasets that can be leveraged to support the needs of different applications. Finally, we benchmark our datasets

Technical Abstract
The virus SARS-CoV-2 can exploit biological vulnerabilities (e.g. host proteins) in susceptible hosts that predispose to the development of severe COVID-19. To identify host proteins that may contribute to the risk of severe COVID-19, we undertook proteome-wide genetic colocalisation tests, and polygenic (pan) and cis-Mendelian randomisation analyses leveraging publicly available protein and COVID-19 datasets...

Lay Summary
Individuals who become infected with the virus that causes COVID-19 can experience a wide variety of symptoms. These can range from no symptoms or minor symptoms to severe illness and death. Key demographic factors, such as age, gender and race, are known to affect how susceptible an individual is to infection. However, molecular factors, such as unique gene mutations and gene expression levels can also have a major impact on patient responses by affecting the levels of proteins in the body...

Figure 3.1: *The first few sentences of the abstract and lay summary of an eLife article, illustrating differences in the language and focus on background information.*

using mainstream summarisation approaches and perform a manual evaluation with domain experts, demonstrating their utility and casting light on the key challenges of this task.

3.1 Introduction

Scientific publications contain information that is essential for the preservation and progression of our understanding across all scientific disciplines. Typically being highly technical in nature, such articles tend to assume a degree of background knowledge and make use of domain-specific language, making them difficult to comprehend for one lacking the required expertise (i.e., a *lay person*). These factors often limit the impact of research to only its direct community (Albert et al., 2015, 2021) and, more dangerously, can cause readers (members of the public, journalists, etc.) to misinterpret research findings (Kuehne and Olden, 2015).

This latter point is especially important for biomedical research which, in addition to having particularly dynamic and confusing terminology (Smith, 2006; Peng et al., 2021), has the potential to directly impact people's decision-making regarding health-related issues, with a pertinent example of this being the widespread misinformation seen during the COVID-19 pandemic (Islam et al., 2020). Aiming to address these challenges, some academic journals choose to publish *lay summaries* that clearly and concisely explain the context and significance of an article using non-specialist language. Figure 3.1 illustrates how simplifying jargon (e.g., “SARS-CoV-2" → “the virus that causes COVID-19") and focusing on background information allows a reader to better understand a complex scientific topic. However, in addition to placing an extra burden on authors, lay summaries are not yet ubiquitous and focus only on newly published articles.

Automatic text summarisation can provide significant value in the generation of scientific lay summaries. Although previous use of summarisation techniques for scientific articles has largely focused on generating a technical summary (e.g., the abstract), only a few have addressed the task of lay summarisation and introduced datasets to facilitate its study (Chandrasekaran et al., 2020; Guo et al., 2021; Zaman et al., 2020). However, compared to datasets ordinarily used for training supervised summarisation models, these resources are relatively small (ranging from 572 to 6,695 articles), presenting a significant barrier to the deployment of large data-driven approaches that require training on large amounts of parallel data. Furthermore, these resources are somewhat fragmented in terms of their framing of the task, making use of article and summary formats that limit their applicability to broader biomedical literature. These factors hinder the progression of the field and the development of usable models that can be used to make scientific content accessible to a wider audience.

To help alleviate these issues, we introduce **two new datasets** derived from different academic journals **within the biomedical domain** - PLOS and eLife (§3.3). Both datasets use the full journal article as the source, enabling the training of models which can be broadly applied to wider literature. PLOS is significantly larger than currently available datasets and makes use of short author-written lay summaries (150-200 words), whereas eLife's summaries are approximately twice as long and written by expert editors who are well-practiced in the simplification of scientific content. Given these differences in authorship and length, we expect the lay summaries of eLife to simplify content to a greater extent, meaning our datasets are able to cater to different audiences and applications (e.g., personalised lay summarisation). We confirm this via an in-depth characterisation of the lay summaries within each dataset, quantifying ways in which they differ from the technical abstract and from each other (§3.4). Finally, we benchmark our datasets with popular summarisation approaches using automatic metrics and conduct an expert-based manual evaluation, highlighting the utility of our datasets and key challenges for the task of lay summarisation (§3.5). This paper also presents a literature review (§3.2), conclusions (§3.6), and a discussion on its limitations (§3.7).

3.2 Related Work

Past attempts to automatically summarise scientific content in layman's terms have been scarce, with the most prominent example being the LaySumm subtask of the CL-SciSumm 2020 shared task series (Chandrasekaran et al., 2020) which attracted a total of 8 submissions. Alongside the task, a training corpus of 572 articles and author-generated lay summaries from a multi-disciplinary collection of Elsevier-published scientific journals was provided, with submissions being evaluated on a blind test set of 37 articles. It was noted by the task organisers that the data provided was insufficient for training a model to produce a realistic lay summary.

Guo et al. (2021) also make use of a single publication source to retrieve lay summaries: The Cochrane Database of Systematic Reviews (CDSR). Their dataset contains the abstracts of 6,695 systematic reviews paired with their respective plain-language summaries, covering various healthcare domains. Although larger than other available datasets for lay summarisation, CDSR is constrained in that it only uses the abstracts of systematic reviews as source documents, and thus models trained using CDSR will be unlikely to generalise well to inputs that are longer than an abstract or the abstracts of other types of publication.

Alternatively, Zaman et al. (2020) introduce a dataset derived from the 'Eureka-Alert' science news website for the combined tasks of simplification and summarisation. Summaries consist of news articles (average length > 600 words) that aim to describe the content of a scientific publication to the non-expert. However, the extensive size of reference summaries is likely to present additional challenges in model training and their news-based format limits their applicability (e.g., in automating lay summarisation for journals).

Compared to previous resources, our datasets contain articles and lay summaries of a format that we consider to be more broadly applicable to wider literature. Additionally, PLOS is significantly larger than those currently available (over $4\times$ larger than CDSR) and eLife contains summaries written by expert editors. Furthermore, our work is the first to provide two datasets with different levels of readability, thus supporting the needs of different audiences and applications. Through each of these factors, we hope to enable the creation of more usable lay summarisation models.

3.3 Our Datasets

We introduce two datasets from different biomedical journals (PLOS and eLife), each containing full scientific articles paired with manually-created lay summaries. For each data source, articles were retrieved in XML format and parsed using Python to retrieve the lay summary, abstract, and article text.[1] In line with previous datasets for scientific summarisation (Cohan et al., 2018), the article text is separated into sections, and the heading of each section is also retrieved. Sentences are segmented using the PySBD rule-based parser (Sadvilkar and Neumann, 2020), which we empirically found to outperform neural alternatives. We separate our datasets into training, validation, and testing splits at a ratio of 90%/5%/5%. Statistics describing the contents of our datasets and that of past lay summarisation datasets are given in Table 3.1.

PLOS The Public Library of Science (PLOS) is an open-access publisher that hosts influential peer-reviewed journals across all areas of science and medicine. Several of these journals require authors to submit an *author summary* alongside their work, defined as a 150-200 word non-technical summary aimed at making the

[1] For each article, we also retrieve a number of *keywords* from the meta-data, providing an indication of the high-level topics covered within the article.

Dataset	# Docs	Doc	Summary	
		# words	# words	# sents
LaySumm	572	4,426.1	82.15	3.8
Eureka-Alert	5,204	5,027.0	635.6	24.3
CDSR	6,695	576.0	338.2	16.1
PLOS	27,525	5,366.7	175.6	7.8
eLife	4,828	7,806.1	347.6	15.7

Table 3.1: *Statistics of lay summarisation datasets, with ours given in* ***bold****. Words and sentences (sents) are average values.*

Metric	Abstract		Lay Summary	
	PLOS	**eLife**	**PLOS**	**eLife**
FKGL↓	15.04	15.57	14.76	10.92
CLI↓	16.39	17.68	15.90	12.51
DCRS↓	11.06	11.78	10.91	8.83
WordRank↓	9.08	9.21	8.98	8.68

Table 3.2: *Mean readability scores for abstracts and lay summaries from our datasets. For all metrics, a lower score indicates greater readability.*

findings of a paper accessible to a wider audience, including non-scientists.[2] The journals in question focus specifically on Biology, Computational Biology, Genetics, Pathogens, and Neglected Tropical Diseases.

eLife eLife is an open-access peer-reviewed journal with a specific focus on biomedical and life sciences. Of the articles published in eLife, some are selected to be the subject of a *digest*, a simplified summary of the work written by expert editors based on both the article itself and questions answered by its author. Similarly to PLOS, these digests aim to explain the background and significance of a scientific article in language that is accessible to non-experts (King et al., 2017).

3.4 Dataset Analysis

We carry out several analyses comparing the lay summaries of our datasets to the respective technical abstracts. Through these analyses, we seek to highlight and quantify the key differences between these two different types of summary, as well as

[2]Source of PLOS author summary definition: https://journals.plos.org/plosgenetics/s/submission-guidelines

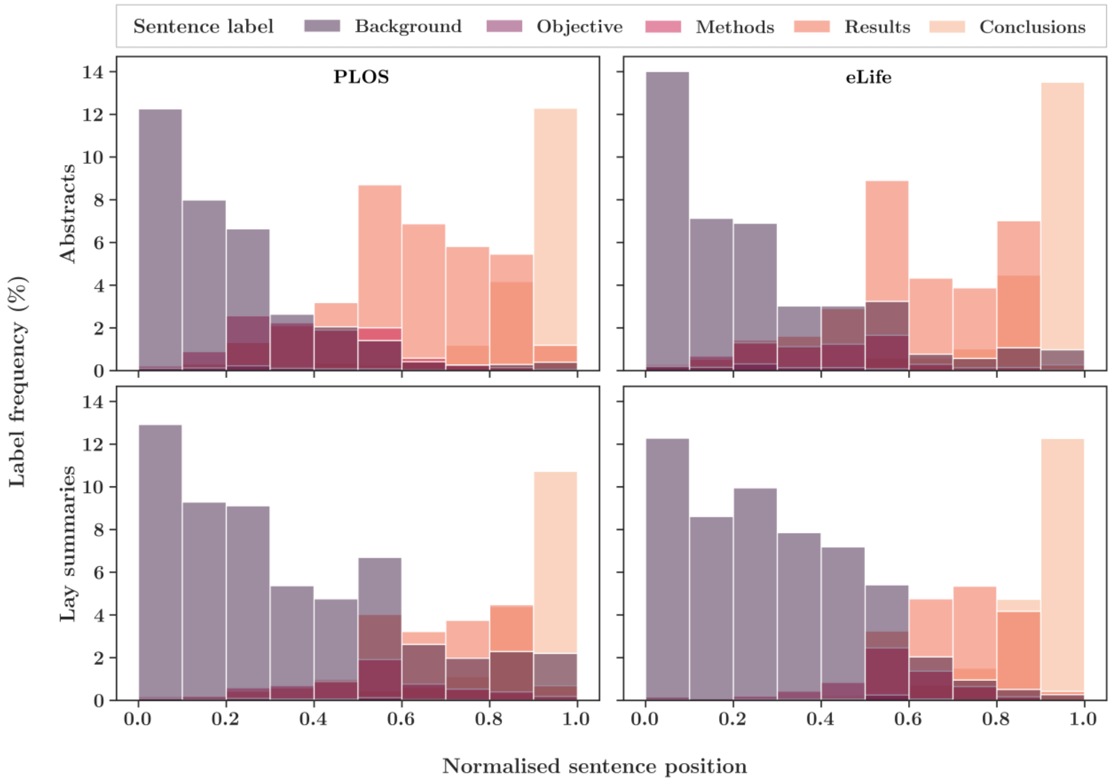


Figure 3.2: *Barplot visualising the rhetorical class distributions in our abstracts and lay summaries.*

those present between the lay summaries of our two datasets. Specifically, we focus on readability (§3.4.1), rhetorical structure (§3.4.2), vocabulary sharing (§3.4.3), and abstractiveness (§3.4.4).

3.4.1 Readability

We assess the readability of our lay summaries and abstracts using several established metrics. Specifically, we employ Flesch-Kincaid Grade Level (FKGL), Coleman-Liau Index (CLI), Dale-Chall Readability Score (DCRS), and WordRank score.[3] FKGL, CLI, and DCRS provide an approximation of the (US) grade level of education required to read a given text. The formula for FKGL surrounds the total number of sentences, words, and syllables present within the text, whereas CLI is based on the number of sentences, words, and characters. Alternatively, DCRS measures readability using the average sentence length and the number of *familiar* words

[3]Computed using the `textstat` and `EASSE` (Alva-Manchego et al., 2019) packages.

Label	**Abstract**		**Lay Summary**	
	PLOS	**eLife**	**PLOS**	**eLife**
Background	35.40	41.05	58.11	55.03
Objective	0.76	1.06	0.54	0.47
Methods	10.26	6.73	6.24	6.23
Results	34.75	30.60	17.86	18.23
Conclusions	18.83	20.55	17.26	18.83

Table 3.3: *Mean percentage of each rhetorical label within our abstracts and lay summaries.*

present, using a lookup table of the 3,000 most commonly used English words. Similarly, WordRank estimates the lexical complexity of a text based on how common the language is, using a frequency table derived from English Wikipedia.

The scores given in Table 3.2 show that the lay summaries of both datasets are consistently more readable than their respective abstracts across all metrics. Although these differences are small in some cases, in line with the findings of previous works (Devaraj et al., 2021), we find them all to be **statistically significant** by way of Mann–Whitney U tests ($p < 0.05$). These results indicate that lay summaries are more readable than technical abstracts in terms of both syntactic structure and lexical intelligibility. Additionally, the lay summaries from eLife obtain lower readability scores than those of PLOS across all metrics, confirming our expectation that they are suitable for less technical audiences.[4]

3.4.2 Rhetorical Structure

Rhetoric is another important factor when assessing the comprehensibility of a text. Specifically, a lay person will require a much larger focus on the background of a scientific article than an expert in order to understand the significance of its findings (King et al., 2017), thus we would expect lay summaries to focus more on such aspects.

To provide further insight into the structural differences between abstracts and lay summaries, we classify all sentences within each based on their rhetorical status. To do this, we make use of PubMed RTC (Dernoncourt and Lee, 2017), a

[4]Manual inspection of the summaries from each dataset also support this.

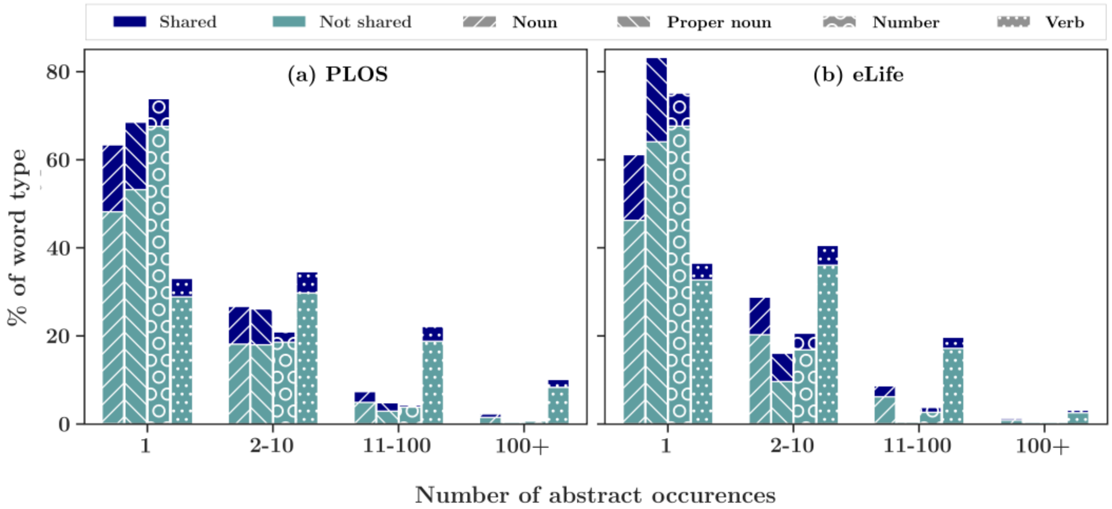


Figure 3.3: *Stacked barplot showing how regularly (on average) abstract content words are shared with the respective lay summaries (as a % of all words of that type), separated by number of abstract occurrences.*

dataset containing the 20,000 biomedical abstracts retrieved from PubMed, with each sentence labelled according to its rhetorical role (roles: Background, Objective, Methods, Results, Conclusions). We use PubMed RTC to train the BERT-based sequential classifier introduced by Cohan et al. (2019) due to its strong reported performance (92.9 micro F1-score), before applying this model to lay summary and abstract sentences from our datasets.

Figure 3.2 provides a visualisation of how the frequency of each rhetorical class changes according to the sentence position within our summaries. For each sub-graph, observing the pattern of most frequent labels (tallest bars) across all positions allows us to get an idea of the dominant rhetorical structure. In Table 3.3, we further quantify the difference in structure by giving the average percentage of each label present in the different summaries.

For both datasets, we see a similar pattern when comparing abstract and lay summary distributions. Specifically, a much greater portion of lay summary sentences is dedicated to explaining the relevant background information (“Background"). This is unsurprising, as such information is essential to understanding the motivation and significance of any work and, thus, would be of great value to a non-expert. This additional focus on “Background" comes at the expense of sentences focusing on “Results" and (to a lesser extent) "Methods", which are less frequent within lay summaries. Again, this is to be expected, as these details are less meaningful to an audience without domain expertise.

3.4.3 Content Words

Aiming to determine what terminology is shared between summary types, we analyse the frequency at which *content words* occur simultaneously within abstracts and lay summaries. We treat nouns, proper nouns, verbs, and numbers as content words, and we extract these from the summaries using ScispaCy (Neumann et al., 2019), a library that specialises in the processing of biomedical texts.[5] Figure 3.3 shows the results of our analysis, visualising the average rate at which different types of content word from the abstracts are shared with lay summaries. We divide our analysis based on the number of abstracts content words occur in, allowing us to observe how the ubiquity of a word within the corpus affects the rate at which it is shared. [6]

Generally, we observe similar patterns for content words between datasets. Firstly, regardless of word type or number of abstract occurrences, we find that abstract content words are rarely shared with lay summaries (i.e., 'shared' % $<$ 'not shared' % for all bars). This is indicative of a clear shift in content and/or vocabulary when it comes to the creation of a lay summary. For each dataset, we can also see that the vast majority of content words of all types, except verbs, occur in 10 or fewer abstracts ($>$ 90% on average for both datasets), with most of these occurring within a single abstract. We found that the majority of these low-frequency content words are highly specific to the topic of a single article or small group of articles. In the case of nouns and proper nouns, we empirically observed these instances to often be highly technical terms (e.g., specific chemicals, lesser-known diseases, etc.), whereas numbers are typically exact numerical figures. It is understandable, therefore, that single-use content words of these types (noun, proper noun, and number) are rarely

[5] `en_core_sci_scibert` model used for POS tagging, derived from SciBERT (Beltagy et al., 2019).

[6] E.g., For nouns within PLOS, 15.3% occur in a single abstract and are shared with lay summaries, and 48.2% occur in a single abstract and are not shared (i.e., total percentage of nouns that occur in a single abstract = 15.3 + 48.2 = 63.5%).

included in the lay summary, as they will likely be meaningless to a lay reader. The pattern exhibited by verbs differs significantly from that of other word types, as they typically occur in a greater number of abstracts (most commonly being present within 2-10). For content words of all types, we observe that the ratio of 'shared' to 'not shared' generally increases in line with the number of abstract occurrences.[7]

3.4.4 Abstractiveness

We follow the example of prior works (Sharma et al., 2019a; See et al., 2017) by calculating the *abstractiveness* of our summaries using n-gram novelty, thus providing a measurement of the degree to which the summary uses different language to describe the content of the article. Specifically, for both abstracts and lay summaries, we compute the percentage of summary n-grams which are absent from their respective article. The results of this analysis are presented in Figure 3.4, where we can observe that lay summaries consistently contain more novel n-grams than abstracts across both datasets. However, the lay summaries of eLife, in addition to being approximately twice as long as those of PLOS (Table 3.1), appear to be significantly more abstractive. Alongside differences in readability (highlighted in §3.4.1), we believe these to be important distinctions that should be considered in determining a suitable dataset for a particular use case or application.

3.5 Experiments and Results

To help facilitate future work, we benchmark our datasets using popular heuristics-based, unsupervised, and supervised summarisation approaches (§3.5.1). Additionally, we provide further insight into these results via a detailed discussion (§3.5.2) and an expert-based manual evaluation (§3.5.3).

[7]Better illustrated by Table B.2 in Appendix B, which gives the exact percentages shown in Figure 3.3.

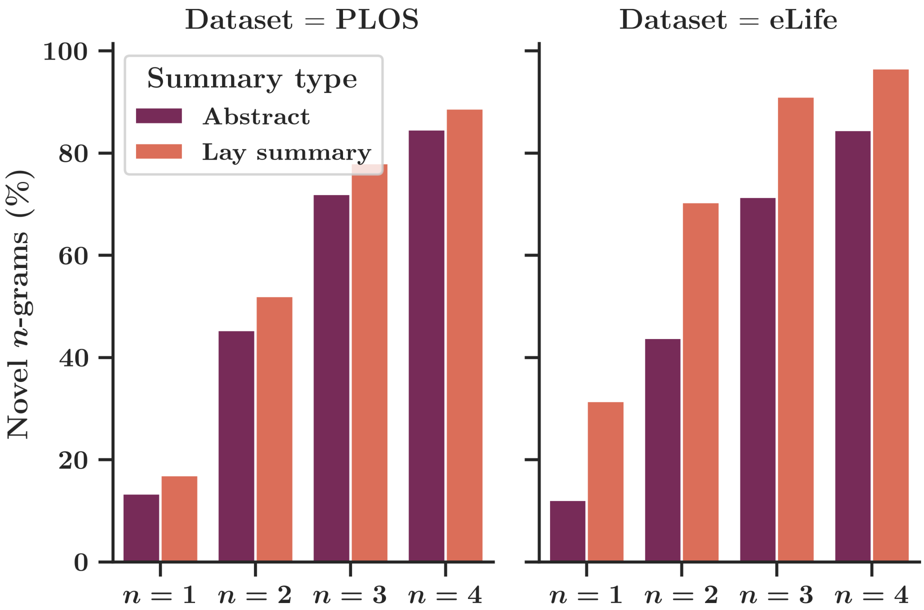


Figure 3.4: *Barplot showing the percentage of novel n-grams for each summary type.*

3.5.1 Baseline Approaches

For our heuristics-based approaches, we include the widely-used Lead-3 baseline which simply uses the first three sentences of the main body of the text. As our lay summaries typically consist of more than three sentences, we also include Lead-k, with k being equal to the average lay summary length for each dataset (Table 3.1). Additionally, we include the scores obtained by the technical abstracts (Abstract) and Oracle_Ext, a greedy extractive oracle (Nallapati et al., 2017) that provides an upper bound for the expected performance of extractive models.[8]

We benchmark four unsupervised extractive approaches: LSA (Steinberger and Jezek, 2004b), LexRank (Erkan and Radev, 2004a), TextRank (Mihalcea and Tarau, 2004), and HipoRank (Dong et al., 2021). For supervised models, we use the transformer-based BART base model (Lewis et al., 2020), which we fine-tune on our own datasets. To assess how the use of additional data in various forms can benefit performance, we include several other variants of this model which are described in the remainder of this subsection.[9]

[8]The extractive oracle is further explained in Appendix B.
[9]For all variants, we truncate the input to 1,024 tokens.

	Approach	PLOS					eLife				
		R-1	R-2	R-L	FKGL	DCSR	R-1	R-2	R-L	FKGL	DCSR
Heuristic	Lead-3	25.46	6.35	22.89	15.08	12.66	17.93	3.66	16.45	13.30	12.65
	Lead-K	35.50	8.69	32.33	14.94	11.88	34.12	6.73	32.06	11.89	10.58
	Abstract	47.07	18.60	43.51	14.98	11.10	28.92	6.19	27.04	15.35	11.87
	Oracle_Ext	56.43	31.24	52.88	15.28	11.20	46.38	11.48	43.82	13.18	10.51
Unsupervised	LSA	35.14	6.92	31.46	17.39	12.41	34.57	5.69	32.27	16.50	11.52
	LexRank	36.53	9.43	33.09	14.90	**10.51**	30.32	5.29	28.40	13.45	9.75
	TextRank	36.03	8.71	32.04	20.42	11.83	31.45	5.28	29.22	19.13	11.28
	HipoRank	42.16	11.41	38.01	14.91	11.82	29.95	5.38	27.44	13.58	12.32
Supervised	Bart	**42.35**	12.96	38.57	14.62	12.07	**46.57**	**11.65**	**43.70**	10.94	9.36
	Bart$_{PubMed}$	42.12	12.70	38.34	14.75	12.09	46.20	11.36	43.21	11.24	9.64
	Bart$_{Cross}$	42.24	**13.52**	**38.63**	**14.24**	12.18	46.22	11.53	43.33	10.81	9.31
	Bart$_{Scaffold}$	39.47	9.73	35.79	14.45	11.74	45.28	10.99	42.51	**10.65**	**9.19**

Table 3.4: *Performance of summarisation models on the test splits of each dataset (**R** = average ROUGE F1-score). The best non-heuristic scores for each metric are given in **bold**.*

Additional training As our datasets remain smaller than those used in other forms of summarisation, we experiment with BART$_{PubMed}$, which is previously trained on the PubMed abstract generation dataset (Cohan et al., 2018) and fine-tuned on our own datasets. Aiming to assess how well models trained on PLOS can generalise to eLife and vice versa, we also include BART$_{Cross}$, which is initially trained on the opposite dataset to that which it is eventually fine-tuned and evaluated on.

Scaffolding We also experiment with artificially enlarging our training data by way of a scaffold task. Inspired by CATTS (Cachola et al., 2020), we remove the article's abstract from the input text and train the model to generate the abstract as a scaffold task to lay summarisation. In addition to showing whether training for abstract generation can benefit lay summarisation, results for this model will provide an indication of the baseline BART model's reliance on the abstract content. Specifically, we include two copies of every article within our training data - one using the abstract as the reference summary and the other using the lay summary. We distinguish between the two by prepending the input document with the control tokens ⟨|`ABSTRACT`|⟩ or ⟨|`SUMMARY`|⟩. Documents within the validation and test splits are prepended the ⟨|`SUMMARY`|⟩ code to induce lay summary generation. This model is denoted by BART$_{Scaffold}$.

3.5.2 Discussion

Table 3.4 presents the performance of the aforementioned approaches on the PLOS and eLife test splits using automatic metrics. In line with common practice for summarisation, we report the F1-scores of ROUGE-1, 2, and L (Lin, 2004). Additionally, we include FKGL and DCRS scores of the generated output (see §3.4.1), providing an assessment of the syntactic and lexical complexity, respectively.

The importance of the abstract Based on the ROUGE scores obtained by the Abstract baseline, we can safely assume that the lay summaries of PLOS are much closer in resemblance to their respective abstracts than those of eLife. The importance of the abstract for lay summary generation is further highlighted by the ROUGE scores of the $\text{BART}_{Scaffold}$ model, which performs notably worse than the standard BART model on PLOS and slightly worse on eLife. These results suggest that having the abstract included within the model input provides significantly more benefit than using abstract generation as an auxiliary training signal.

Extractive vs abstractive In general, we would expect abstractive methods to have greater application for the task of lay summarisation due to their ability to transform (and thus, simplify) an input text. However, abstractive approaches have a tendency to generate hallucinations, resulting in factual inconsistencies between the source and output that damages their usability (Maynez et al., 2020). Therefore, extractive approaches may still have utility for the task, especially if the comprehensibility of selected sentences is directly considered.

For ROUGE scores, we find that extractive baselines (i.e., all unsupervised and heuristic approaches) perform significantly better on PLOS than on eLife, aligning with our previous analysis (§3.4.4) which identified PLOS as the less abstractive dataset. Interestingly, readability scores achieved by extractive models on PLOS match and sometimes exceed those of abstractive BART models, although they are inconsistent. For eLife, abstractive methods (i.e., BART models) generally obtain superior scores for both ROUGE and readability metrics. In fact, the ROUGE scores achieved by BART exceed those obtained by Oracle_Ext, further indicating that abstractive methods have greater potential for this dataset.

Use of additional data As previously mentioned, artificially creating more data via an abstract-generation scaffold task results in a decrease in ROUGE scores for both datasets, indicating a reliance on the abstract content for lay summarisation. We also find that pretraining BART on Pubmed (BART$_{PubMed}$) does very little to affect the performance, suggesting that habits learned for abstract generation do not transfer well to lay summarisation. Similarly, BART$_{Cross}$ achieves a performance close to that of the standard BART model. Overall, these results indicate that additional out-of-domain training does provide much benefit for lay summarisation, and alternative modelling approaches that make better use of available data may be a more promising route for future work.

3.5.3 Human evaluation

To further assess the usability of the generated abstractive summaries, we perform an additional human evaluation of our standard Bart baseline model using two domain experts.[10]

Our evaluation uses a random sample of 10 articles from the test split of each dataset. Alongside each model-generated summary, judges are presented with both the abstract and reference lay summary of the given article. Using a 1-5 *Likert* scale, the annotators are asked to rate the model output based on three criteria: (1) *Comprehensiveness* – to what extent does the model output contain all information that might be necessary for a non-expert to understand the high-level topic of the article and the significance of the research; (2) *Layness* – to what extent is the content of the model output comprehensible (or readable) to a non-expert, in terms of both structure and language; (3) *Factuality* – to what extent is the model output factually consistent with the two other provided summaries. We choose not to provide judges with the full article text in an effort to minimise the complexity of the evaluation and the cognitive burden placed upon them.

[10]Both judges have experience in scientific research and hold at least a bachelor's degree in Biomedical Science.

Dataset	Comp.	Layness	Factuality
PLOS	3.7	3.0	3.0
eLife	3.1	3.0	3.0

Table 3.5: *Mean evaluator ratings (1-5) obtained by* Bart *outputs for each metric.*

Table 3.5 presents the average ratings from our manual evaluation. We calculate Krippendorff's α to measure inter-rater reliability, where we obtain values of 0.78 and 0.54 for PLOS and eLife, respectively. In addition to providing ratings, evaluators also provided comments on the general performance on each criterion for both datasets, providing further insights into model performance.

Comprehensiveness We can see from Table 3.5 that model outputs on PLOS are judged to be more comprehensive than on eLife. From evaluators' comments, we understand that this largely results from extensive use of abstract content for PLOS, which is sometimes copied directly (or with minor edits) to the lay summary. For eLife, it was observed that new information (i.e., not contained in the reference abstract or lay summary) was often introduced which was irrelevant or confusing, potentially affecting the understanding of a lay reader.

Layness Interestingly, given the previously highlighted differences in readability, the average layness of the model output is judged to be equal for both datasets (3.0), suggesting a reasonable degree of content simplification. However, evaluators' comments indicate that model outputs for each dataset were penalised for different reasons. For PLOS, the aforementioned use of abstract content often resulted in the inclusion of jargon terms that a lay reader would struggle to interpret. Alternatively, the language of eLife outputs was observed to be better suited to a lay audience but was sometimes simplified to a point that it could be misconstrued and mislead a reader, occasionally containing grammatical errors, typos, and repeated content.

Factuality Again, we find an equal average score of 3.0 given for factuality, suggesting the model struggles to produce factually correct outputs for both datasets. In fact, we found no output from either dataset was given a perfect score by both annotators, indicating that simplifying technical content accurately is a consistent problem. Evaluators' comments highlight contradictions, unclear phrasing, and misrepresentation of entities as key contributing factors to factual inconsistencies. We regard this as an integral obstacle to overcome in the development of usable lay summarisation models and an essential focus for future research.

3.6 Conclusion

In this work, we have introduced PLOS and eLife, two new datasets for the lay summarisation of biomedical research articles. Compared to currently available resources, these datasets possess source article and summary formats that are more broadly applicable to wider literature, with PLOS also being larger by a significant margin. A thorough analysis of our lay summaries highlights key differences between datasets, enabling them to cater to the needs of different audiences and applications. Specifically, in addition to being approximately twice as long as those of PLOS, we find eLife summaries to be both more readable and abstractive, thus better suited to a less technical audience. To facilitate future research, we benchmark our datasets with popular summarisation models using automatic metrics and conduct an expert-based human evaluation, providing further insight into the intricacies of model performance on our datasets and highlighting key challenges for the task of lay summarisation.

3.7 Limitations

Although we introduce the largest dataset available to the task of lay summarisation, our datasets remain smaller than those available for other forms of summarisation (e.g., abstract generation), where there exists datasets containing 100,000+ articles. This is largely due to the fact that lay summaries are less ubiquitous that other forms of summary (e.g., the abstract), only being used in a relatively small number of journals, of which only some are open-access and available to be utilised for such purposes as dataset creation.

On a related note, another potential limitation of our datasets is the fact they only cover a single broad domain - biomedicine. Again, this comes down to the availability of data, and the fact that the use of lay summaries is much less common in other scientific domains (e.g. Computer Science). There is, however, a reason for this disparity in the adoption of lay summaries between domains, as it is generally considered more important that the public have an awareness and understanding of research breakthroughs in health-related areas such as biomedicine. Therefore, we believe it is in these domains that automatic lay summarisation can provide the greatest benefit, although we also hope to address the lay summarisation of other domains in future work.

3.8 Acknowledgements

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

3.9 Updated Limitations

Additional limitations of this work are outlined below for the purpose of this thesis:

- For the human evaluation of the summaries included in this work, expert judges are not provided with the full article text, but instead rely on their expertise to make ratings according to specific criteria. This was done in an effort to reduce the already signficant cognitive burden and time requirements needed for judges to conduct the evaluation, but is an inherent limitation when judging specific criteria (e.g. Factuality).

- Similarly, although it would have been desirable to include more than two judges for the human evaluation in this work, this was not possible due to budget constraints. In light of these, constraint the domain expertise of select judges was highly prioritised.

- Although not stated in this chapter, human annotators were recruited from within the University of Sheffield Biomedical Sciences Department, and were compensated in the form of Amazon vouchers.

- As outlined in §3.5.3, evaluators' comments when making judgements for the Layness criterion indicated that model outputs for each dataset were often penalised for different reasons. This indicates that greater refinement of the evaluation guidelines could have provided a more effective results.

- In light of developments in the field of summarisation since the publication of this work, including an LLM based approach as a baseline would now be appropriate.

Chapter 4

Publication III: Enhancing Biomedical Lay Summarisation with External Knowledge Graphs

Abstract

Previous approaches for automatic lay summarisation are exclusively reliant on the source article that, given it is written for a technical audience (e.g., researchers), is unlikely to explicitly define all technical concepts or state all of the background information that is relevant for a lay audience. We address this issue by augmenting eLife, an existing biomedical lay summarisation dataset, with article-specific knowledge graphs, each containing detailed information on relevant biomedical concepts. Using both automatic and human evaluations, we systematically investigate the effectiveness of three different approaches for incorporating knowledge graphs within lay summarisation models, with each method targeting a distinct area of the encoder-decoder model architecture. Our results confirm that integrating graph-based domain knowledge can significantly benefit lay summarisation by substantially increasing the readability of generated text and improving the explanation of technical concepts.

4.1 Introduction

Lay summarisation consists of generating a concise summary that illustrates the significance of a longer technical (or otherwise specialist) text and is comprehensible to the non-expert (Kuehne and Olden, 2015). A lay summary should contain minimal jargon and technical details (e.g., methodology), instead focusing largely on the simplification of key technical concepts and the explanation or relevant background information, thus allowing readers without technical knowledge to grasp the general topic and main ideas of an article (Srikanth and Li, 2021; Goldsack et al., 2022a). However, since the original article is intended for a technical audience who already possess some domain knowledge, it is unlikely to explicitly include all the information necessary for the lay summary, such as background details or definitions. As a result, lay summaries are often highly abstractive, adopting a simpler lexicon than the original article (Goldsack et al., 2022a), and are typically written by experts who possess the knowledge required to simplify and explain the contents of the article (King et al., 2017).

Despite this disparity between lay summary and source article, automatic approaches to lay summarisation have typically relied solely upon the source article as input (Chandrasekaran et al., 2020; Guo et al., 2021; Luo et al., 2022a). Aiming to address this, we propose enhancing lay summarisation models with *external domain knowledge*, conducting the first study on **knowledge graph-enhanced lay summarisation** with a focus on biomedical articles. We augment eLife (Goldsack et al., 2022a), an existing high-quality lay summarisation dataset, with article-specific knowledge graphs containing information on the technical concepts covered within the articles and the relationships between them (exemplified in Figure 4.1), thus providing a structured representation of the domain knowledge that an expert human author might draw upon when writing a lay summary (§4.3). In doing this, we hypothesise that a model's ability to simplify and explain technical concepts for a lay audience will improve.

Although other forms of summarisation task (e.g., news articles) have seen significant research in the enhancing models using knowledge graphs (Huang et al.,

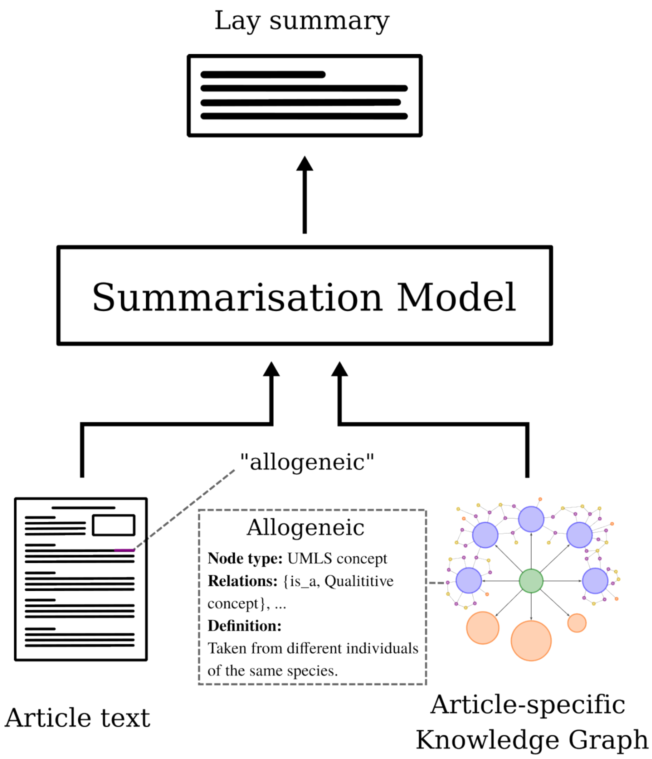


Figure 4.1: *Overview of the "knowledge graph-enhanced Lay Summarisation" task formulation, exemplifying graph-based external information.*

2020; Zhu et al., 2021a; Lyu et al., 2022), this has yet to be explored for lay summarisation. To our knowledge, there is no work on determining the most effective way to incorporate graph-based knowledge for lay summarisation or other summarisation tasks. Therefore, we systematically investigate the effectiveness of **three different methods for injecting graph-based information into lay summarisation models** (§4.4) assessing them with both automatic and human evaluation (§4.5 & §4.6). Our results demonstrate that the integration of graph-based domain knowledge can significantly improve automatic lay summarisation enabling models to generate substantially more readable text and to better explain technical concepts.

4.2 Related Work

4.2.1 Lay Summarisation

The task of lay summarisation is a relatively novel one, introduced by the LaySumm subtask of the CL-SciSumm 2020 shared task series (Chandrasekaran et al., 2020). Introducing a multi-domain corpus of 572 article-lay summary pairs, the task attracted a total of 8 participants. The winning system, proposed by Kim (2020), adopted a hybrid approach, using a PEGASUS-based (Zhang et al., 2020) model to generate an initial abstractive lay summary before augmenting this with sufficiently readable article sentences extracted by a BERT-based model (Devlin et al., 2019b).

Subsequent lay summarisation work has focused almost exclusively on the introduction and benchmarking of new corpora (all from the biomedical domain), rather than introducing specific modelling approaches for the task. Guo et al. (2021) introduce CDSR, a dataset derived from the Cochrane Database of Systematic Reviews, whereas Goldsack et al. (2022a) introduce PLOS and eLife, two datasets derived from different biomedical journals (the Public Library of Science and eLife journals, respectively).[1] Both studies benchmark their datasets with widely-used summarisation approaches, with BART variants (Lewis et al., 2020) invariably achieving the strongest performance. In another highly related work, Luo et al. (2022a) address the task of readability-controlled summarisation using data derived from PLOS, training a BART-based model to produce both the abstract and lay summary of an article in a controlled setting.

In contrast to previous works, we investigate the unexplored approach of modelling and integrating structured domain knowledge into lay summarisation models using article-specific knowledge graphs.

4.2.2 Knowledge Graph-Enhanced Text Generation

In recent years, the utilisation of knowledge graphs (KGs) containing external knowledge for text generation has seen increased interest, particularly when it comes

[1] A version of these datasets with different test sets is also used within the BioLaySumm 2023 shared task (Goldsack et al., 2023a), that ran in parallel with this work.

to the modelling of commonsense knowledge. In particular, works focusing on tasks such as dialogue generation (Zhou et al., 2018; Tang et al., 2023), commonsense-reasoning (Liu et al., 2021), story generation (Guan et al., 2019; Tang et al., 2022), essay generation (Yang et al., 2019) have all seen the introduction of commonsense KG-enhanced models.

Some recent work has also focused on using KGs for abstractive summarisation, but tending towards modeling internal knowledge. Aiming to improve the faithfulness and informativeness of summaries, Huang et al. (2020), Zhu et al. (2021a), and Lyu et al. (2022) all utilise OpenIE to construct fact-based knowledge graphs from source documents (news articles). Huang et al. (2020) and Zhu et al. (2021a) extract graph node features using graph attention networks (Veličković et al., 2017), before incorporating these into the summarisation model decoder using an attention mechanism. Lyu et al. (2022) instead make use of additional semantic loss measures to attempt to capture extracted facts using an adapted pointer-generator network.

In contrast to previous works, we apply KG-based techniques to biomedical lay summarisation (as opposed to news article summarisation), a domain with additional challenges, including the extensive length of input articles and the presence of complex technical concepts that need to be simplified or explained. Furthermore, due to the unique requirements of the task, we innovate by constructing knowledge graphs largely using *external* domain knowledge sources rather than from the article itself, as is the case with previous approaches.

4.3 Article Knowledge Graphs

We augment eLife, an existing dataset for biomedical lay summarisation with heterogeneous article-specific knowledge graphs (KGs). Each KG contains structured information on the complex biomedical concepts covered within the article and the relationships between them. In order to localise this information and provide an indication of where in an article a concept is mentioned, we also choose to model the article's section-based document structure within our graphs through the use of section-specific nodes. In the following, we describe in detail: 1) our process

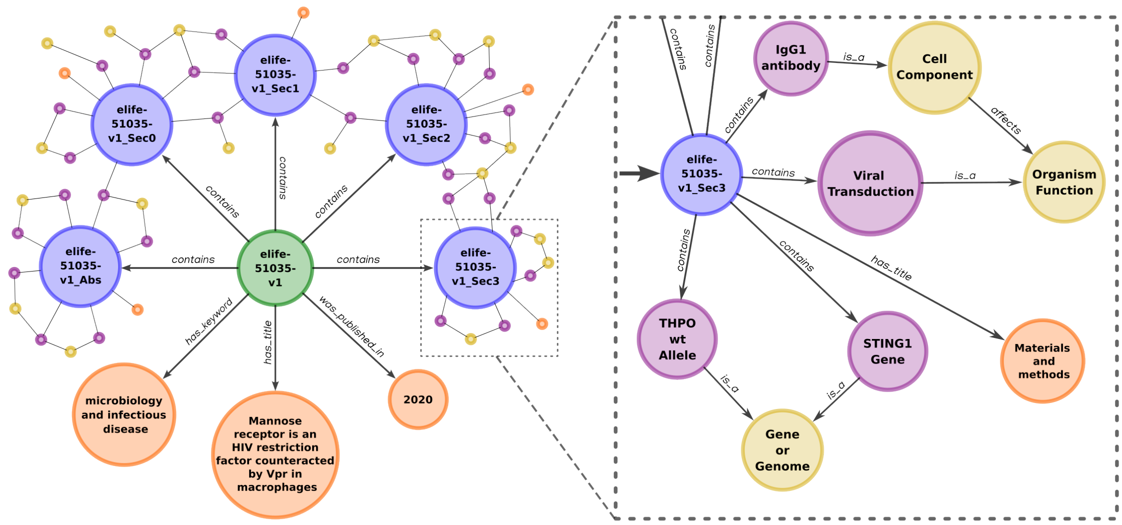


Figure 4.2: *Example of article knowledge graph structure. Graph nodes are coloured as follows: Document, Section, Metadata, Concept, Semantic type.*

for extracting the knowledge that is used by our model (§4.3.1), and 2) how we structure that knowledge within a knowledge graph (§4.3.2). The methods by which we integrate graph-based knowledge into summarisation models are discussed in §4.4.

4.3.1 Knowledge Extraction

To extract relevant domain knowledge for an article, we draw upon the Unified Medical Language System (UMLS) (Bodenreider, 2004). This rich and actively-maintained resource has long been used as a key knowledge source for NLP in the biomedical domain (McCray et al., 2001; Demner-Fushman et al., 2010; Kang et al., 2021) and is comprised of three primary components: the Metathesaurus, the Semantic Network, and the Specialist Lexicon and Lexical Tools. The Metathesaurus is an extensive multi-lingual vocabulary database containing information on a large number of biomedical concepts, including their various names and definitions. The Semantic Network defines a set of *semantic types* that represent broad subject categories into which *all* concepts in the Metathesaurus can be assigned. Additionally, high-level relationships that occur between different semantic types are also defined. To extract the UMLS concepts mentioned within a given article, we utilise MetaMap

(Aronson and Lang, 2010), one of the Lexical Tools provided alongside UMLS for this exact purpose, that is widely used in previous work (Sang et al., 2018; Sharma et al., 2019b; Lai et al., 2021). For all articles in eLife, we apply MetaMap to each section in turn, retrieving all mentioned UMLS concepts. We restrict MetaMap to only a select number of English vocabularies without prohibitive access restrictions, but otherwise run it using default settings.

In line with observations made in previous works (Lai et al., 2021), we found that MetaMap, whilst succeeding in linking biomedical entities mentioned in the text with their corresponding UMLS concepts, also frequently returned a number of irrelevant concepts. Therefore, we adopt a text overlap-based approach to filter the original pool of extracted concepts for a given section, which we empirically found to eliminate the vast majority of the noise.[2]

For each remaining UMLS concept, we retrieve all semantic types with which it is associated, in addition to the formal definitions of both the concept and semantic types. Notably, these definitions are used as an integral component within all three KG-enhancement methods.[3] In order to confirm their suitability for a lay audience, we calculate and compare their average readability scores with those reported by Goldsack et al. (2022a) for both the technical abstracts and lay summaries of eLife articles. The results of this analysis, given in Table 4.1, show that the UMLS definitions obtain scores that are overall much closer to those of the lay summaries than the abstracts, actually exceeding them in two out of the four metrics (FKGL and WordRank). An example of the definition format used for text augmentation is given in Figure C.2 in the Appendix. In the next section, we describe how we represent all extracted information within article-specific knowledge graphs.

4.3.2 Graph Construction

Each graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V}$ is a set of nodes (or entities) and $\mathcal{E}$ is a set of edges. Each edge $e_{ij} \in \mathcal{E}$ defines a relation r_{ij} between entities $v_i, v_j \in \mathcal{V}$, and thus can be

[2] More details on MetaMap vocabularies, noise reduction process, and the average article KG statistics are provided in Appendix C.

[3] Concepts without a formal UMLS definition are also removed from the final pool.

Metric	Abstract	Lay Summary	Definitions
FKGL↓	15.57	10.92	10.55
CLI↓	17.68	12.51	13.02
DCRS↓	11.78	8.83	10.36
WordRank↓	9.21	8.68	8.6

Table 4.1: *Mean readability scores for abstracts, lay summaries, and key UMLS definitions for eLife. FKGL = Flesch-Kinkaid Grade Level, CLI = Coleman-Liau Index, DCRS = Dale-Chall Readability Score.*

represented as a triplet $e_{ij} = (v_i, r_{ij}, v_j)$. All graphs are heterogeneous, containing multiple types of entities and relations. Figure 4.2 presents a visualisation of an article knowledge graph.[4] Each type of node is described below:

- **Document node** – the central *root node*, which is the ancestor of all other nodes in the graph. We label this node simply with the unique ID assigned to each article.
- **Section node** – each section node represents a specific titled section (e.g. Introduction) of the document, including the abstract. To label these nodes, we concatenate the article ID with "_Abs" for the abstract or "_Sec$\{i\}$" for other sections, where $\{i\}$ is the index of the section (zero-based).
- **Metadata node** – identify additional information relating to the article or its specific sections. This includes article and section titles, article keywords, and the date of publication.
- **Concept node** – nodes representing UMLS concepts. These are labelled with their unique UMLS identifier (CUI).
- **Semantic type node** – nodes representing semantic types from the Semantic Network. These are labelled with their unique Semantic Type identifier (TUI).

In addition to the 54 different relationship types defined within the semantic network (e.g., *affects* in Figure 4.2), we define several relations in order to represent the graph structure and additional metadata. Specifically, we define the relations *contains*, *was_published_in*, *has_title*, and *has_keyword*.

[4]Note that, for visual clarity, this example contains significantly fewer concept and semantic type nodes than are present in the actual article graph.

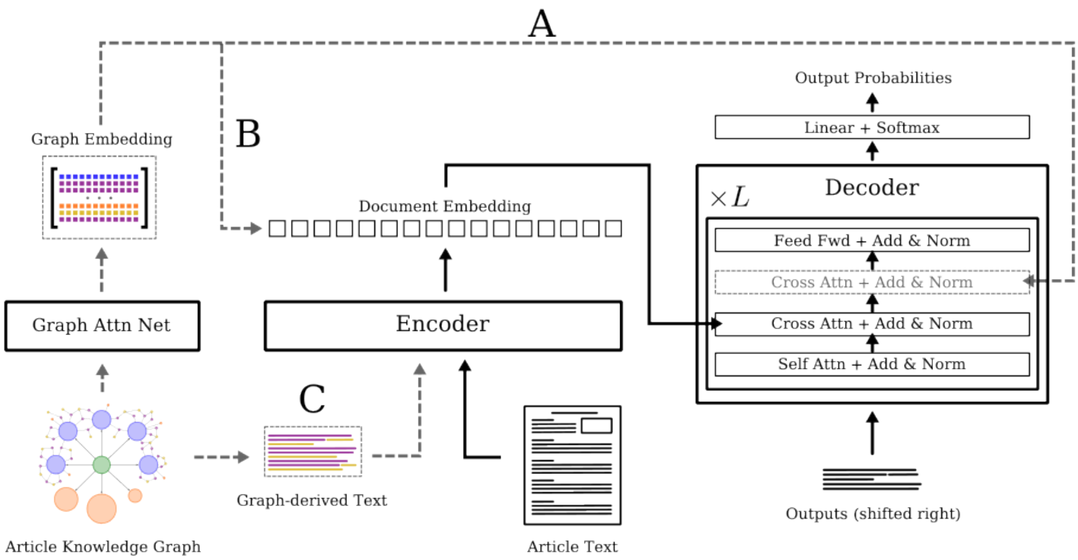


Figure 4.3: *Visualisation of how our various knowledge-enhancement approaches incorporate external knowledge from article knowledge graphs into a transformer-based encoder-decoder architecture (as described in §4.4). **A**) Decoder cross-attention, **B**) Document embedding enhancement, **C**) Article text augmentation.*

4.4 Knowledge-Enhanced Lay Summarisation Approaches

We investigate the effectiveness of three different methods for incorporating external knowledge from article graphs into encoder-decoder-based summarisation models. Our experiments are carefully designed so as to target a distinct aspect within the model architecture (i.e., the input, the encoder, and the decoder) with each selected method, taking inspiration from models that have recently been proven effective in the domain of news article summarisation Zhu et al. (2021a); Pasunuru et al. (2021). Figure 4.3 provides a visualization of how each of these approaches fits into this architecture. To allow the ingestion of the full input article, we make use of Longformer Encoder-Decoder (Beltagy et al., 2020) as our base model for all experiments. This BART-based model replaces standard transformer self-attention with a sparse attention mechanism that scales linearly to the sequence length, enabling the processing of longer texts (such as research articles). We describe each knowledge-enhancement approach in detail below.

(A) Decoder cross-attention. We make use of a Graph Attention Network (GAT) (Veličković et al., 2017) to obtain an embedding of the article graph $\mathcal{G}$ in parallel with the base model encoder.

$$H^{\mathcal{G}} = \text{GAT}(\mathcal{G}) \tag{4.1}$$

These Graph Neural Network (GNN) models produce a final set of node features (i.e., a graph embedding) by using attention layers to efficiently aggregate over the features of neighboring nodes, and are widely used in current literature to aggregate graph-based information for NLG tasks (Huang et al., 2020; Zhu et al., 2021b; Liu et al., 2021). During the decoding phase, we follow previous works (Zhu et al., 2021a) by forcing our model to attend to the KG embedding $H^{\mathcal{G}}$. Specifically, in every transformer layer of the decoder, we include a second cross-attention mechanism that occurs directly after the standard encoder cross-attention (see arrow A in Figure 4.3) and attends to the output of the GAT-based model.

(B) Document embedding enhancement. Again, we obtain an embedded graph representation $H^{\mathcal{G}}$ using the GAT model, but rather than attending to graph embeddings during decoding, we follow Pasunuru et al. (2021), combining the embedded node information into the final document embedding (i.e., the output of the encoder). Specifically, we concatenate the document and graph embeddings, before passing them through an additional encoder layer. For a given input document X, this process can be formalised as follows:

$$H^X = \text{Encoder}(X) \tag{4.2}$$

$$H^C = [H^X; H^{\mathcal{G}}] \tag{4.3}$$

$$H^* = p \cdot \text{EncoderLayer}(H^C) + (1 - p) \cdot H^C \tag{4.4}$$

where H^X is an embedding of document X, H^C is the concatenated document and graph embeddings, H^* is the final 'enhanced' document embedding that is subsequently attended to during decoding, and p is a scaling factor controlling the extent to which the additional encoder layer output is incorporated in the final enhanced document embedding. Note that p is treated as a hyperparameter to the model, for which a value of 0.25 was found to provide the strongest validation set performance.[5]

(C) Article text augmentation. We also experiment with simply augmenting the input text with textual explanations of the key concepts (and their relations) derived from the graph. Whilst this may be arguably the most 'natural' way for a PLM-based model to interpret external information, this approach leads to an exponential increase in the number of tokens required to describe each element, hence it is restricted to a set of few concepts. We select only those entities which are likely to be most central to the topic of the article (and, therefore, relevant to the lay summary). Specifically, we take the concept nodes that are mentioned in the article abstract and use the graph relations and retrieved definitions to provide a textual explanation of these salient concepts and their semantic types, which is then prepended to the article text. This takes the format of "*{concept_name} = {concept_definition}. {concept_name} is a {semtype_name}.*" repeated for each selected concept, followed by semantic type definitions formatted "*{semtype_name} = {semtype_definition}*" repeated for all mentioned semantic types.

4.5 Experimental Setup

4.5.1 Data

We derive knowledge graphs for all articles in eLife (Goldsack et al., 2022a), a dataset for biomedical lay summarisation containing 4,828 article-summary pairs. Target summaries are expert-written lay summaries (i.e., summaries with a non-expert target audience) and inputs are the full text of the corresponding biomedical research articles. As explained in §4.1, we believe this task to be particularly suitable for domain knowledge augmentation due to the contrast in the level of expertise of

[5]We also find that a large value of p causes significant degradation in performance, suggesting that the original document information is lost.

the target audience between source and target which causes a discrepancy in the language used (specifically, reducing or explaining jargon terms) and the level of background information required.[6]

4.5.2 Baselines

As a baseline model, we include BART (Lewis et al., 2020), the state-of-the-art benchmark reported by Goldsack et al. (2022a) for eLife, as well as in other previous lay summarisation works (Guo et al., 2021). Additionally, we include the reported performance of $\text{BART}_{scaffold}$ (Goldsack et al., 2022a), a variant of BART trained to produce both the abstract and lay summary of an article in a controlled setting, which is equivalent to the model proposed by Luo et al. (2022a).[7]

4.5.3 Implementation and Training

Each knowledge-based approach is implemented by manually adapting the Longformer implementation from Huggingface (Wolf et al., 2020) and, following previous work on lay summarisation (Chandrasekaran et al., 2020; Luo et al., 2022b; Goldsack et al., 2022a), uses the full article text as input. For GAT-based models, we make use of the Deep Graph Library package (Wang et al., 2019) to implement a 3-layer GAT with 4 attention heads at each layer. For article graphs, we vary our node initialisation approach based on node type (as defined in §4.3.2). Specifically, we initialise concept and semantic type node features, with the embeddings of their textual definitions; document and section nodes with the embeddings of their title text (with title metadata nodes being subsequently ignored); and remaining metadata nodes (publication date and keywords) with embeddings on their textual content. All embeddings are generated using SciBert (Beltagy et al., 2019), a language model specifically trained on research papers from Semantic Scholar (Ammar et al., 2018) that is widely used for scientific data (Cohan et al., 2019; Cai et al., 2022; Goldsack et al., 2023b). Furthermore, all node embedding features are concatenated with

[6]This discrepancy is evidenced in the analysis provided by Goldsack et al. (2022a).

[7]Note that the original code for Luo et al. (2022a) is not yet available at the time of writing and their results are reported on a different dataset and thus are not comparable.

Model	Relevance				Readability		Factuality
	R-1↑	**R-2↑**	**R-L↑**	**BeS↑**	**CLI↓**	**DCRS↓**	**BaS↑**
BART	46.57	11.65	43.70	84.94	11.7	9.36	-2.39
BART$_{scaffold}$	45.28	10.99	42.51	84.65	11.32	9.19	-2.57
Longformer	47.23	13.20	44.44	85.11	11.72	9.09	-2.56
– text-aug	**48.58***	14.24*	**45.71***	**85.4***	10.94*	8.72*	-2.45*
– doc-enhance	48.10*	**14.43***	45.52*	85.33*	**10.72***	**8.50***	**-2.35***
– decoder-attn	48.30*	13.93	45.45*	85.39*	10.99*	8.75*	-2.48*

Table 4.2: *Average performance of models on eLife test split.* ***R*** *= ROUGE F1,* ***BeS*** *= BERTScore F1,* ***CLI*** *= Coleman-Liau Index,* ***BaS*** *= BARTScore. * denotes that KG-enhanced model results are statistically significant with respect to the base model (Longformer) by way of Mann-Whitney U test.*

one-hot features according to node type, as well as Random Walk Positional Encodings (Dwivedi et al., 2021). Following initialisation with `allenai/led-base-16384` checkpoint, we train all models on A100 GPUs and retain the checkpoint with the best validation set performance (more details provided in the Appendix).

4.5.4 Evaluation Setup

We conduct both automatic and human evaluations to provide a comprehensive assessment of how each knowledge-enhancement method affects the overall performance.

Automatic evaluation. For each model, we report the average scores of several automatic metrics on the test split of eLife. As is common practice, we report widely-used summarisation metrics: BERTScore (Zhang et al., 2019) and the F1-scores of ROUGE-1, 2, and L (Lin, 2004).

To assess the readability of generated summaries, we report Flesh-Kincaid Grade Level (FKGL) and Dale-Chall Readability Score (DCRS), both of which compute an estimate of the US-grade level required to comprehend a text.[8]

Additionally, we evaluate factuality using BARTScore (Yuan et al., 2021), which has been shown to have a strong alignment with human judgments of factual consistency in a recent study focusing specifically on long documents (Yee Koh et al.,

[8]Computed using the `textstat` package.

2022). Following Yee Koh et al. (2022), we adapt BARTScore to use Longformer (thus allowing it to process the entire document as input) and fine-tune it on eLife.

Human evaluation. To provide a comprehensive assessment of the summaries generated by each knowledge-enhanced model, we conduct a human evaluation focusing on readability and factuality. Specifically, making use of 5 randomly sampled articles from the eLife test set, we ask human judges to evaluate each sentence within a generated summary along the following binary criteria: 1) *Factuality* - is the sentence factually correct (with respect to the source article); and 2) *Readability* - would a layperson be able to understand this sentence.[9] To help determine the factuality of the sentence, the annotator has access to the PDF of the source article as well as the reference lay summary.[10]

4.6 Experimental Results

4.6.1 Automatic Evaluation

Table 4.2 presents the performance of different models using the described automatic evaluation metrics on the test set of eLife. In addition to applying KG-enhancement methods in isolation, we also experiment with combining different methods, which we largely find to be detrimental to model performance. Discussion and results (Table C.3) of combined methods are provided in the Appendix. We discuss the performance of individually applied methods below, focusing on each aspect of automatic evaluation in turn.

Relevance Longformer can be seen to outperform the standard BART model in terms of relevance metrics, indicating that processing the entire document provides some benefit for lay summarisation. Additionally, all three knowledge enhancement methods significantly obtain improved scores across almost all relevance metrics

[9] An average of 68.5 sentences evaluated per model.

[10] Following Yee Koh et al. (2022), we encourage judges to use text-based search within the article to quickly identify relevant passages rather than asking them to read each article in full, reducing the cognitive burden placed upon them.

(with the exception of R2 for the 'decoder attention' model). This provides a strong indication that the addition of graph-based domain knowledge provides models with relevant external information, enabling them to produce lay summaries that are closer in resemblance to the high-quality references.

Readability For readability metrics, it can first be noted that Longformer-based models obtain lower CLI and DCRS scores than those BART-based models. The calculation of CLI is based on the number of characters, words, and sentences it contains, whereas DCRS is based on the frequency of "familiar" (i.e., commonly-used) words, suggesting that Longformer produces summaries that are less syntactically and lexically complex.

We observe that the application of all knowledge enhancement methods results in improved scores for both metrics, with the document enhancement approach achieving the largest gains. This indicates that all knowledge enhancement methods are able to successfully influence the phrasing and structure of the summaries being generated by increasing the usage of more common (i.e., less technical) terminology. As reported in Table 4.1, the average CLI and DCRS scores for the reference lay summaries of eLife are 12.51 and 8.83, respectively.

Factuality For BARTScore (BAS), we again see a statistically significant improvement over the base Longformer model for all KG-enhancement methods, with the greatest improvement being obtained by the doc-enhance method. In order to gain further insight into these results, we also calculate the BARTScore values obtained by reference summaries, getting a mean score of -2.39, which is similar to that of all tested models (and identical to that of BART). This suggests that all models are able to produce summaries with generative probabilities similar to that of the reference summaries. However, given that one model (doc-enhance) actually outscores the reference summaries, further analysis is needed to gain an understanding of the difference in the factual correctness of summaries produced by each method, for which we turn to our human evaluation.

Model	**#**	**Readability**	**Factuality**
Longformer	73	78.08	60.96
– text-aug	65	96.92*	68.46
– doc-enhance	67	97.01*	55.97
– decoder-attn	69	95.65*	63.77

Table 4.3: *Average percentage of generated sentences positively classified by judges for each high-level binary criteria. # = total number of sentences generated across all summaries. * denotes that KG-enhanced model results are statistically significant with respect to the base model (Longformer) by way of Mann-Whitney U test.*

4.6.2 Human Evaluation

Given the challenging and time-consuming nature of evaluating the factuality of technical biomedical sentences against the source article, we carefully plan our human evaluation so as to ensure reliability in our results. We employed two annotators to evaluate generated sentences following the procedure laid out in §4.5, both of whom are experts in NLP and familiar with common model shortcomings (e.g., hallucinations). Table 4.3 presents the total percentage of sentences that were positively classified for both readability and factuality averaged across evaluators, who achieve a Cohan's κ of 0.42.

Discussion The results in Table 4.3 suggest that the application of all KG-enhancement methods causes a notable increase in the readability of the text produced by the model, with all models scoring significantly higher than the base Longformer model. Alternatively, the results for factuality show that, although there is a slight variance in performance between KG methods, none of them are judged to be by a statistically significant margin. These results indicate that all methods are able to effectively introduce relevant external information into the model, enabling it to produce text that is easier for a lay audience to comprehend without significantly compromising the factual correctness of the base model.

Case Study To gain a better insight into how knowledge enhancement methods influence the readability of generated summaries, we present a case study in Figure 4.4 in which we compare the explanations of specific technical concepts generated

a. [Meiosis]
Longformer - During meiosis, the DNA in one of the chromosomes is copied and then the two copies are recombined so that each new generation will have a single copy of the gene that encodes the protein encoded by that gene.**1/2**
w/ text-aug - ... a process known as meiosis ... two copies of each chromosome are then exchanged between the newly formed cells, which results in a unique set of genes being passed on to the next generation.**2/2**

b. [Glabrous skin / Mechanoreceptors]
Longformer - The orientation of an object depends largely on how its edges activate mechanoreceptors in the glabrous skin of the fingertips.**0/2**
w/ doc-enhance - The fingertip's surface is covered by a ... layer of skin known as the glabrous skin.**2/2** These cells are responsible for sensing touch, and they are also responsible for detecting the orientation of objects that touch them.**2/2**

c. [Slow wave sleep]
Longformer - Most studies of sleep have focused on ... slow wave sleep, in which the brain's activity alternates between periods of alternating periods of slow and fast sleep.**0/2**
w/ decoder-attn - Slow wave sleep is characterized by rhythmic waves of electrical activity in the brain, which are thought to be part of the process by which the brain consolidates memories.**2/2**

Figure 4.4: *A case study comparing how the application of each method affects the explanation of specific technical concepts within the human evaluation sample. Colours and superscript are used to denote the number of evaluators who judged the sentence as readable for a lay audience (e.g.,* **2/2** *= 2 out of 2 evaluators).*

by KG-enhanced models and the base Longformer model, alongside their annotator ratings.[11]

These examples demonstrate how KG-enhancement methods improve the model's handling of technical concepts, thus making them easier to understand for a lay reader. Specifically, examples show how methods can influence the model to generate an explanation in instances where the base model fails to provide one (b) or improve the explanation in instances where the base model's is difficult to understand (a and c).

4.7 Conclusion

This papers presents the first study on the use of knowledge graphs to enhance lay summarisation, augmenting the biomedical lay summarisation dataset eLife with article-specific knowledge graphs containing domain-specific external knowledge on relevant technical concepts. We compare three distinct approaches for incorporating graph-based knowledge into encoder-decoder summarisation models, placing an emphasis on the readability and factual correctness of the generated output. Our

[11] An extended version of this case study is given in Figure C.1 in the Appendix.

results suggest that integrating external knowledge has the potential to substantially improve lay summarisation, particularly for the generation of readable text and explanation of technical concepts. We would like to see future work investigate the use of additional graph representations, as well as their integration into larger models that adopt different architectures (e.g., decoder-only).

Limitations

One possible limitation of our work is derived from the use of resources from UMLS (i.e., UMLS concept names, semantic types and relations, definitions, etc.). Accessing these resources requires an individual license with the US National Library of Medicine (NLM), and their subsequent distribution is restricted by this license agreement. Therefore, it is likely that we will have to confirm the license status of those who wish to have access to the knowledge-graph resources used in this work. In an attempt to reduce any potential impacts this will have on the ability to share our resources, we only make use of only a select number of vocabularies less restrictive licences. More details on the vocabularies used are provided in the Appendix.

4.8 Acknowledgements

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

4.9 Updated Limitations

Additional limitations of this work are outlined below for the purpose of this thesis:

- Within this work, the terms "factual consistency" and "factuality" are used inter-changeably for the discussion of our evaluation and results. These terms are intended to represent identical meanings within this thesis - how factually aligned the generated summary is with the source document. This reflects their somewhat inconsistent usage in broader summarisation literature.
- For the human evaluation included in this work, it would have been desirable to recruit Biomedical domain experts as judges. In the absence of domain experts, judges with expertise in NLP and familiarity with common errors made by encoder-decoder neural networks were prioritised.
- Although not stated in this chapter, human annotators were recruited from within the University of Sheffield Computer Science Department, and were compensated in the form of Amazon vouchers.
- In light of developments in the field of summarisation since the publication of this work, including an LLM based approach as a baseline would now be appropriate.

Chapter 5

Publication IV: Leveraging Large Language Models for Zero-shot Lay Summarisation in Biomedicine and Beyond

Abstract

In this work, we explore the application of Large Language Models to zero-shot Lay Summarisation. We propose a novel two-stage framework for Lay Summarisation based on real-life processes, and find that summaries generated with this method are increasingly preferred by human judges for larger models. To help establish best practices for employing LLMs in zero-shot settings, we also assess the ability of LLMs as judges, finding that they are able to replicate the preferences of human judges. Finally, we take the initial steps towards Lay Summarisation for Natural Language Processing (NLP) articles, finding that LLMs are able to generalise to this new domain, and further highlighting the greater utility of summaries generated by our proposed approach via an in-depth human evaluation.

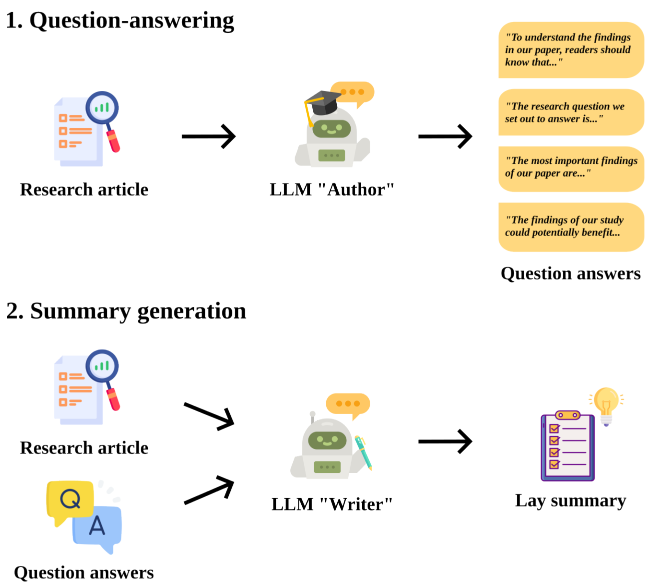


Figure 5.1: *A Visualisation of our two-stage Lay Summarisation framework, based on the real-life process of the eLife Journal.*

5.1 Introduction

The goal of Lay Summarisation is to create a summary that effectively communicates the key concepts and findings of a technical article to a non-expert audience (Goldsack et al., 2022a). To achieve this, a well-written lay summary must, without hype or exaggeration, convey these findings clearly and concisely using everyday language, ensuring that it remains accessible to readers without specialised knowledge (King et al., 2017). In recent years, the task of automatic Lay Summarisation has attracted increased interest in the research community, driven by the need to broaden access to scientific research. However, due to a lack of availability of publically accessible reference data, research for this task has been restricted to a few select domains, with Biomedicine being the most prominent.

In recent years, the rise of Large Language Models (LLMs) has drastically altered the research landscape for Natural Language Generation (NLG), including for tasks such as Summarisation (Goyal et al., 2022). For Lay Summarisation specifically, the results of the recent BioLaySumm 2023 shared task suggest that LLMs have the potential to also push the boundaries of task performance (Goldsack et al., 2023a), with two of the top three submissions adopting them within their approach (Turbitt et al., 2023; Sim et al., 2023).

At the forefront of this changing landscape are current LLMs' significantly improved zero-shot capabilities, emerging from the increased scale of models (Kaplan et al., 2020) and breakthrough techniques such as instruction-tuning (Wei et al., 2022). Importantly, these zero-shot capabilities have significantly widened the scope of LLM applications, enabling us to explore tasks that were previously impossible due to a lack of training data. For Lay Summarisation, these models now provide the opportunity to explore domains outside of Biomedicine and improve accessibility to research in all domains. However, little is known about how best to utlise LLMs to generate lay summaries in a zero-shot setting, or how best to evaluate generated summaries in the absence of references.

In this work, we attempt to address these questions and lay the groundwork for future Lay Summarisation research in unexplored domains. We propose a novel two-stage prompting framework for zero-shot Lay Summary generation inspired by real-world practices, before thoroughly evaluating the performance of our method in two domains: Biomedicine and Natural Langauge Processing (NLP). Our results show that summaries generated using the proposed framework are increasingly preferred by both human and LLM judges as we increase the size of the underlying LLM. Furthermore, we find that LLMs can effectively generalise to the previously-unexplored domain of NLP articles, and that the summaries generated with our proposed framework provide lay readers with a more well-rounded understanding of article contents in this domain.

5.2 Task Decomposition Framework

We explore the use of a two-stage framework for Lay Summarisation, based on the practices of the journal eLife. In a recent study, Goldsack et al. (2022a) found that the lay summaries from eLife, crafted internally by journal editors and hired writers in collaboration with article authors, were notably more accessible than the author-written lay summaries from PLOS (the Public Library of Science).[1] To

[1]eLife lay summaries are on average longer, more readable, and more abstractive than those of PLOS, positively impacting their comprehensibility.

produce their lay summaries, eLife adopts a multi-stage process involving different actors. King et al. (2017) describe this process, whereby the paper authors are asked to answer a set of questions concerning the motivation and impact of their work, aiming to extract information that may not be explicit with the article itself but is essential to lay audience understanding. Upon inquiry with the team at eLife, we were provided with these questions, which inform our approach.

As visualised in Figure 5.1, our framework is based on the introduction of an additional "question-answering" stage in the summary generation process, with the answers generated from this stage used as an additional data source during summary generation. By introducing this additional stage, we aim to simulate the process that produces the high-quality lay summaries of eLife, and make explicit the important questions that should be addressed in the final summary. A detailed description of each stage is provided below:

1. Question-answering Given the article text, the LLM is asked to play the role of the paper author and answer four questions about the article. As previously mentioned, these questions are derived from those that eLife asks of authors. Specifically, the questions are: 1) What background information would someone who is completely unfamiliar with your field need to know to understand the findings in your paper? 2) What exact research question did you set out to answer and why? 3) What are the most important findings of your paper? and 4) Who might eventually benefit from the findings of your study, and what would need to be done before we could achieve these benefits?[2]

2. Summary generation Given the article text and the "author's" answers from the previous stage, we instruct the LLM to play the role of a writer and generate the lay summary based on guidelines derived from the analyses of Goldsack et al. (2022a). These guidelines describe the desirable length, language, and structure of the generated summary.

[2]Alongside each question, additional guidance for answering is also presented. We provide a full explanation of this, and the prompts used for each agent, in Appendix D.1.

Model	# Params	QA	Relevance				Readability		Factuality	H2H	
			R-1↑	R-2↑	R-L↑	BeS↑	FKGL↓	DCRS↓	BaS↑	PoLL↑	PoH↑
Longformer	124M	✓	47.23	13.20	44.44	85.11	11.72	9.09	-2.56	-	-
Phi2	2.7B	✗	41.66	9.10	39.39	83.27	13.08	9.05	-3.06	0.65	0.80
		✓	35.06	7.97	32.51	83.30	12.61	9.62	-3.77	0.35	0.20
Mistral	7B	✗	44.02	10.39	41.26	84.15	14.08	10.42	-3.60	0.60	0.75
		✓	44.60	9.95	41.77	84.29	14.17	10.19	-3.58	0.40	0.25
Mixtral	46B	✗	45.59	11.02	42.85	84.17	14.08	9.69	-3.06	0.40	0.45
		✓	45.49	10.41	42.90	83.95	13.92	9.57	-3.12	0.60	0.55
LLAMA3	70B	✗	45.59	11.22	43.03	84.72	10.37	8.43	-3.34	0.20	0.35
		✓	44.90	10.68	41.93	84.93	11.99	9.34	-3.27	0.80	0.65
DBRX	132B	✗	44.77	11.09	42.04	84.25	13.48	9.69	-3.10	0.15	0.30
		✓	44.18	9.94	41.54	84.29	14.31	10.17	-3.24	0.85	0.70
DeepSeek	671B	✗	45.31	10.48	42.27	84.37	12.40	10.42	-4.26	0.05	-
		✓	43.14	8.33	40.21	84.31	11.57	10.32	-4.44	0.95	-

Table 5.1: *Average performance of models on eLife test split.* ***R*** *= ROUGE F1,* ***BeS*** *= BERTScore F1,* ***FKGL*** *= Flesch-Kincaid Grade Level,* ***DCRS*** *= Dale-Chall Readability Score,* ***BaS*** *= BARTScore,* ***PoLL*** *= Panel of LLM evaluators,* ***PoH*** *= Panel of Human evaluators.*

5.3 Biomedical Lay Summarisation

To provide a comprehensive assessment of the performance of our proposed method on an existing dataset, we conduct several experiments using the test set of eLife dataset (Goldsack et al., 2022a), which contains 241 Biomedical research articles paired with expert-written lay summaries.

We compare our two-stage method against a standard one-stage lay summary generation prompt across several popular open-source LLMs of various sizes: Phi2, Mistral-7B, Mixtral 8×7B, LLAMA3-70B, DBRX, and DeepSeek. For each LLM, we generate lay summaries using both our proposed two-stage QA-based prompting method and a generic lay summary generation prompt: "Generate a summary of the following article that is suitable for non-experts".[3]

Evaluation We adopt the automatic evaluation metrics used in the recent BioLaySumm Shared Task (Goldsack et al., 2023a), assessing models along the dimensions of Relevance, Readability, and Factuality. Specifically, we include ROUGE-1, 2, and L (Lin, 2004) and BERTScore (Zhang et al., 2019) to measure relevance to references. For readability, we Flesch-Kincaid Grade-Level (FKGL) and Dale-Chall Readability Score (DCRS) are used. Finally, a version of BARTScore (Yuan et al., 2021), adapted to process long inputs (using sparse attention) and fine-tuned on the

[3] An additional discussion of prompts is given in Appendix D.1 and details of additional experiments, including an input selection analyses and ablation study, is given in Appendix D.2.

eLife dataset, is used to measure Factuality.

Finally, we perform a head-to-head comparison using a sample of 20 randomly selected test set instances for each LLM, using a panel of both lay human judges and LLM judges. Specifically, for a given sample instance, judges are provided with the lay summaries generated by each method and asked to decide which summary they find more useful as a layperson wanting to understand the findings and significance of the article.[4] For LLMs judges, we follow Verga et al. (2024) and employ a panel of 3 smaller LLMs - namely Command R, Haiku, and GPT-3.5. For human judges, we employ 4 lay people with no experience in biomedicine. In both cases, judges indicate their preferences individually and preferences are then aggregated using majority vote.[5] We report the proportion of each summary type preferred by both sets of judges.

Results and discussion Table 5.1 presents the metrics scores obtained by both the standard one-stage prompt and the proposed two-stage prompt for each LLM (where a ✓ in the **QA** column denotes the two-stage prompt). Interestingly, we observe different patterns occuring for smaller ($\leq$ 7B) and larger ($\geq$ 46B) LLMs. For the smaller LLMs - Phi2 and Mistral - we see a significant drop in the scores of automatic metrics when the proposed two-stage framework is used, particularly ROUGE scores. Additionally, we find that both human and LLM judges prefer the standard one-stage prompt in this setting. However, for some larger LLMs - Mixtral and LLAMA3 - we observe that, despite roughly comparable metric scores, both human and LLM judges tend to prefer summaries generated by the proposed two-stage framework. Furthermore, the extent to which judges prefer these summaries increases as model size increases.

Overall, these results indicate that: 1) the proposed two-stage framework produces lay summaries that are increasingly preferred by lay people as model size increases, but presents too complex a task for smaller LLMs; 2) automatic metrics such as

[4]To mitigate potential positional bias, we swap the order in which summaries are provided to judges randomly, such that 50% of each ordering is used.

[5]When a tie in voting occurs between human judges, we take the preference of the judge who obtained the highest agreement with other judges over all instances. A further discussion on evaluator agreement is provided in Appendix D.2.

Model	QA	Readability		Factuality	H2H
		FKGL↓	DCRS↓	BaS↑	PoLL↑
Mixtral	✗	13.88	9.44	-3.17	0.40
	✓	14.59	9.48	-3.28	0.60
LLAMA3	✗	10.60	8.37	-3.53	0.30
	✓	12.47	9.47	-3.59	0.70
DBRX	✗	13.23	9.84	-3.22	0.25
	✓	14.91	10.07	-3.34	0.75

Table 5.2: *Average performance of models on ACL paper set.* ***FKGL*** *= Flech-Kincaid Grade Level,* ***DCRS*** *= Dale-Chall Readability Score,* ***BaS*** *= BARTScore.*

ROUGE fail to capture all aspects of the summary that inform both human and LLM preferences for higher quality summaries.

5.4 Lay Summarisation for NLP

In this section, we explore the application of our framework to NLP, a domain with a high publication rate that has recently been gaining significantly more attention from non-expert audiences (i.e., the general public). We collect 100 randomly sampled articles from the proceedings of ACL 2023 to use as a test set. To delve deeper into the quality of summaries produced by each zero-shot prompt, we conduct a reference-less automatic evaluation, followed by a carefully designed human evaluation that utilises both lay people and experts to provide a comprehensive assessment of lay summary quality.

Evaluation In the absence of reference lay summaries, we repeat our evaluation protocol from Section 5.3 using only reference-less metrics. To gain further insight into the specific differences between summaries generated in each setting, we use our lay human evaluators to perform an alternative human evaluation using a random sample of 20 articles from the ACL paper set, carefully designed to assess how the summaries differ in their utility. Specifically, our 4 lay participants are tasked with answering the following set of questions about the article, based on only the lay summary: 1) What problem is the article tackling? 2) How did the authors tackle the problem? 3) What are the key findings of the article? 4) Why are these findings

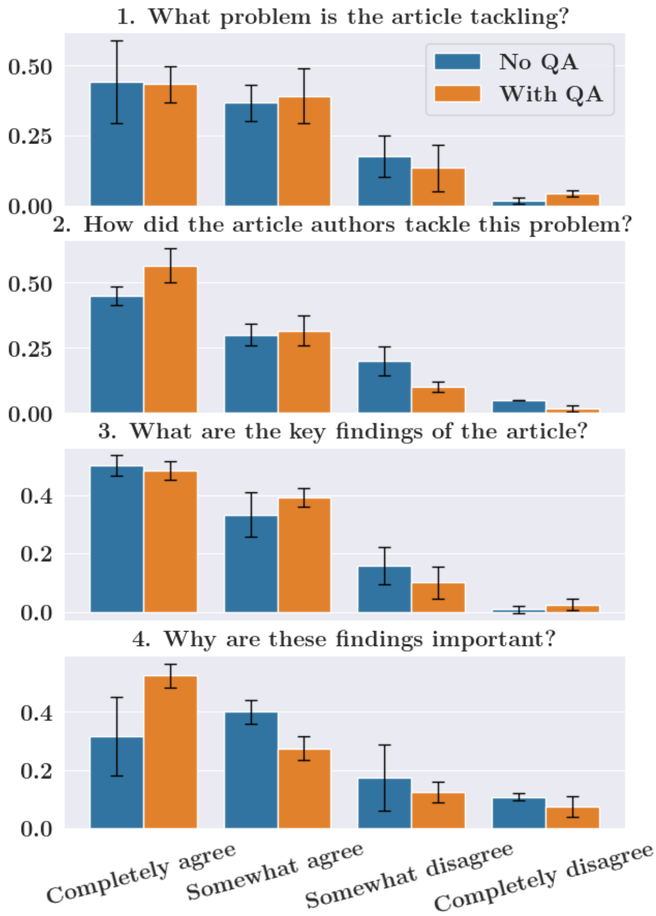


Figure 5.2: *Human evaluation results as the proportion of answer votes for each question.*

significant?[6] Subsequently, we employ 3 NLP experts to judge the answers of lay participants by classifying an answer based on the extent to which they agree with it.[7] This is done using the labels "Completely agree", "Somewhat agree", "Somewhat disagree", and "Completely disagree".

Results and Discussion Table 5.2 presents the automatic metric scores obtained by the larger LLMs, for which human judges preferred summaries generated by the proposed framework. We find that scores obtained for NLP articles appear largely comparable with those obtained for biomedical articles (Table 5.1), suggesting that

[6] To avoid any cross-contamination of knowledge between generated summaries, each lay annotator only assesses 1 generated summary per article. To ensure a fair comparison between models, our 4 annotators are split into 2 groups of 2, with each group assessing 10 summaries generated by each model.

[7] All expert judges are PhD students studying NLP.

both approaches generalise to the NLP domain.

Figure 5.2 presents the results of our human evaluation, visualising the total proportion of votes both models received for each classification label. Firstly, it is evident that the answers of lay participants for both methods receive overwhelmingly more positive classifications (agree) than negative classifications (disagree) from expert annotators, suggesting that both methods produce reasonably good lay summaries when applied to NLP articles. However, we find that the two-stage approach (with QA) almost consistently receives fewer votes for the "disagree" categories across all questions (except for question 1), indicating that these summaries are generally better in enabling laypeople to provide answers that demonstrate some level of understanding. Furthermore, particularly significant gains for positive labels can be seen for questions relating to article methodology and significance (2 and 4, respectively; suggesting that answers extracted during the QA stage of the framework are particularly pertinent for these aspects.

5.5 Conclusion

In this work, we study the utility of LLMs for zero-shot Lay Summarisation and proposing a novel two-stage framework inspired by professional practices. Our experiments on biomedical articles suggest that the benefit provided by our framework increases in line with the size of the LLM, as demonstrated by increased preference of lay human and LLM judges. However, standard automatic metrics such as ROUGE fail to capture such preferences. Our experiments on NLP article attest to the ability of models to generalise to an unexplored domain, and further illustrate how the proposed approach enables a more well-rounded understanding of article contents.

5.6 Limitations

- In this work, we have attempted to comprehensively address various common settings in which LLMs may be deployed for the task of Lay Summarisation. However, due to the inherent flexibility of prompt-based interaction, there will undoubtedly be variations that remain unexplored.

- Furthermore, it would have been desirable to assess the performance of an even wider range of models, particularly given the sheer number of LLM that are currently available and the rate at which they are released. However, in our model selection, we aim to ensure that we were assessing the performance of several widely-used models.

- As the DeepSeek model results were collected and added at a later date to other models, it was unfortunately impossible to include it within the preference-based human evaluation.

- In assessing the ability of models to generalise to non-biomedical articles, we could have selected any from a large number of viable domains. NLP was selected based on the its high rate of publication and high levels of interest it is currently attracting from lay audiences.

- Despite the fact that LLM and human judges both display an increased preference for the proposed two-stage framework, an additional analysis of inter-annotator agreement (given in Appendix D) highlights relatively low of sample-level agreement.

Chapter 6

Publication V: From Facts to Insights: A Study on the Generation and Evaluation of Analytical Reports for Deciphering Earnings Calls

Abstract

This paper explores the use of Large Language Models (LLMs) in the generation and evaluation of analytical reports derived from Earnings Calls (ECs). Addressing a current gap in research, we explore the generation of analytical reports with LLMs in a multi-agent framework, designing specialized agents that introduce diverse viewpoints and desirable topics of analysis into the report generation process. Through multiple analyses, we examine the alignment between generated and human-written reports and the impact of both individual and collective agents. Our findings suggest that the introduction of additional agents results in more insightful reports, although reports generated by human experts remain preferred in the majority of cases. Finally, we address the challenging issue of report evaluation, we examine the limitations and strengths of LLMs in assessing the quality of generated reports in different settings, revealing a significant correlation with human experts across

multiple dimensions.

6.1 Introduction

Earnings Calls (ECs), critical quarterly meetings conducted by publicly traded companies to discuss financial performance with professional analysts, have been extensively studied for various prediction tasks. These tasks include volatility prediction (Sawhney et al., 2021; Niu et al., 2023), analyst decision prediction (Keith and Stent, 2019), financial risk prediction (Qin and Yang, 2019), and earnings surprise prediction (Koval et al., 2023), highlighting ECs' significance in investment decision-making. Because ECs' typical duration of about one hour, another prominent research area in this domain is summarizing lengthy EC transcripts (Mukherjee et al., 2022). Post-EC, two types of summaries emerge: *Journalistic Reports*, in which journalists concisely summarize the key financial takeaways from the meeting, and *Analytical Reports*, in which professional analysts offer a considerably more extensive and multifaceted analysis of meeting events, financial performance, and implications on investment strategies. Whilst the automatic generation of journalistic reports has been addressed in previous studies (Mukherjee et al., 2022), no work to our knowledge has explored the task of generating analytical reports, despite numerous potential benefits. For example, the automatic generation of analytical reports could reduce the burden placed on analysts, enable the immediate distribution of key information to a broad range of stakeholders, and introduce novel insights through scalable interpretation of complex data. However, given the inherent complexity of generating analytical reports, success in this challenging task requires a methodology that can enable an in-depth analysis across multiple varied and important technical aspects, such as expectations on future operations and managers' attitudes during ECs.

One promising candidate for such an approach comes in the form of multi-agent frameworks: an exciting avenue of recent research that explores how multiple role-playing LLM "Agents" can be deployed to cooperatively solve a task. When deployed for generation tasks such as Software development (Qian et al., 2023) and Trivia-based Creative Writing (Wang et al., 2024), the introduction of role-based division has enabled complicated requirements to be broken down into simple subtasks and processes, reducing the cognitive and contextual burden placed on the underlying models. Furthermore, the utilization of role-playing LLM agents has provided expanded opportunities for domain specialization, the leveraging of external data/tools, and the incorporation of diverse viewpoints, all of which could add significant value to the generation of analytical reports. Notably, such an approach also bears a closer resemblance to a human writing process which, from a cognitive science perspective, is a complex, cyclic, and multi-step procedure, often requiring strategic discourse planning and multiple iterations to effectively achieve a communicative goal (Flower and Hayes, 1981).

In this work, we explore the utility of a multi-agent framework for generating analytical reports using LLMs. We establish the details of our novel multi-agent framework and specially designed agents (§6.2), before performing a thorough characterization of the generated reports under different settings (§6.3), highlighting key differences to human-authored reports and the additional insights offered by feedback agents. Finally, we assess the capabilities of LLMs in evaluating the quality of generated reports (§6.4), establishing both promising directions and limitations, and laying the groundwork for future research on this task.

Overall, the following three research questions (RQ) are addressed:

(RQ1): How do generated analytical reports differ from human-authored analytical reports?

(RQ2): Can a multi-agent approach be used to generate more insightful analytical reports?

(RQ3): How effective are LLM-based evaluation methods in assessing the quality of analytical reports?

6.2 Multi-Agent Report Generation

We explore a novel framework for generating analytical reports through collaborative multi-agent conversation, leveraging the capacity of LLMs to refine their output

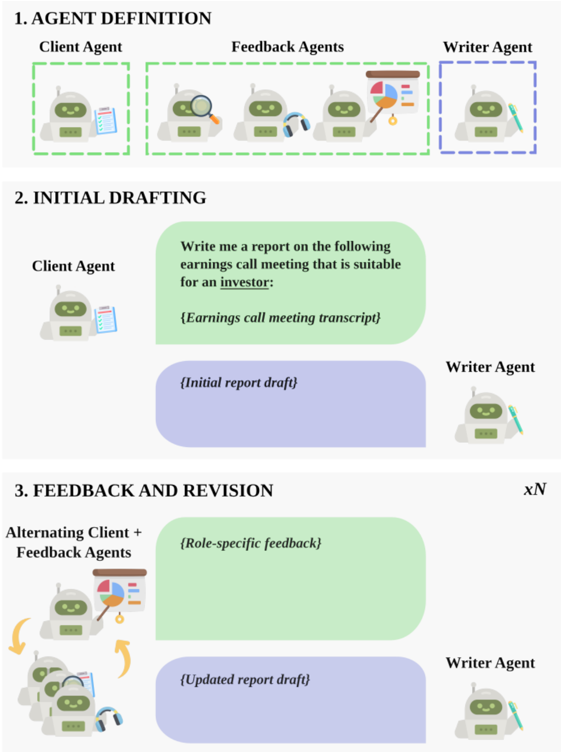


Figure 6.1: *An overview of our multi-agent framework.*

based on feedback. This framework, developed using Microsoft's AutoGen (Wu et al., 2023), assigns each LLM-powered agent a distinct role through an initialization prompt, dictating their contribution to the conversation. For all experiments and agents, we employ the `gpt-4-1106-preview` model via the ChatGPT API as the underlying LLM. Figure 6.1 illustrates the framework, whereby a single agent with the role of **Writer** (✍) is tasked with drafting and revising the report, guided by feedback from other agents. The process encompasses three stages:

Stage 1. Agent Definition Before generation commences, it is imperative to define the agents involved: 1) a **Client** (📋) agent, providing the initial task brief (as shown in Figure 6.1) and subsequent feedback representing the audience's perspective; 2) "Feedback" agents, offering insights based on their specific roles. In this study, we explore the integration of three feedback agents alongside the Writer and Client agents. Details of the prompts used to initialize and generate responses for each agent are provided in Appendix E.2.

Analyst (📈) agent tasked with extracting and analyzing historical financial data, the Analyst agent leverages the AlphaVantage API to gather earnings performance data for the previous quarter. This information, presented alongside the preceding conversation context to the LLM, allows the agent to draw additional insights about the company's current financial performance through comparison to previous performance, enabling the formulation of pertinent feedback.[1]

Psychologist () agent analyzes external data, specifically phonetic statistics from earnings call (EC) audio recordings, to offer additional insights on the level of confidence vocally expressed by management (e.g., CEO, CFO, etc.). Following Qin and Yang (2019) who show that such statistics are useful in the prediction of financial risk, PRAAT features are derived from the utterances of the management team during the EC, enriching the feedback provided to the LLM and the discourse on management's attitudes towards present or future financial performance.[2]

Editor () agent ensures the generated report is suitable for the intended audience (in terms of content, style, and structure) and for checking that important information is maintained through revisions.

Stage 2. Initial Drafting Upon receiving the task brief from a Client agent, the Writer agent generates the initial draft of the report.

Stage 3. Feedback and Revision In an iterative process, each agent furnishes feedback aimed at enhancing the report, concentrating on elements relevant to their roles. Following each feedback round, the Writer updates the report. This cycle concludes when the preset maximum of N iterations is reached or upon the Client agent's determination that the report is complete.[3]

[1] https://www.alphavantage.co/

[2] Audio recordings are collected from Seeking Alpha (https://seekingalpha.com/) and force-aligned with transcripts using the Aenaes library (https://github.com/readbeyond/aeneas).

[3] We set the value of N to 10 for all experiments, but find that it rarely reaches this threshold without being stopped by the Client.

6.3 Generated Report Characterization

To gain a full understanding of the style, content, and utility of generated reports, we conduct several in-depth experiments and analyses. Adopting multiple configurations of our multi-agent framework, we generate reports for a sample of 60 EC transcripts used in previous work (Mukherjee et al., 2022), basing our sample size on previous work for multi-agent text generation tasks (Chan et al., 2023; Wang et al., 2024).

6.3.1 Generated vs. Human-Authored Reports

Given that our analytical reports are generated in a zero-shot setting, they is no guarantee that they will closely resemble those written by human experts. Therefore, to answer **RQ1**, we start by analyzing the similarities and differences between generated reports and those produced by human experts. For human-authored reports, we use a sample of 26 equality research reports from the Bloomberg Terminal that are authored by professional analysts at J.P. Morgan, to which we were granted restricted access.

Content To identify the key topics of discussion within each report type, we employ the aspect-extraction method outlined by Tulkens and van Cranenburgh (2020), we use SpaCy (Honnibal and Montani, 2017) to extract the 250 most frequent nouns in the earnings call transcripts of ECTSumm and analyze their presence in both generated and reference (JP Morgan) reports. Figure 6.2 presents the results of the content analysis. Using the Figure we can see that, for example, the topic of "growth" appears in 95% of analytical reports generated with all agents (All Agents), but only in 65% of the human-authored reports of JP Morgan. To enhance clarity, we've divided the figure into three segments based on the similarity of aspect occurrences between the generated reports (with all agents) and the reference reports. The top segment represents aspects that occur more frequently in the generated reports, the middle segment includes aspects with similar frequencies, and the bottom segment comprises aspects that occur more frequently in the reference reports. The Figure shows that, although generated and human-authored reports can be seen to discuss

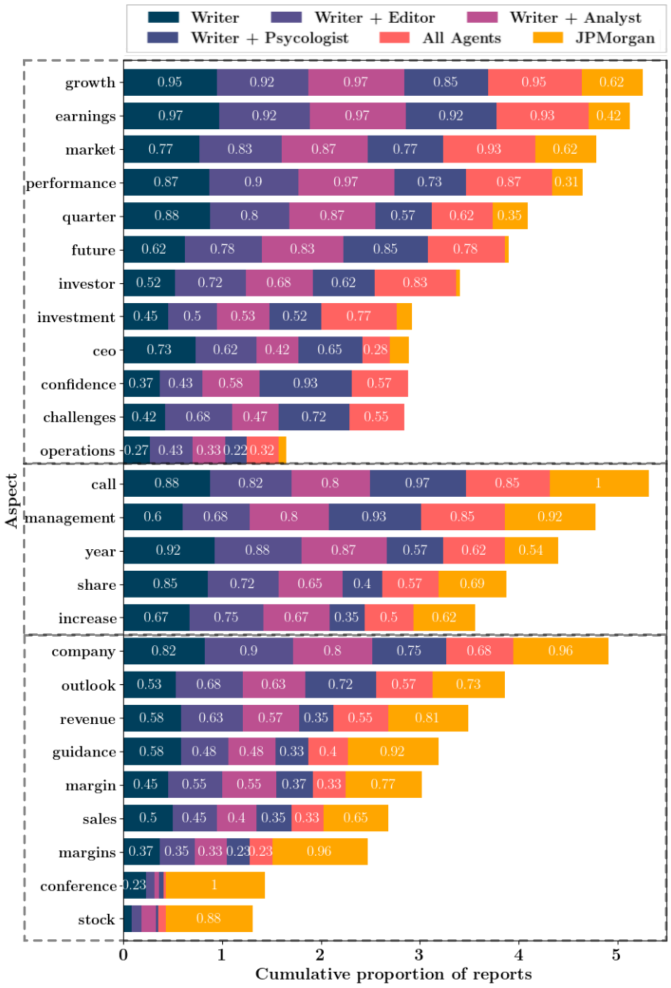


Figure 6.2: *A visualization of the proportion of aspect occurrences for different report types, including at least the 10 most common aspects for each.*

aspects like "share", "management", and "increase" at a similar rate, there exists a significant divergence in content emphasis. For instance, human-authored reports place a greater focus on financial statistics, with aspects like "margin(s)", "revenue", "sales", and "stock" occurring more frequently. In contrast, generated reports can be seen to emphasize aspects such as "performance", "future", "earnings", and "market" which, whilst different from references, remain indicative of analytical discussion. Furthermore, generated reports introduce several aspects that very rarely occur in reference reports (or do not occur at all), including "investor", 'investment", and "confidence" implying they address their intended audience more explicitly.

Agents	# Sents	FKGL	CLI	ARI	Abst
✍️	24.35	12.88	16.42	16.87	41.74
✍️🔍	22.90	13.67	17.55	17.83	48.03
✍️📈	21.43	13.44	17.32	17.24	49.46
✍️🎧	20.03	15.71	19.03	20.26	57.95
✍️📈🔍	19.65	14.76	18.33	19.10	53.40
✍️🎧🔍	19.68	15.69	19.18	20.11	56.87
✍️🎧📈🔍	18.58	15.11	18.98	19.46	56.72
References	19.25	7.26	8.54	8.85	47.14

Table 6.1: *Readability with different feedback agent configurations.*

Style Table 6.1 presents the statistical results of several metrics selected to measure the stylistic properties of reports. Here, we employ readability metrics Flesch-Kincaid Grade Level (FKGL) Kincaid et al. (1975), Coleman–Liau index (CLI) Coleman and Liau (1975), and Automated Readability Index (ARI) Senter and Smith (1967), to assess text complexity Goldsack et al. (2023c). These widely used metrics are calculated based on such factors as the number of characters (ARI and CLI), syllables (FKGL), words (all), and sentences (all) present in a given text. Higher scores indicate greater document complexity. To provide insights on document length and content novelty, we also calculate the number of sentences using NLTK (Bird et al., 2009) and the abstractiveness (% of unigrams used that do not occur in the source transcript).

Here, we can see that, although both generated and human-authored reports are generally of a similar length and level of abstractiveness, there is a significant divergence in the scores of readability metrics. Expert-written reports obtain readability scores ranging from 7-10, deemed suitable for a majority of marketing materials, and indicative of shorter statement-like sentences (i.e., containing fewer syllables or characters). Contrastingly, readability scores of generated reports span a range of 12-20, indicative of longer more complex sentences and aligning with materials intended for a highly skilled readership, such as academic publications.

6.3.2 Impact of Feedback Agents

Again utilizing Figure 6.2 and Table 6.1, we can begin to assess the impact each agent has on the output report. Firstly, the incorporation of both the Editor and the Analyst agents is shown to have a similar effect on content, increasing the rate at which aspects like "outlook", "market", "management", and "future" are discussed when compared to the Writer agent alone. This implies that both the additional financial statistics introduced by the Analyst and the critical feedback of the Editor induce a broader and more speculative analysis of company performance, causing the content to transition from focusing primarily on reporting the facts and figures from the transcript to a more speculative and potentially more insightful discussion. Table 6.1 supports this, showing an increase in abstractiveness (Abst) upon the introduction of each agent, suggesting that report content becomes less based on the source transcript and more based on external data and agent discussion. Additionally, readability metrics (FKGL, CLI, and ARI) can also be seen to increase, denoting the use of longer and more complex words/sentences.

For the Psychologist agent, Figure 6.2, shows a significant decrease in the reporting of aspects relating to financial performance figures ("year", "share", "quarter", "increase", "revenue", "margins") in favor of aspects relating to the attitude and confidence of management ("management", "confidence", "future", "outlook"), areas that this agent was designed to focus on. In addition to the change of focus, Table 6.1 suggests a significant change in the style of reporting. More than any other agent, the Psychologist causes readability and abstractiveness metrics to rise, demonstrating its ability to influence the generated text through the introduction of novel content.

6.3.3 Insightfulness of Generated Reports

To answer **RQ2**, and determine how effective a multi-agent approach is at providing insights that are potentially useful to an investor, we conduct an in-depth human evaluation utilizing domain experts. We employ three evaluators, all of whom are pursuing postgraduate studies in Finance, and ask them to assess reports generated by the Writer (✍️) alone with those produced using all agents (✍️🎧📈🔍) for 32

Report characteristic	Description
Financial takeaways	The key financial details from the meeting (i.e., numerical statistics relating to company performance for the quarter).
Financial context	Any additional information (e.g., financial details from previous quarters) that helps to contextualize the current financial performance.
Management attitudes	Information on how management (e.g., CEO, CFO, etc..) feels about the company's financial performance.
Management expectation	Details about how the company is expected to perform in the future/next quarter.
Possible future events	Details surrounding any noteworthy events/scenarios that are likely to occur in the future.

Table 6.2: *Human evaluation assessment criteria descriptions.*

randomly-selected earnings call instances, allowing us to see the impact of our designed feedback agents. Specifically, the evaluators' task is broken down into the assessment of the following key report characteristics: 1) Financial takeaways, 2) Financial context, 3) Management attitudes, 4) Management expectations, 5) Possible future events. For each characteristic, evaluators assign one of the following labels, identifying both the presence and insightfulness of relevant information: 1) Reported and insightful, 2) Reported and reasonable, 3) Reported but not useful, and 4) Not reported. Our evaluation characteristics and labels were developed in collaboration with professional analysts, and are crucial in capturing report value to potential investors. Full descriptions of our characteristics and labels are provided in Tables 6.2 and E.7 (Appendix E.3), respectively.[4]

Evaluation results The distribution of evaluator ratings for each system is depicted in Figure 6.3. Firstly, it can be observed that, for both systems, the percentage of positive labels (i.e., "reported and insightful" and "reported and useful") significantly outweighs the more negative labels for all characteristics. Additionally, there are almost no cases of the "reported but not useful" label for either system, attesting to the quality of the information that is included in the report. Of all the characteristics that are discussed in the reports generated by both systems, Financial takeaways are the most frequently included, whereas Management attitudes are the most frequently absent.

[4] A discussion of inter-annotator agreement is also provided in Appendix E.3.

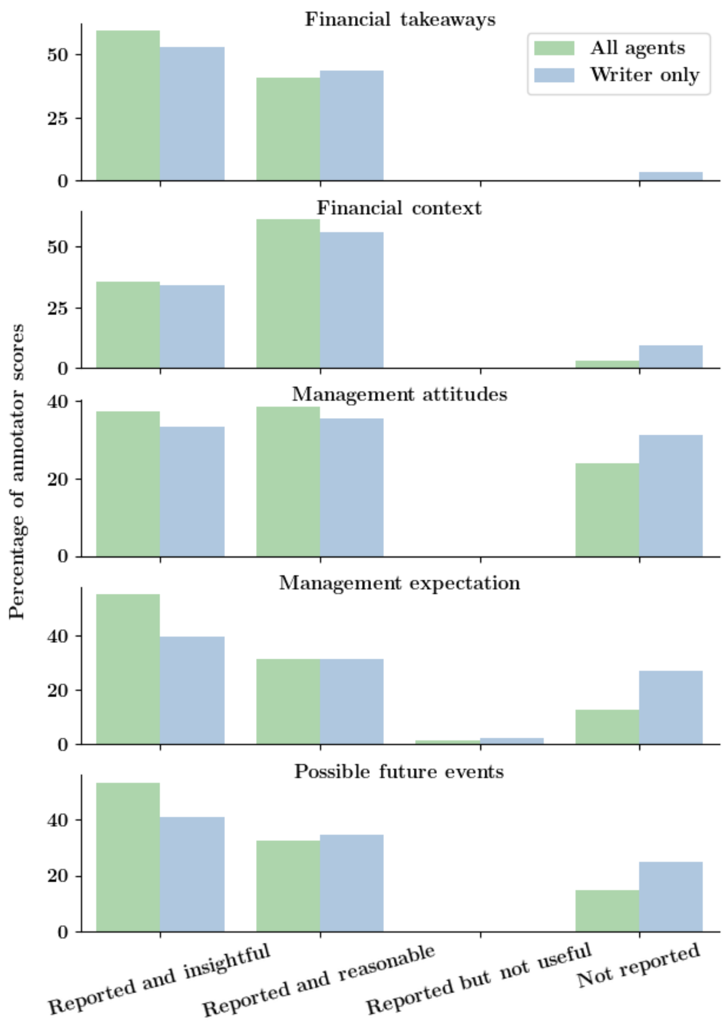


Figure 6.3: *Characteristic-based human evaluation results.*

Results also indicate that the inclusion of all agents leads to a decrease in the percentage of characteristics labeled as "not reported" and an increase in the percentage of characteristics deemed "reported and insightful" in all instances. Particularly notable improvements are observed for the management expectations and possible future events characteristics, illustrating how the additional feedback agents significantly improve the speculation and anticipation of future performance in generated reports, factors that are particularly relevant to investment decision-making.

Case study To gain a better understanding of insights generated by the multi-agent framework, we perform a closer examination of reports and annotator classifications,

A. Financial takeaways

Writer only - Motorola Solutions showcased a robust financial performance in the third quarter of 2021. The company reported a revenue increase of 13% and a surge in earnings per share (EPS) by 21%, signaling strong demand across businesses. Operating margins also expanded by 150 basis points. A record backlog of $11.4 billion was achieved, marking a 7% year-over-year increase.**1/3**

All agents - Motorola Solutions persists in its upward financial trajectory, as illustrated by pivotal metrics: - *Revenues*: A robust 13% increase substantiates the company's firm market footing and demand for its solutions. - *Earnings Per Share (EPS)*: The 21% rise in EPS reflects burgeoning profitability, a reassuring signal for shareholders. - *Operating Margins*: The expansion by 150 basis points testifies to enhanced operational rigor and disciplined cost governance. - *Record Backlog*: The $11.4 billion figure – up 7% year-over-year – fortifies visibility into future revenue streams.**3/3**

b. Management attitudes

Writer only - As the economy continues to transition, Dynex remains cautiously optimistic about capitalizing on emerging opportunities while maintaining its established principles of responsible money management and transparency.**1/3**

All agents - The managers' communication styles, marked by confidence and clarity, supplemented the strong financial results by providing an indication of management's collective poise and tactful approach to stewardship, which might positively influence investor sentiment.**3/3**

c. Management expectation

Writer only - In regards to the outlook for Q4 of fiscal '22, RPM projected consolidated sales growth in the low teens versus the prior year, despite anticipated supply chain challenges and currency headwinds due to the strengthening U.S. dollar. Notwithstanding these obstacles, it was suggested that RPM could achieve record EBIT on a consolidated basis in Q4.**1/3**

All agents - *Growth Prospects*: RPM foresees low teens growth in consolidated sales and adjusted EBIT for Q4. This would continue the trend of sustained growth even when set against a challenging economic backdrop, including impacts from the Ukrainian conflict.**3/3**

Figure 6.4: *A case study comparing how the introduction of feedback agents changes the discussion of key aspects. Colors and superscript are used to denote the number of evaluators who judged the sentence as "reported and insightful (e.g.,* **3/3** *= 3 out of 3 evaluators).*

presenting examples of our findings in the form of a case study in Figure 6.4. The given examples demonstrate how additional agent feedback can make the reporting improve the reporting of important information. Looking at each of the provided instances, example A shows how the same financial takeaways are presented alongside additional insights for investors (a likely influence of the Editor); example B illustrates how novel content relating to the confidence of management is introduced by the Psychologist's audio analysis; and example C demonstrates how past financial from the Analyst can contextualize relevant facts.

6.4 Evaluating Analytical Reports with LLMs

In this section, we address **RQ3** and attempt to establish how future works may address the challenge of evaluating generated analytical reports. Specifically, we first describe the challenges of a conventional reference-based evaluation (§6.4.1),

Characteristic	**GPT-4**			**Gemini-pro**			**Mistral-medium**		
	γ	ρ	τ	γ	ρ	τ	γ	ρ	τ
Financial Takeaways	0.375	0.160	0.412	0.156	0.018	0.014	0.139	0.205	0.192
Financial Context	0.597	0.455	0.397	0.341	0.330	0.292	0.758	0.437	0.397
Management Attitudes	0.570	0.524	0.463	0.248	0.301	0.266	0.463	0.558	0.492
Management Expectation	0.529	0.511	0.441	0.643	0.598	0.521	0.670	0.661	0.581
Future Events	0.472	0.379	0.327	0.179	0.194	0.167	0.422	0.382	0.330
Average	0.509	0.405	0.408	0.313	0.288	0.252	0.490	0.449	0.398

Table 6.3: *Correlation statistics of LLMs vs. human evaluators (averaged) for each report characteristic.*

before exploring the potential of LLMs for reference-free evaluation (§6.4.2).

6.4.1 Challenges of Reference-based Evaluation

Significantly, the more conventional method of assessing generated reports against references using automatic metrics faces several inherent limitations in the context of analytical reports. Firstly, the scarcity of available data samples is compounded by the fact that these reports, being internally generated by corporate entities, pose challenges in establishing public benchmarks. The small dataset of 26 reports we've gathered is subject to strict redistribution restrictions, precluding us from making them publicly accessible for benchmark creation. Even if such access were feasible, these reports typically adhere to in-house guidelines and practices, contributing to disparities in content and style between human and machine-generated reports such as those outlined in §6.3.1. Accordingly, any novel insights provided during generation are unlikely to be adequately captured or rewarded through reference-based comparison. Furthermore, these instances are based on Earnings Calls (ECs) considerably older (2012-2016) than others utilized in this study (2019-2022), which played a role in our granted access. This raises questions about their usefulness as a potential point of comparison, considering a possible evolution of financial reporting practices.

6.4.2 Reference-free Evaluation with LLMs

Given the limitations described in §6.4.1, we explore the use of LLMs for the reference-free evaluation of generated outputs, a direction that has proved promising in previous studies (Liu et al., 2023; Luo et al., 2023; Chan et al., 2023). Utilizing their respective APIs, we experiment with GPT-4 (OpenAI, 2023), Gemini-pro (Gemini, 2023), and Mistral-medium (Jiang et al., 2024), instructing each to embody a financial expert and reenact evaluations performed by experts. We assess the performance of LLM evaluators in two popular human evaluation settings: 1) a characteristic-based setting and 2) a preference-based (ranking) setting. The prompts used for each setting are provided in Appendix E.2, Table E.3.

Characteristic-based Evaluation For a characteristic-based evaluation, LLM evaluators are tasked with performing the evaluation described in §6.3.3. Here, we adhere to established evaluation methodologies from previous studies (Zhong et al., 2022; Chan et al., 2023) and employ Pearson (γ), Spearman (ρ), and Kendall (τ) correlation coefficients between LLM and human evaluators.[5] Table 6.3 presents our findings. Here, we can see that GPT-4 and Mistral obtain a good level of correlation with human experts, whereas Gemini-pro performs slightly worse in terms of average correlation scores.

Looking closer at individual characteristics, all models can be seen to achieve a strong level of correlation (> 0.5) for at least one of the listed characteristics and GPT-4 maintains at least a moderate level (> 0.3) of correlation across all characteristics. Contrastingly, Gemini and Mistral achieve a broader range of correlation scores, with particularly strong scores for some aspects (e.g., Management Expectations), but weaker scores for others (e.g., Financial Takeaways).

Overall, these results indicate that LLMs have significant potential in the evaluation of analytical reports when assessing fine-grained characteristics, but that performance is likely to differ depending on the LLM. Although GPT-4 can be considered the best all-round evaluator, the fact that different LLMs achieve stronger correlations for specific characteristics is something that future works should consider when designing their evaluation.

[5]To calculate correlation, we convert our labels into numeric scores ranging from 1-4, with 4 being the most positive classification (reported and insightful) and 1 being the most negative (not reported).

Report	GPT-4		Gemini-pro		Mistral-med	
	#1	#2	#1	#2	#1	#2
Generated	100.0	70.83	87.5	100.0	91.67	16.67
Reference	0	29.17	12.5	0.0	8.33	83.33

Table 6.4: *The % of preference annotations given by each LLM evaluator for generated (using the Writer, Analyst, and Editor) and reference reports. #1 = generated report given first in prompt, #2 = reference report given first in prompt.*

Preference-based Evaluation In addition to the characteristic-based evaluation, we perform a preference-based evaluation utilizing the professional analytics reports we collect from JP Morgan (described in §6.3.1). Specifically, evaluators are required to compare the reference report and the report generated by our system with all possible agents (✍️📈🔍)[6] Here, we integrate argument quality evaluation principles Gretz et al. (2020), which hinge on *whether evaluators would recommend a friend to use that argument as is in a speech supporting/contesting the topic, regardless of personal opinion.* In our case, evaluators indicate which report they would recommend to someone who would be making an investment decision based on the information released in the EC.

After conducting this evaluation using the same three experts as in our characteristic-based evaluation, we find that there remains a large preference for human-authored reports over generated reports, with human annotators preferring the reference reports of JP Morgan 83.33% of the time (on average). To gain further insights into human preferences, we conducted in-depth interviews with human annotators, which revealed the general preference for reference reports was attributed to their detailed, forward-looking evaluations, comprehensive risk assessments, and specific financial performance forecasts. These reports adeptly juxtapose company guidance against market expectations and consider the implications of company policies and regional market dynamics. This feedback serves as a cornerstone for future research on generating analytical reports for ECs.

Table 6.4 presents the results of LLM evaluators for this preference-based evaluation. Given that previous work (Wang et al., 2023) has identified a tendency of

[6] Note that, due to these ECs being from earlier dates, their audio recordings are unavailable and we were unable to include our Psychologist agent in report generation process for these instances.

LLMs to favor the first displayed candidate in a ranking scenario, we perform this experiment with both possible candidate orderings. Interestingly, we find that only Mistral exhibits the strong positional bias described by Wang et al. (2023). For GPT-4 and Gemini-pro, the stronger bias is shown toward generated outputs, with both models overwhelmingly favoring them regardless of candidate orderings, starkly contrasting with human experts. Furthermore, in all cases, the models are largely inconsistent across both runs. These factors highlight serious limitations in using LLMs to assess the overarching quality of analytical reports for ECs, particularly in ranking scenarios involving both human- and machine-generated outputs.

6.5 Related Work

6.5.1 LLMs as Task-solving Agents

The deployment of multiple LLMs to collaboratively work on a task has recently emerged as a trend in NLP research. While one branch of this research has typically sought to answer if adopting such an approach can improve collective reasoning Du et al. (2023); Liang et al. (2023) another has explored how the unique opportunities afforded by this approach might allow us to attempt yet more complex tasks (Qian et al., 2023; Wu et al., 2023; Chan et al., 2023; Wang et al., 2024). Of these, our work is in closer alignment with the latter. Flexible and generic frameworks such as Autogen (Wu et al., 2023), utilized in this work, have recently emerged, enabling the development of agent-based approaches that are easily customizable and utilize external tools. Related studies have taken the initial steps in exploring challenges such as assessing generated text by employing multiple agents (Chan et al., 2023), establishing the importance of diverse agent roles/personas. Similarly, Wang et al. (2024) leverages multiple language model personas to enhance performance in tasks demanding knowledge and reasoning, such as creative writing based on trivia and solving logic puzzles. In contrast to these prior efforts, our focus centers on a more specialized task, necessitating the development of agents with domain expertise and benefiting from the incorporation of external data.

6.5.2 Earnings Call Processing

As mentioned in §6.1, ECs have proved a popular topic of study for previous work, due largely to their significance in investor decision-making. Of these works, the most related to ours is that of Mukherjee et al. (2022) which introduces and benchmarks ECTSumm, a dataset for the generation of journalistic EC reports. However, in contrast to this work, we focus on generating analytical reports using a multi-agent framework, a notably more challenging task.[7]

Another more well-explored branch of EC processing focuses on the utilization of transcripts as the source documents for predictive NLP tasks, including volatility prediction (Sawhney et al., 2021; Niu et al., 2023), analyst decision prediction (Keith and Stent, 2019), financial risk prediction (Qin and Yang, 2019), and earnings surprise prediction (Koval et al., 2023). In contrast to these works, we tackle a complex generation task, although we inspiration from their findings. For instance, the inclusion and design of our Psychologist agent is influenced by the work of Qin and Yang (2019), who demonstrate the efficacy of features based on EC audio in the prediction of financial risk.

6.6 Conclusion

This study explores the novel task of generating analytical reports for ECs. Following an investigation of the key distinctions between analytical reports and previously studied journalistic reports, we address the generation of analytical reports using an LLM-based multi-agent framework. We perform a thorough characterization of generated reports, revealing key divergences with human-authored reports while also highlighting the ability of agents to introduce useful insights. Finally, we address the open challenge of evaluation using LLMs, establishing the utility of different setups, and laying the groundwork for future research. Here, our findings illustrate a detrimental tendency for LLMs to favor generated over human-authored reports, but reveal that LLMs largely achieve good alignment with human experts when it comes

[7] We provide an analysis of the differences between journalistic and analytical reports in detail in Section E.1 of the Appendix.

to evaluating fine-grained criteria. While our framework aims to generate insightful analytical reports, there remains a significant opportunity to explore generative techniques that can produce novel insights. Future research could greatly benefit from incorporating real-time financial data, news, and market trends for more useful analyses.

Limitations

Given the flexibility of multi-agent frameworks, there are undoubtedly many alternative options in terms of agent design and interaction that could prove beneficial and are worthy of investigation. However, as this work represents a first exploration of this task, we primarily concentrate our research efforts on establishing knowledge and practices that will benefit future work, for example, in the form of our generated report characterization (§6.3), and in the investigation of LLM-based evaluation (§6.4).

As discussed in the paper, another limitation in exploring the task of generating analytical reports in open research is the lack of suitable reference reports, an issue that naturally arises from their typically corporate origins. We attempt to address this issue by exploring the use of Large Language Models for reference-less human-style evaluation and hope that these findings will have a positive impact on future research on this task.

6.7 Updated Limitations

Additional limitations of this work are outlined below for the purpose of this thesis:

- Figure 6.3 presents the results of the characteristic-based human evaluation of generated reports, both with and without additional feedback agents included in our framework. Although this Figure demonstrates that additional feedback agents improve performance, an additional comparison with the results of this analysis for human-written reports would be valuable in future work. As such a comparison was not possible in this work due to a lack of human-written report for the source articles used in this evaluation, a separate preference-based evaluation with human-written report was conducted using a separate evaluation set (discussed in §6.4.2).

- It should be further highlighted that alignment of LLM and average human judgments when assessing "financial takeaways" (Table 6.3) is significantly weaker of other metrics. This is a result of relatively greater disagreement between human annotators for this criterion, which is highlighted in Appendix E. Future work on this topic would benefit from placing additional focus on the evaluation this criterion, given its importance to the overall quality of generated reports.

Chapter 7

Conclusion

This thesis has focused on the application of summarisation approaches to technical and scientific texts. In doing this, an emphasis was placed on developing approaches that can broaden access to wider audiences, improving understanding of technical texts through diverse analyses, and exploring how relevant external knowledge can be used to improve model performance.

7.1 Thesis Summary

This section provides a summary of the inner chapters of this thesis, and how the research questions associated with each corresponding publication are addressed.

7.1.1 Publication I: *Domain-driven and Discourse-guided Scientific Summarisation*

Presented in Chapter 2, this publication addresses the following research questions:

Does scientific domain of an article has a strong influence over the rhetorical structure of its abstract?

By performing a detailed rhetorical-role annotation of abstracts drawn from multiple arXiv categories (e.g., physics, computer science, mathematics), this work reveals significant differences in how authors in each field organize their abstracts. For instance, computer-science abstracts allocate a larger proportion of sentences to "method" and "results," whereas mathematics abstracts emphasize "background" and "motivation." This empirical mapping of domain to abstract structure demonstrates definitively that the scientific domain exerts a strong guiding influence on abstract structure.

Can scientific domain be leveraged in the context of abstract generation to improve both the content and structure of the generated output?

Building on this insight, the paper introduces DodoRank, which constructs for each domain a lightweight "rhetorical content model" (i.e., a template of sentence-level functions and their typical ordering). During summarisation, DodoRank fills in these structural blueprints by selecting the sentences with the greatest centrality from the source that correspond to each rhetorical role. By anchoring both what content to include and where to place it, DodoRank is able to produce abstracts whose structure more closely matches human-written examples and whose content better covers each rhetorical function. Empirical results on the multi-domain arXiv corpus show that this discourse-aware, domain-driven strategy yields higher ROUGE scores and more coherent, well-balanced abstracts than domain-agnostic baselines.

7.1.2 Publication II: *Making Science Simple: Corpora for the Lay Summarisation of Scientific Literature*

Presented in Chapter 3, this publication introduced, analysed, and benchmarked two novel datasets for Lay Summarisation in the Biomedical domain, addressing the following research questions:

How do expert-authored lay summaries differ from technical abstracts in terms of readability, rhetorical structure, vocabulary overlap, and abstractiveness?

Differences between expert-authored lay summaries and technical abstracts were quantified through a multi-faceted analysis. Readability metrics (FKGL, CLI, DCRS, WordRank) consistently showed lay summaries to be easier to read than abstracts—and eLife digests to be simpler still (e.g., eLife FKGL 10.92 vs. abstract 15.57). Rhetorical-role classification revealed lay summaries devote roughly 55–58% of sentences to background (vs. 35–41% in abstracts), while trimming emphasis on results and methods. Vocabulary-overlap analysis indicated that most content words in abstracts (nouns, proper nouns, numbers) are not carried into lay summaries, evidencing significant lexical simplification. Finally, n-gram novelty confirmed lay summaries are markedly more abstractive than abstracts—particularly the longer, editor-written eLife digests.

How effective are current summarisation methods at producing clear, accurate lay summaries of biomedical research?

The effectiveness of current summarisation methods was assessed via automatic and human evaluation. Transformer-based BART models achieved ROUGE-1 scores of ∼42.4 on PLOS and ∼46.6 on eLife—surpassing most extractive and heuristic baselines on eLife but only matching them on PLOS—while generating outputs with FKGL around 14 and DCRS around 12. However, expert judgments (1–5 scale) rated BART's outputs only moderately for comprehensiveness (3.7 PLOS; 3.1 eLife), layness (3.0 both), and factuality (3.0 both), highlighting persistent challenges in accuracy and clarity for non-expert audiences.

7.1.3 Publication III: *Enhancing Biomedical Lay Summarisation with External Knowledge Graphs*

Chapter 4 presents this publication, which addresses the following research question:

Can current approaches for lay summarisation be improved by augmenting source articles with external domain knowledge in the form of knowledge graphs?

The eLife dataset is augmented with paper-specific knowledge graphs containing background information on relevant Biomedical concepts, and the performance of three different methods for including these graphs within transformer encoder-decoder models for Biomedical Lay Summarisation is compared. All three KG-enhanced variants produce summaries that human judges find significantly more readable and better at explaining technical concepts, with additional case studies indicating that the use of external knowledge improves the ability of models to generate readable explanations of technical concepts, thus improving the overall quality of lay summaries.

7.1.4 Publication IV: *Leveraging Large Language Models for Zero-shot Lay Summarisation in Biomedicine and Beyond*

This publication, included in Chapter 5, explores the use of Large Language Models (LLMs) for the task of zero-shot Lay Summarisation in the Biomedical and Natural Language Processing domains. It addresses the following research question:

Can zero-shot lay summarisation with LLMs be improved by introducing a prompting framework that emulates real-world practices?

Using multiple open-source LLMs of various sizes and architectures, the performance of LLMs using a baseline prompt and a novel two-stage prompting technique, proposed based on the real-world practices of the eLife journal were compared. The results of this comparison indicate that humans increasingly prefer the lay summaries of the two-stage framework as LLM size increases, and cast light on the inability of standard summarisation evaluation metrics to capture such preferences.

7.1.5 Publication V: *From Facts to Insights: A Study on the Generation and Evaluation of Analytical Reports for Deciphering Earnings Calls*

Chapter 6 includes this final publication, addressing the novel task of generating analytical reports based on the transcripts of financial Earnings Call meetings. It covers the following research questions:

How do generated analytical reports differ from human-authored analytical reports?

In terms of content, generated reports tend to emphasize forward-looking and interpretive aspects - e.g., "performance", "future", "earnings", "market", and investor-focused terms like "confidence" and "investment"—whereas human-authored reports concentrate more heavily on concrete financial statistics such as "margins", "revenue", "sales", and "stock". Secondly, in terms of style, generated reports exhibit substantially higher complexity: readability scores (Flesch–Kincaid Grade Level, Coleman–Liau, ARI) range roughly 12–20—indicative of long, dense sentences—while expert-written reports score around 7–10, aligning with concise, statement-like prose. Generated texts also show greater abstractiveness (more novel wording) and a similar overall length.

Can a multi-agent approach be used to generate more insightful analytical reports?

When comparing reports by the Writer-only system against those produced by the full multi-agent setup (Writer + Client + Analyst + Psychologist + Editor), expert annotators rated the multi-agent reports as "reported and insightful" more often and marked "not reported" far less frequently across all five key characteristics (financial takeaways, financial context, management attitudes, expectations, and future events). The most pronounced improvements occurred for management expectations and possible future events, showing that specialized agents (Analyst and Psychologist) effectively broaden and deepen speculative analysis beyond mere transcript facts.

How effective are LLM-based evaluation methods in assessing the quality of analytical reports?

Using GPT-4, Gemini-pro, and Mistral-medium as "expert" evaluators on the same five dimensions:

- GPT-4 achieves moderate correlations with human judgments (Pearson $\gamma \approx 0.51$, Spearman $\rho \approx 0.41$), performing best on financial context ($\gamma = 0.597$) and worst on financial takeaways ($\gamma = 0.375$).

- Mistral-medium yields similar overall strength ($\gamma \approx 0.49$, $\rho \approx 0.45$), excelling on context ($\gamma = 0.758$) but lagging on takeaways ($\gamma = 0.139$).

- Gemini-pro underperforms across the board.

These results suggest LLMs can reliably assess fine-grained characteristics - especially context and expectations - but their accuracy varies markedly by model and by evaluation criterion.

7.2 Discussion of Research Objectives

In this section, each of the research objectives established in Chapter 1 are discussed:

(1) Developing and enabling the development of novel summarisation approaches that allow the contents of technical texts to be accessed by a wider audience. This objective is addressed directly within Publications II, III, and IV, all of which focus on developing and testing different approaches for the task of Lay Summarisation. Within these Publications, a broad range of techniques were utilised, primarily surrounding contemporary transformer-based summarisation models, including long-context models, knowledge graph-augmented generation, and zero-shot inference with LLMs. Finally, in Publication V, the highly related task of analytical report generation is explored, with the objective being to generate speculative reports that are more accessible and useful to a potential investor than the source text on which they are based.

Moreover, in Publication II, future work on this task is enabled by the introduction of two novel high-quality datasets. Subsequently, these datasets have been widely adopted, forming the basis of the two edition of the BioLaySumm shared task (Goldsack et al., 2023a, 2024a). As of January 2025, this paper has received 62 citations (according to Google Scholar) and the datasets have been downloaded over 11,924 times from the HuggingFace platform where they are hosted.[1]

(2) Improving our understanding of technical texts and their summaries through in-depth, varied linguistic analyses. Numerous analyses of technical texts and their summaries were conducted across the published work presented in this thesis. In Publication I, the rhetorical structures of research article abstracts were analysed, illustrating how abstracts differ from domain to domain. In Publication II, an extensive characterisation of the summaries introduced within the proposed Lay Summarisation datasets was performed, focusing on their readability, rhetorical structure, use of content words, and abstractiveness. In Publication III, a deeper understanding of the Biomedical concepts covered within the eLife dataset was gained through the introduction and analysis of paper-specific knowledge graphs. Finally, in Publication V, the differences in both the readability and terminology of journalistic and analytical reports was established for financial Earnings Calls (given in Appendix E). Additionally, the same differences between generated and human-authored analytical reports was subsequently analysed.

(3) Exploring how the use of relevant external knowledge can be used to improve the summarisation of technical texts. External knowledge, created or retrieved from a source other than the input text, has formed the backbone of many of the approaches used within this thesis. In Publication I, a sample of research articles was used to derive sentence-level rhetorical content models of abstracts from different scientific domains, before using these content models to inform the generation of new abstracts. In Publication III, the Unified Medical Language System (UMLS) was leveraged to create external knowledge graphs that illustrate the complex Biomedical concepts that were mentioned within an article and the relationships between them, before testing several summarisation approaches which utilise these knowledge graphs. In Publication IV, a framework based upon knowledge of the real-world practices of eLife was proposed, utilising resources

[1] https://huggingface.co/datasets/tomasg25/scientific_lay_summarisation

provided by eLife staff members for prompt design. Finally, in Publication V, the Psychologist and Analyst roles within the proposed multi-agent framework were specifically designed to leverage two important sources of external knowledge - the Earnings Call audio recording and API-retrieved financial statistics, respectively.

7.3 Future Directions

The work presented in this thesis raises several compelling avenues for future research:

- How can contemporary LLMs and LLM frameworks be further leveraged to improve the generation of lay summaries for technical text? Although the strong performance of zero-shot LLMs is demonstrated in Publication IV, avenues related to LLM fine-tuning and the inclusion of external knowledge within a Retrieval-Augmented Generation (RAG) framework remain unexplored, but are likely to provide performance benefits.

- How can the evaluation of generations based on technical texts be improved? Publication V highlights that, in particular settings and for particular criteria, LLMs can prove an effective proxy for humans in the evaluation of generated analytical reports. Based on these findings, future work is likely to benefit from a rigourous exploration of the use of LLMs for evaluation when generating text based on technical documents, particularly for novel tasks that lack reference outputs.

- What other generation tasks based on technical texts are now open to use with contemporary approaches? In this thesis, the task of summarisation is largely focused on partly due to the availability of parallel training and evaluation corpora. However, the recent increase in the zero-shot abilities of LLMs has somewhat alleviated this constraint within the NLG community, and opened up many more exciting directions that can be explored. Publication V exemplifies this, where the novel task of analytical report generation is addressed, but many other interesting tasks and avenues remain unexplored within the context of technical texts, many of which have the potential to aid the comprehension of expert and non-expert readers alike.

Bibliography

Toufique Ahmed and Premkumar Devanbu. 2023. Few-shot training llms for project-specific code-summarization. In *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering*, ASE '22, New York, NY, USA. Association for Computing Machinery.

Mathieu Albert, Elise Paradis, and Ayelet Kuper. 2015. Interdisciplinary promises versus practices in medicine: The decoupled experiences of social sciences and humanities scholars. *Social Science Medicine*, 126:17–25.

Mathieu Albert, Paula Rowland, Farah Friesen, and Suzanne Laberge. 2021. Barriers to cross-disciplinary knowledge flow: The case of medical education research. pages 2212–277X. Perspectives on Medical Education.

Fernando Alva-Manchego, Louis Martin, Carolina Scarton, and Lucia Specia. 2019. EASSE: Easier automatic sentence simplification evaluation. 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): System Demonstrations*, pages 49–54, Hong Kong, China. Association for Computational Linguistics.

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, Rodney Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler Murray, Hsu-Han Ooi, Matthew Peters, Joanna Power, Sam Skjonsberg, Lucy Lu Wang, Chris Wilhelm, Zheng Yuan, Madeleine van Zuylen, and Oren Etzioni. 2018. Construction of the literature graph in semantic scholar. In *Proceedings of the 2018 Conference of* *the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 84–91, New Orleans - Louisiana. Association for Computational Linguistics.

Alan R Aronson and François-Michel Lang. 2010. An overview of MetaMap: historical perspective and recent advances. *J. Am. Med. Inform. Assoc.*, 17(3):229–236.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. 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 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *Preprint*, arXiv:2004.05150.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, 32(Database issue):D267–70.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel Weld. 2020. TLDR: Extreme summarization of scientific documents. In *Findings of the Association for Com-*

putational Linguistics: EMNLP 2020, pages 4766–4777, Online. Association for Computational Linguistics.

Hengxing Cai, Xiaochen Cai, Junhan Chang, Sihang Li, Lin Yao, Wang Changxin, Zhifeng Gao, Hongshuai Wang, Li Yongge, Mujie Lin, Shuwen Yang, Jiankun Wang, Mingjun Xu, Jin Huang, Xi Fang, Jiaxi Zhuang, Yuqi Yin, Yaqi Li, Changhong Chen, Zheng Cheng, Zifeng Zhao, Linfeng Zhang, and Guolin Ke. 2025. SciAssess: Benchmarking LLM proficiency in scientific literature analysis. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2335–2357, Albuquerque, New Mexico. Association for Computational Linguistics.

Xiaoyan Cai, Sen Liu, Libin Yang, Yan Lu, Jintao Zhao, Dinggang Shen, and Tianming Liu. 2022. COVIDSum: A linguistically enriched SciBERT-based summarization model for COVID-19 scientific papers. *Journal of Biomedical Informatics*, 127:103999.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *Preprint*, arXiv:2308.07201.

Muthu Kumar Chandrasekaran, Guy Feigenblat, Eduard Hovy, and Anita de Waard. 2020. Overview and insights from the shared tasks at scholarly document processing 2020: CL-SciSumm, LaySumm and. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 214–224. unknown.

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. Booookscore: A systematic exploration of book-length summarization in the era of LLMs. In *The Twelfth International Conference on Learning Representations*.

Yi Cheng, Siyao Li, Bang Liu, Ruihui Zhao, Sujian Li, Chenghua Lin, and Yefeng Zheng. 2021. Guiding the growth: Difficulty-controllable question generation through step-by-step rewriting. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 5968–5978, Online.

Sumit Chopra, Michael Auli, and Alexander M. Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98, San Diego, California. Association for Computational Linguistics.

Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Dan Weld. 2019. Pretrained language models for sequential sentence classification. 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 3693–3699, Hong Kong, China. Association for Computational Linguistics.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.

Ed Collins, Isabelle Augenstein, and Sebastian Riedel. 2017. A supervised approach to extractive summarisation of scientific papers. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 195–205, Vancouver, Canada. Association for Computational Linguistics.

Danish Contractor, Yufan Guo, and Anna Korhonen. 2012. Using argumentative zones for extractive summarization of scientific articles. In *Proceedings of COLING 2012*, pages 663–678, Mumbai, India. The COLING 2012 Organizing Committee.

Peng Cui and Le Hu. 2021. Sliding selector network with dynamic memory for extractive summarization of long documents. In *Proceedings of the 2021 Conference* *of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5881–5891.

Peng Cui, Le Hu, and Yuanchao Liu. 2020. Enhancing extractive text summarization with topic-aware graph neural networks. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5360–5371.

Dina Demner-Fushman, James Mork, Sonya Shooshan, and Alan Aaronson. 2010. Wumls content views appropriate for nlp processing of the biomedical literature vs. clinical text. In *Journal of biomedical informatics, 43(4)*, page 587–594.

Franck Dernoncourt and Ji Young Lee. 2017. PubMed 200k RCT: a dataset for sequential sentence classification in medical abstracts. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 308–313, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Ashwin Devaraj, Iain Marshall, Byron Wallace, and Junyi Jessy Li. 2021. Paragraph-level simplification of medical texts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4972–4984, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. 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.

Yue Dong, Andrei Mircea, and Jackie Chi Kit Cheung. 2021. Discourse-Aware unsupervised summarization for long scientific documents. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1089–1102, Online. Association for Computational Linguistics.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *Preprint*, arXiv:2305.14325.

Shiran Dudy, Steven Bedrick, and Bonnie Webber. 2021. Refocusing on relevance: Personalization in NLG. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5190–5202, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Vijay Prakash Dwivedi, Anh Tuan Luu, Thomas Laurent, Yoshua Bengio, and Xavier Bresson. 2021. Graph neural networks with learnable structural and positional representations.

Wafaa S. El-Kassas, Cherif R. Salama, Ahmed A. Rafea, and Hoda K. Mohamed. 2021a. Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, 165:113679.

Wafaa S El-Kassas, Cherif R Salama, Ahmed A Rafea, and Hoda K Mohamed. 2021b. Automatic text summarization: A comprehensive survey.

G. Erkan and D. R. Radev. 2004a. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

Günes Erkan and Dragomir R Radev. 2004b. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479.

Liana Ermakova, Frederique Bordignon, Nicolas Turenne, and Marianne Noel. 2018. Is the Abstract a Mere Teaser? Evaluating Generosity of Article Abstracts in the Environmental Sciences. *Frontiers in Research Metrics and Analytics*, 3.

Linda Flower and John R Hayes. 1981. A cognitive process theory of writing. *College composition and communication*, 32(4):365–387.

Team Gemini. 2023. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Tomas Goldsack, Zheheng Luo, Qianqian Xie, Carolina Scarton, Matthew Shardlow, Sophia Ananiadou, and Chenghua Lin. 2023a. Overview of the biolaysumm 2023 shared task on lay summarization of biomedical research articles. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 468–477, Toronto, Canada. Association for Computational Linguistics.

Tomas Goldsack, Carolina Scarton, and Chenghua Lin. 2025. Leveraging large language models for zero-shot lay summarisation in biomedicine and beyond. *Preprint*, arXiv:2501.05224.

Tomas Goldsack, Carolina Scarton, Matthew Shardlow, and Chenghua Lin. 2024a. Overview of the BioLaySumm 2024 shared task on the lay summarization of biomedical research articles. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 122–131, Bangkok, Thailand. Association for Computational Linguistics.

Tomas Goldsack, Yang Wang, Chenghua Lin, and Chung-Chi Chen. 2024b. From facts to insights: A study on the generation and evaluation of analytical reports for deciphering earnings calls. *Preprint*, arXiv:2410.01039.

Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. 2022a. Making science simple: Corpora for the lay summarisation of scientific literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10589–10604, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. 2022b. Making science simple: Corpora for the lay summarisation of scientific literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10589–10604, Abu Dhabi. Association for Computational Linguistics.

Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. 2023b. Domain-Driven and Discourse-Guided Scientific Summarisation. In *Advances in Information Retrieval*, pages 361–376, Cham. Springer Nature Switzerland.

Tomas Goldsack, Zhihao Zhang, Chen Tang, Carolina Scarton, and Chenghua Lin. 2023c. Enhancing biomedical lay summarisation with external knowledge graphs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8016–8032.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of GPT-3. *CoRR*, abs/2209.12356.

Naomi Graetz. 1982. Teaching efl students to extract structural information from abstracts. *International Symposium on Language for Special Purposes*.

Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2020. A large-scale dataset for argument quality ranking: Construction and analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7805–7813.

Jian Guan, Yansen Wang, and Minlie Huang. 2019. Story Ending Generation with Incremental Encoding and Commonsense Knowledge. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6473–6480.

Yue Guo, Wei Qiu, Yizhong Wang, and Trevor Cohen. 2021. Automated Lay Language Summarization of Biomedical Scientific Reviews. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(1):160–168.

Yufan Guo, Anna Korhonen, Maria Liakata, Ilona Silins, Lin Sun, and Ulla Stenius. 2010. Identifying the information structure of scientific abstracts: An investigation of three different schemes. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, pages 99–107, Uppsala, Sweden. Association for Computational Linguistics.

Yufan Guo, Ilona Silins, Ulla Stenius, and Anna Korhonen. 2013. Active learning-based information structure analysis of full scientific articles and two applications for biomedical literature review. *Bioinformatics*, 29(11):1440–1447.

Annemarie Hamlin, Chris Rubio, and Michele DeSilva. 2015. *"Audience Analysis". Technical Writing*. PressBooks.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Julie E. Howe and Michael S. Wogalter. 1994. The understandability of legal documents: Are they adequate? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 38(8):438–442.

Luyang Huang, Lingfei Wu, and Lu Wang. 2020. Knowledge graph-augmented abstractive summarization with semantic-driven cloze reward. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5094–5107, Online. Association for Computational Linguistics.

Md Saiful Islam, Tonmoy Sarkar, Sazzad Hossain Khan, Abu-Hena Mostofa Kamal, S M Murshid Hasan, Alamgir Kabir, Dalia Yeasmin, Mohammad Ariful Islam, Kamal Ibne Amin Chowdhury, Kazi Selim Anwar, Abrar Ahmad Chughtai, and Holly Seale. 2020. COVID-19–Related Infodemic and Its Impact on Public Health: A Global Social Media Analysis. *The American Journal of Tropical Medicine and Hygiene*, 103(4):1621–1629.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep

Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Frances Johnson. 1995. Automatic abstracting research. *Libr. Rev.*, 44(8):28–36.

Jiaxin Ju, Ming Liu, Huan Yee Koh, Yuan Jin, Lan Du, and Shirui Pan. 2021. Leveraging information bottleneck for scientific document summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4091–4098, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tian Kang, Adler Perotte, Casey Ta, and Chunhua Weng. 2021. Umls-based data augmentation for natural language processing of clinical research literature. In *Journal of the American Medical Informatics Association JAIMA 28(4)*, pages 812–823, Online.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeff Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *ArXiv*, abs/2001.08361.

Katherine Keith and Amanda Stent. 2019. Modeling financial analysts' decision making via the pragmatics and semantics of earnings calls. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 493–503, Florence, Italy. Association for Computational Linguistics.

Seungwon Kim. 2020. Using pre-trained transformer for better lay summarization. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 328–335, Online. Association for Computational Linguistics.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

Stuart R F King, Emma Pewsey, and Sarah Shailes. 2017. Plain-language Summaries of Research: An inside guide to eLife digests. *eLife*, 6:e25410.

Anastassia Kornilova and Vladimir Eidelman. 2019. BillSum: A corpus for automatic summarization of US legislation. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 48–56, Hong Kong, China. Association for Computational Linguistics.

Ross Koval, Nicholas Andrews, and Xifeng Yan. 2023. Forecasting earnings surprises from conference call transcripts. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8197–8209, Toronto, Canada. Association for Computational Linguistics.

Lauren M Kuehne and Julian D Olden. 2015. Lay summaries needed to enhance science communication. *Proceedings of the National Academy of Sciences*, 112(12):3585–3586.

Tuan Lai, Heng Ji, ChengXiang Zhai, and Quan Hung Tran. 2021. Joint biomedical entity and relation extraction with knowledge-enhanced collective inference. 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 6248–6260, Online. Association for Computational Linguistics.

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.

Maria Liakata. 2010. Zones of conceptualisation in scientific papers: a window to negative and speculative statements. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pages 1–4, Uppsala, Sweden. University of Antwerp.

Maria Liakata, Simon Dobnik, Shyamasree Saha, Colin Batchelor, and Dietrich Rebholz-Schuhmann. 2013. A Discourse-Driven content model for summarising scientific articles evaluated in a complex question answering task. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 747–757, Seattle, Washington, USA. Association for Computational Linguistics.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *Preprint*, arXiv:2305.19118.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Shuaiqi Liu, Jiannong Cao, Yicong Li, Ruosong Yang, and Zhiyuan Wen. 2024. Low-resource court judgment summarization for common law systems. *Information Processing Management*, 61(5):103796.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Ye Liu, Yao Wan, Lifang He, Hao Peng, and Philip S. Yu. 2021. Kg-bart: Knowledge graph-augmented bart for generative commonsense reasoning. 35:6418–6425.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. 7th International Conference on Learning Representations.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2022a. Readability controllable biomedical document summarization. In *Findings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi. Association for Computational Linguistics.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2022b. Readability controllable biomedical document summarization. In *Findings of the Association for Compu-*

tational Linguistics: EMNLP 2022, pages 4667–4680, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Chatgpt as a factual inconsistency evaluator for text summarization. *Preprint*, arXiv:2303.15621.

Yuanjie Lyu, Chen Zhu, Tong Xu, Zikai Yin, and Enhong Chen. 2022. Faithful abstractive summarization via fact-aware consistency-constrained transformer. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, page 1410–1419, New York, NY, USA. Association for Computing Machinery.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan Thomas Mcdonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of The 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Alexa McCray, Olivier Bodenreider, James Malley, and Allen Browne. 2001. Evaluating umls strings for natural language processing. In *Proceedings of AMIA Symposium*, page 448–452.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Rajdeep Mukherjee, Abhinav Bohra, Akash Banerjee, Soumya Sharma, Manjunath Hegde, Afreen Shaikh, Shivani Shrivastava, Koustuv Dasgupta, Niloy Ganguly, Saptarshi Ghosh, and Pawan Goyal. 2022. ECTSum: A new benchmark dataset for bullet point summarization of long earnings call transcripts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10893–10906, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents.

In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 3075–3081. AAAI Press.

Ramesh Nallapati, Bowen Zhou, Cicero Nogueira dos santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *Preprint*, arXiv:1602.06023.

National Science Board (NSB). 2021. Publications output: U.s. and international comparisons. Science and Engineering Indicators 2022 NSB-2021-4, National Science Foundation, Alexandria, VA. Accessed 29 June 2025.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

Hao Niu, Yun Xiong, Xiaosu Wang, Wenjing Yu, Yao Zhang, and Weizu Yang. 2023. KeFVP: Knowledge-enhanced financial volatility prediction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11499–11513, Singapore. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Ramakanth Pasunuru, Mengwen Liu, Mohit Bansal, Sujith Ravi, and Markus Dreyer. 2021. Efficiently summarizing text and graph encodings of multi-document clusters. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4768–4779, Online. Association for Computational Linguistics.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *International Conference on Learning Representations*.

Keqin Peng, Chuantao Yin, Wenge Rong, Chenghua Lin, Deyu Zhou, and Zhang Xiong. 2021. Named entity aware transfer learning for biomedical factoid question answering. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.

Chen Qian, Xin Cong, Wei Liu, Cheng Yang, Weize Chen, Yusheng Su, Yufan Dang, Jiahao Li, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2023. Communicative agents for software development. *Preprint*, arXiv:2307.07924.

Yu Qin and Yi Yang. 2019. What you say and how you say it matters: Predicting stock volatility using verbal and vocal cues. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 390–401, Florence, Italy. Association for Computational Linguistics.

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).

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

Nipun Sadvilkar and Mark Neumann. 2020. PySBD: Pragmatic sentence boundary disambiguation. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 110–114, Online. Association for Computational Linguistics.

Evan Sandhaus. 2008. The new york times annotated corpus.

Shengtian Sang, Zhihao Yang, Lei Wang, Xiaoxia Liu, Hongfei Lin, and Jian Wang. 2018. Sematyp: a knowledge graph based literature mining method for drug discovery. *BMC Bioinformatics*, pages 1471–2105.

Ramit Sawhney, Arshiya Aggarwal, and Rajiv Ratn Shah. 2021. An empirical investigation of bias in the multimodal analysis of financial earnings calls. In *Proceedings of the 2021 Conference of the North American Chapter of the Association* *for Computational Linguistics: Human Language Technologies*, pages 3751–3757, Online. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

RJ Senter and Edgar A Smith. 1967. Automated readability index. Technical report, Technical report, DTIC document.

Eva Sharma, Chen Li, and Lu Wang. 2019a. BIGPATENT: A Large-Scale dataset for abstractive and coherent summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2204–2213, Florence, Italy. Association for Computational Linguistics.

Soumya Sharma, Bishal Santra, Abhik Jana, Santosh Tokala, Niloy Ganguly, and Pawan Goyal. 2019b. Incorporating domain knowledge into medical NLI using knowledge graphs. 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 6092–6097, Hong Kong, China. Association for Computational Linguistics.

Mong Yuan Sim, Xiang Dai, Maciej Rybinski, and Sarvnaz Karimi. 2023. CSIRO Data61 team at BioLaySumm task 1: Lay summarisation of biomedical research articles using generative models. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 629–635, Toronto, Canada. Association for Computational Linguistics.

Barry Smith. 2006. From concepts to clinical reality: An essay on the benchmarking of biomedical terminologies. *Journal of Biomedical Informatics*, 39(3):288–298.

Neha Srikanth and Junyi Jessy Li. 2021. Elaborative simplification: Content addition and explanation generation in text simplification. In *Findings of the Association* *for Computational Linguistics: ACL-IJCNLP 2021*, pages 5123–5137, Online. Association for Computational Linguistics.

J. Steinberger and K. Jezek. 2004a. Using latent semantic analysis in text summarization and summary evaluation. In *Proceedings of the 7th International Conference ISIM*.

Josef Steinberger and Karel Jezek. 2004b. Using latent semantic analysis in text summarization and summary evaluation. pages 93–100. 7th International Conference ISIM.

John Swales. 1990. *Genre analysis: English in academic and research settings*. Cambridge University Press.

Chen Tang, Chenghua Lin, Henglin Huang, Frank Guerin, and Zhihao Zhang. 2022. EtriCA: Event-triggered context-aware story generation augmented by cross attention. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5504–5518.

Chen Tang, Hongbo Zhang, Tyler Loakman, Chenghua Lin, and Frank Guerin. 2023. Enhancing dialogue generation via dynamic graph knowledge aggregation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4604–4616.

Simone Teufel. 1999. *Argumentative Zoning: Information Extraction from Scientific Text*. Ph.D. thesis, University of Edinburgh.

Simone Teufel and Marc Moens. 2002. Articles summarizing scientific articles: Experiments with relevance and rhetorical status. *Comput. Linguist.*, 28(4):409–445.

Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2009. Towards discipline-independent argumentative zoning: evidence from chemistry and computational linguistics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3*, EMNLP '09, pages 1493–1502, USA. Association for Computational Linguistics.

Stéphan Tulkens and Andreas van Cranenburgh. 2020. Embarrassingly simple unsupervised aspect extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3182–3187, Online. Association for Computational Linguistics.

Oisín Turbitt, Robert Bevan, and Mouhamad Aboshokor. 2023. MDC at BioLay-Summ task 1: Evaluating GPT models for biomedical lay summarization. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 611–619, Toronto, Canada. Association for Computational Linguistics.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2017. Graph attention networks. *6th International Conference on Learning Representations*.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. 2024. Replacing judges with juries: Evaluating llm generations with a panel of diverse models. *Preprint*, arXiv:2404.18796.

Minjie Wang, Da Zheng, Zihao Ye, Quan Gan, Mufei Li, Xiang Song, Jinjing Zhou, Chao Ma, Lingfan Yu, Yu Gai, Tianjun Xiao, Tong He, George Karypis, Jinyang Li, and Zheng Zhang. 2019. Deep graph library: A graph-centric, highly-performant package for graph neural networks. *arXiv preprint arXiv:1909.01315*.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large language models are not fair evaluators. *Preprint*, arXiv:2305.17926.

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2024. Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. *Preprint*, arXiv:2307.05300.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian

Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing. *Preprint*, arXiv:1910.03771.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation framework.

Wen Xiao and Giuseppe Carenini. 2019. Extractive summarization of long documents by combining global and local context. 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 3011–3021, Hong Kong, China. Association for Computational Linguistics.

Pengcheng Yang, Lei Li, Fuli Luo, Tianyu Liu, and Xu Sun. 2019. Enhancing topic-to-essay generation with external commonsense knowledge. In *Proceedings* *of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2002–2012, Florence, Italy. Association for Computational Linguistics.

Michihiro Yasunaga, Jungo Kasai, Rui Zhang, Alexander R Fabbri, Irene Li, Dan Friedman, and Dragomir R Radev. 2019. ScisummNet: A large annotated corpus and Content-Impact models for scientific paper summarization with citation networks.

Huan Yee Koh, Jiaxin Ju, He Zhang, Ming Liu, and Shirui Pan. 2022. How far are we from robust long abstractive summarization? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2682—-2698, Abu Dhabi. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Farooq Zaman, Matthew Shardlow, Saeed-Ul Hassan, Naif Radi Aljohani, and Raheel Nawaz. 2020. HTSS: A novel hybrid text summarisation and simplification architecture. *Inf. Process. Manag.*, 57(6):102351.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint*.

Hao Zheng and Mirella Lapata. 2019. Sentence centrality revisited for unsupervised summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6236–6247.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuan-Jing Huang. 2020. Extractive summarization as text matching. In *Proceedings of* *the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6197–6208.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Commonsense knowledge aware conversation generation with graph attention. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18, International Joint Conferences on Artificial Intelligence Organization*, pages 4623–4629.

Chenguang Zhu, William Hinthorn, Ruochen Xu, Qingkai Zeng, Michael Zeng, Xuedong Huang, and Meng Jiang. 2021a. Enhancing factual consistency of abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 718–733, Online. Association for Computational Linguistics.

Tianyu Zhu, Wen Hua, Jianfeng Qu, and Xiaofang Zhou. 2021b. Summarizing Long-Form document with rich discourse information. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, pages 2770–2779, New York, NY, USA. Association for Computing Machinery.

Appendices

Appendix A

Publication 1

A.1 Example Model Outputs

Figure A.1 compares the outputs of DodoRank and HipoRank on a sample article taken from the arXiv test set.

Gold-standard Summary
finite - size scaling , finite - size corrections , and boundary effects for critical systems have attracted much attention in recent years . Here derive exact finite - size corrections for the free energy @xmath0 and the specific heat @xmath1 of the critical ferromagnetic ising model on the @xmath2 square lattice with brascamp - kunz (bk) boundary conditions [j. math . phys . *15 * , 66 (1974)] and compared such results with those under toroidal boundary conditions .when the ratio @xmath3 is smaller than 1 the behaviors of finite - size corrections for @xmath1 are quite different for bk and toroidal boundary conditions ; when @xmath4 is larger than 3 , finite - size corrections for @xmath1 in two boundary conditions approach the same values . in the limit @xmath5we obtain the expansion of the free energy for infinitely long strip with bk boundary conditions .our results are consistent with the conformal field theory prediction for the mixed boundary conditions by cardy [nucl . phys .* b 275 * , 200 (1986)] although the definitions of boundary conditions in two cases are different in one side of the long strip .l 0.2 cm pacs numbers : 05.50+q , 75.10-b

DodoRank Output
@xcite , in the present paper we derive exact finite - size corrections for the free energy @xmath0 and and the specific heat @xmath1 of the critical ferromagnetic ising model on the @xmath2 square lattice with brascamp - kunz (bk) boundary conditions @xcite and compared such results with those under toroidal boundary conditions @xcite .ii we show how to lead the partition function of ising model under brascamp - kunz (bk) boundary conditions to the form of partition function with twisted boundary conditions . in sec .we obtain the expansion of the free energy for infinitely long strip with bk boundary conditions .our results are consistent with the conformal field theory prediction for the mixed boundary conditions by cardy @xcite although the definitions of boundary conditions in two cases are different in one side of the long strip .iii asymptotic expansions of the free energy is presented . in sec .we find that when the ratio @xmath3 is smaller than 1 the behaviors of finite - size corrections for @xmath1 are quite different for bk and toroidal boundary conditions ; when @xmath4 is larger than 3 , finite - size corrections for @xmath1 in two boundary conditions approach the same values . in the limit @xmath5

HipoRank Output
([expansionpsi]) @xmath219 c .- k . in the study of phase transitions and critical phenomena , it is extremely important to understand finite - size corrections to thermodynamical quantities . in experiments and in numerical studies of model systems . japan . * 70 * , 1537 - 1542 (2001) . * fig . hu , c .- y . , j. phys . our results are summarized and discussed in sec . @xmath206 , @xmath207 and @xmath208 and the kronecker s function can again be reduce to the bernoulli polynomials . ii we show how to lead the partition function of ising model under brascamp - kunz (bk) boundary conditions to the form of partition function with twisted boundary conditions . in sec . lin , and j .- a . e * 64 * , 016127 (2001) ; h. watanabe , et al . , it is essential to take into account finite size effects in order to extract correct infinite - volume predictions from the data . therefore , in recent decades there have many investigations on finite - size scaling , finite - size corrections , and boundary effects for critical model systems @xcite .

Figure A.1: *Comparison of the output of* DodoRank *and* HipoRank *on an instance from the arXiv test set.*

Appendix B

Publication II

Availability of data sources Both PLOS and eLife are open-access journals, with an emphasis on making scientific research accessible to a wide audience. PLOS articles are available to be mined, reused, and shared by anyone, as per their data mining policy.[1] eLife articles are available under the permissive CC-BY 4.0 license, and thus also available to be retrieved and shared for these purposes.[2] The data for PLOS and eLife was retrieved on 7/03/22 and 11/03/22, respectively. All articles and summaries are in English only. Our datasets are made available to the community to facilitate future research.

Additional data processing details Here we provide some additional details regarding the dataset creation process, building on the description given in §3.3. For eLife, the XML files retrieved were found to include multiple versions of the same article, identifiable by the article id which includes a version number. For these, we removed duplicates and kept only the most recent versions.

For both datasets, prior to extraction, we remove all Tables, Figures, and sections marked with the tag "supplementary-material". We also do not extract sections with the heading "acknowledgments". During sentence segmentation, we use a regular expression to identify and temporarily replace all "et al." occurrences with unique placeholder tokens, which are then replaced following segmentation. Following

[1] https://plos.org/text-and-data-mining
[2] https://elifesciences.org/inside-elife/6933fe8e/resources-for-developers

segmentation, we again use a regular expression to identify and remove all sentences which began with "DOI:" followed by a URL from both abstracts and lay summaries, as these were found to commonly occur at the end of both.

Comparison to previous datasets We provide a comparison of the readability and abstractiveness of lay summaries for all lay summarisation datasets in Table B.1 and Figure B.1, respectively.

Supplementary figure details Table B.2 gives the exact percentages that are visualised in Figure 3.3, allowing for a more detailed analysis of content word sharing (e.g., calculating the ratio 'shared' to 'not shared' for different content word types).

Baseline model details Here we provide additional experimental details for our baselines approaches (Table 3.4). Oracle_Ext is a greedy oracle, which means it repeatedly extracts the next article sentence that will maximise the mean ROUGE scores (1, 2, and L) of the extracted summary, up to the maximum length (equal to the average lay summary length for a given dataset - Table D.4).

For all BART models, we make use of the `huggingface` library (Wolf et al., 2019). Specifically, we use the "facebook/bart-base" model for baselines BART, BART_{Cross}, and $\text{BART}_{Scaffold}$, and we use the "mse30/bart-base-finetuned-pubmed" model for BART_{PubMed}. Training was run (using 4x NVIDIA Tesla V100 SXM2 GPUs) for all models with AdamW optimisation (Loshchilov and Hutter, 2019) and an early stopping patience of 25 epochs, with the best model being selected by performance on the validation set (ROUGE-2).

All unsupervised baselines were run with default configurations.

Automatic evaluation For the calculation of ROUGE scores, we use the `rouge-score` Python package.[3] For FKGL and DCSR, we use the `textstat` Python package.[4]

Human evaluation comments Comments on the general model performance for each criterion provided by each annotator for our human evaluation are given in

[3] https://pypi.org/project/rouge-score/
[4] https://github.com/shivam5992/textstat

Dataset	FKGL	CLI	DCRS	WordRank
LaySumm	$14.81_{\pm 2.91}$	$16.28_{\pm 2.81}$	$11.63_{\pm 1.16}$	$9.04_{\pm 0.60}$
Eureka-Alert	$13.16_{\pm 2.07}$	$14.09_{\pm 1.56}$	$9.77_{\pm 0.85}$	$8.79_{\pm 0.80}$
CDSR	$12.72_{\pm 2.16}$	$14.17_{\pm 0.88}$	$9.073_{\pm 0.88}$	$8.54_{\pm 0.32}$
PLOS	$14.76_{\pm 2.33}$	$15.90_{\pm 2.01}$	$10.91_{\pm 0.85}$	$8.98_{\pm 0.32}$
eLife	$10.91_{\pm 1.44}$	$12.52_{\pm 1.35}$	$8.94_{\pm 0.53}$	$8.68_{\pm 0.31}$

Table B.1: *Comparison of the lay summary readability scores for all lay summarisation datasets.*

Figures B.2 and B.3 for PLOS and eLife, respectively.

Lay summary examples Full examples of lay summaries and their respective technical abstracts are given in Figures B.4 and B.5 for PLOS and eLife, respectively.

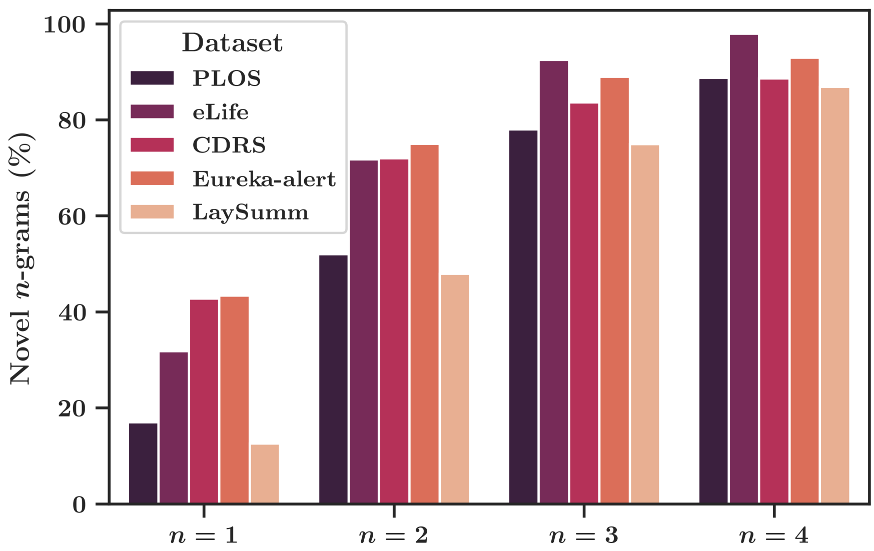


Figure B.1: *Comparison of lay summary n-gram novelty for all lay summarisation datasets.*

Comprehensiveness
Annotator 1: The model outputs summarised the important information, however it also cut out quite a lot of background info which is key for understanding the science. Overall, the comprehensiveness was enough for a non-expert to grasp the overall gist of the studies.
Annotator 2: The model seems to use parts of the abstract, and therefore seems quite comprehensive. It also does a decent job of a final “summary" sentence to the lay summary to summarize/put into context.

Layness
Annotator 1: Some abstracts contained a lot of jargon which would be confusing/off putting to a non-expert. Although I know some scientific words cannot be substituted, it would be good to have an explanation of the more complex words in brackets, for example.
Annotator 2: Due to the overlap with the reference abstract, the output is comprehensive but probably confusing to a lay audience, in some cases there is no introduction/background on scientific jargon (e.g. we cannot expect a lay audience to understand complex scientific techniques, cellular or molecular machinery). Use of Genus species nomenclature is also likely to confuse lay audiences, where a common name could be used instead, not as well as (e.g. ‘Egyptian mosquito’ instead of ‘Aedes aegypti’, also known as the Egyptian mosquito’).

Factuality
Annotator 1: The majority of the statements were factually correct, although sometimes the meaning of the simplified language could be misinterpreted, which would result in a similar outcome to factually wrong statements.
Annotator 2: Some minor factual errors throughout, and mixups between gene symbols (e.g. where one letter will be changed, PMN to PMA), there are also come cases where it will pull out a % but mix up what gene / condition it is related to, ultimately leading to the formation of a sentence which is factually incorrect.

Figure B.2: *Human evaluation comments for PLOS.*

Comprehensiveness
Annotator 1: Overall the information contained within the model-generated summaries effectively conveyed the information in the references. However, there was a few occasions where new elements/concepts were introduced that could confuse the reader and affect their understanding (these were sometimems factual statements, sometimes seemingly made up). The model abstract did provide enough information for a general understanding of the topic and would be sufficient as a brief overview.
Annotator 2: The model usually picks up on the core points of the abstract but can often introduce extra information which is either off-topic or factually incorrect. The model seems to start off by introducing the topic well but struggles to hit the "what is the significance of this research?" question.

Layness
Annotator 1: The language was, in the majority, well suited to a lay person and terminology was adapted accordingly. However, at times the information was simplified to a point where it could be misconstrued which, with scientific information, is a potential risk. At times, jargon still remained and I could imagine some people being confused by this. There were a few grammatical errors and poor sentence structure, typos, repetition etc.
Annotator 2: Sometimes the model introduces extra information which is not suitable for a lay audience, for example: references to genome sequencing, progenitor cells, endoplasmid reticulum. There are also instances where misinterpretation by the model may mislead the lay audience, for example there was an output where Norepinephrine was said to be "a.k.a. dopamine", which is not factually correct.

Factuality
Annotator 1: There were a few summaries which contained incorrect information, things that are well-known in the scientific community were poorly conveyed. At times, new information was introduced which contradicted earlier statements, and those of the reference abstracts/lay summaries. Of course, some information was correct. I would be concerned about the level of misinformation which could arise from these summaries, if used to educate a lay audience.
Annotator 2: This seemed to vary based on the abstract and how well the output started, for example if the model introduced the topic well, it would lead to more factual points. However, there were some generated summaries which were factually incorrect from the start and this lead to more errors.

Figure B.3: *Human evaluation comments for eLife.*

Technical Abstract

Rabies is a uniformly fatal disease, but preventable by timely and correct use of post exposure prophylaxis (PEP). Unfortunately, many health care facilities in Pakistan do not carry modern life-saving vaccines and rabies immunoglobulin (RIG), assuming them to be prohibitively expensive and unsafe. Consequently, Emergency Department (ED) health care professionals remain untrained in its application and refer patients out to other hospitals. The conventional Essen regimen requires five vials of cell culture vaccine (CCV) per patient, whereas Thai Red Cross intradermal (TRC-id) regimen requires only one vial per patient, and gives equal seroconversion as compared with Essen regimen. This study documents the cost savings in using the Thai Red Cross intradermal regimen with cell culture vaccine instead of the customary 5-dose Essen intramuscular regimen for eligible bite victims. All patients presenting to the Indus Hospital ED between July 2013 to June 2014 with animal bites received WHO recommended PEP. WHO Category 2 bites received intradermal vaccine alone, while Category 3 victims received vaccine plus wound infiltration with Equine RIG. Patients were counseled, and subsequent doses of the vaccine administered on days 3, 7 and 28. Throughput of cases, consumption utilization of vaccine and ERIG and the cost per patient were recorded. Government hospitals in Pakistan are generally underfinanced and cannot afford treatment of the enormous burden of dog bite victims. Hence, patients are either not treated at all, or asked to purchase their own vaccine, which most cannot afford, resulting in neglect and high incidence of rabies deaths. TRC-id regimen reduced the cost of vaccine to 1/5th of Essen regimen and is strongly recommended for institutions with large throughput. Training ED staff would save lives through a safe, effective and affordable technique.

Lay Summary

Rabies is a killer disease caused by the rabies virus that is present in the saliva of rabid animals, mainly the dog. Once symptoms become apparent, death is inevitable. However, rabies can be prevented if correct post exposure prophylaxis (PEP) is instituted as soon as possible after a dog bite and before symptoms of rabies begin. In Pakistan, government hospitals treat 50-70 new bite victims each day. Many still dispense the free but poor quality nerve tissue vaccine that is often ineffective and fraught with serious adverse reactions. Hospital administrators consider PEP too expensive to be administered free of cost. The Indus Hospital (TIH), Karachi is a private teaching hospital which provides free medical care to all. From July 2013-June 2014, 2,983 new bites were seen in the ED, and rather than use the Essen regimen of five full vial intramuscular doses per patient over 28 days, we administered the WHO-approved Thai Red Cross-intradermal (TRC-id) 4-dose regimen. The use of the TRC-id regimen resulted in 80% cost savings over the Essen regimen. In resource-poor settings, we advocate training of ED personnel in TRC-id regimen, which, ultimately, will result in less vaccine consumption, more patient compliance and complete treatment, resulting in more lives being saved.

Figure B.4: *PLOS lay summary example.*

Technical Abstract

Adult stem cells are responsible for life-long tissue maintenance. They reside in and interact with specialized tissue microenvironments (niches). Using murine hair follicle as a model, we show that when junctional perturbations in the niche disrupt barrier function, adjacent stem cells dramatically change their transcriptome independent of bacterial invasion and become capable of directly signaling to and recruiting immune cells. Additionally, these stem cells elevate cell cycle transcripts which reduce their quiescence threshold, enabling them to selectively proliferate within this microenvironment of immune distress cues. However, rather than mobilizing to fuel new tissue regeneration, these ectopically proliferative stem cells remain within their niche to contain the breach. Together, our findings expose a potential communication relay system that operates from the niche to the stem cells to the immune system and back. The repurposing of proliferation by these stem cells patch the breached barrier, stoke the immune response and restore niche integrity.

Lay Summary

Most, if not all, tissues of an adult animal contain stem cells. These stem cells regenerate and repair damaged tissues and organs for the entire lifetime of an animal, contributing to a healthy life. They divide to make daughter cells that become either new stem cells or specialized cells of that organ. Adult stem cells exist in specific areas within tissues known as niches, where they interact with surrounding cells and molecules that inform their behavior. For example, cells and molecules within these niches can signal stem cells to remain in a 'dormant' state, but upon injury, they can mobilize stem cells to form new tissue and repair the wound. So far, it has been unclear how stem cells sense damage and stress and direct their efforts away from their normal duties towards repair. Here, Lay et al. studied the stem cells in the mouse skin that are responsible to regenerate hair. Every hair follicle contains a niche (the 'bulge'), where these stem cells live and share their environment with cells that anchor the hair. The niche tethers to the stem cells through specific adhesion molecules that also help the niche to form a tight seal to prevent bacteria from entering. Lay et al. removed one of the adhesion molecules called E-cadherin, which caused a breach in the niche's barrier. The stem cells sensed their damaged niche, prepared to multiply, and sent out stress signals to the immune system. The immune cells then arrived at the niche and sent signals back to the stem cells, prodding them to multiply and patch the barrier, while at the same time, keeping the inflammation in check. This remarkable ability of the stem cells to recruit immune cells and initiate a dialogue with them enabled the stem cells to divert their attention from regenerating hair and instead directing it towards the site of the tissue damage. Other stem cells, such as those in the lung or gut, may have similar mechanisms to detect and respond to physical damage. It will be interesting to investigate the underlying mechanism of how immune cells are involved in balancing stem cell regenerative capacity and response to physical damage. A better knowledge of these processes could help to regenerate tissues or even entire organs.

Figure B.5: *eLife lay summary example.*

	Word type	**# of abstract occurrences**			
		1	**2-10**	**11-100**	**100+**
PLOS	Noun	15.3 / 48.2	8.7 / 18.1	2.5 / 4.9	0.8 / 1.5
	Proper noun	15.3 / 54.2	8.2 / 18.0	1.9 / 2.9	0.2 / 0.3
	Number	6.3 / 67.6	2.4 / 18.6	0.5 / 3.8	0.2 / 0.6
	Verb	4.2 / 28.9	4.9 / 29.7	3.4 / 18.9	1.8 / 8.3
eLife	Noun	14.9 / 46.2	8.6 / 20.3	2.5 / 6.2	0.5 / 0.9
	Proper noun	19.2 / 64.1	6.4 / 9.6	0.3 / 0.4	0.0 / 0.0
	Number	7.6 / 67.6	3.8 / 16.8	1.2 / 2.6	0.2 / 0.2
	Verb	3.8 / 32.7	4.5 / 36.0	2.6 / 17.2	0.6 / 2.6

Table B.2: *Statistics for bars in Figure 3.3. For each table cell, the overall percentage is split into '% that are shared with lay summaries' / '% that are not shared with lay summaries'.*

Appendix C

Publication III

Node type	Average Count
Document	1.00
Section	5.90
Metadata	7.49
Concept	364.93
SemType	63.47

Table C.1: *The average node type frequency statistics for a single article in the train split.*

Graph Statistics Table C.1 presents the average node type frequencies in the graph of a given article. Additionally, Tables C.4 and C.5 present the average semantic type (SemType) node frequencies, and Table C.6 presents the average relation frequencies (all of which are located at the end of the Appendix).

MetaMap - UMLS vocabularies. As mentioned in §4.3, we restrict MetaMap to a select number of English vocabularies with access restrictions lower than level

Model	Abstractiveness (%)
Longformer	22.95
– text-aug	20.99
– doc-enhance	20.00
– decoder-attn	24.46

Table C.2: *The average abstractiveness (measured in terms of novel 1-grams) of summaries generated using different graph integration methodologies.*

Model	Relevance				Readability		Factuality
	R-1↑	R-2↑	R-L↑	BeS↑	CLI↓	DCRS↓	BaS↑
Longformer	47.23	13.20	44.44	85.11	11.72	9.09	-2.56
− text-aug + doc-enhance	46.28	13.45	43.77	84.72	11.37	8.06	-2.56
− text-aug + decoder-attn	47.27	13.93	44.52	84.95	10.47	8.15	-2.57
− doc-enhance + decoder-attn	28.85	3.7	27.58	73.60	2.68	1.56	-6.48

Table C.3: *Average performance of models with combinations of KG-enhancement methods on eLife test split.* ***R*** *= ROUGE F1,* ***BeS*** *= BERTScore F1,* ***CLI*** *= Coleman-Liau Index,* ***BaS*** *= BARTScore.*

4.[1] Specifically, we allow MetaMap to use the following vocabularies: Alcohol and Other Drug Thesaurus (AOD), Diseases Database (DDB), CRISP Thesaurus (CSP), DrugBank (DRUGBANK), Diagnostic and Statistical Manual of Mental Disorders - Fifth Edition (DSM-5), Human Phenotype Ontology (HPO), NCI Thesaurus (NCI), MedlinePlus Health Topics (MEDLINEPLUS), LOINC (LNC), MeSH (MSH), RxNorm (RXNORM), and Gene Ontology (GO).

MetaMap - Noise reduction. We found that MetaMap, in addition to accurately identifying UMLS concepts that were mentioned in a passage of text, often also returned a large number of unwanted concepts. These typically were only distantly relevant to the text - for example, multi-word concepts with one matching word. Therefore, for each section of an article, we employed a simple word overlap-based to filter out unwanted retrieved concepts. Specifically, after removing stopwords from the main text and lemmatisation concept names, we retain only the concepts for which all words in its name were found to appear in the section text. We empirically found this approach to significantly outperform word embedding-based approaches, which often failed to filter out irrelevant multi-word concepts if a single word appeared in the main text.

Additional implementation and training details. We employ Longformer Encoder Decoder (LED) with the `allenai/led-base-16384` Huggingface checkpoint as our base model for all knowledge enhancement approaches, using the default parameters for model training and an input limit of 8192.

[1]Information on all possible MetaMap vocabularies and their access restrictions can be found on the following page: https://www.nlm.nih.gov/research/umls/sourcereleasedocs/index.html

For the “decoder cross-attention" method, we replicate the standard Huggingface LED multi-head decoder cross-attention implementation (where the head no. is determined by the model configuration), making use of the GAT-produced graph embedding as the key and value matrices (in place of the encoder output), with the query matrix being the output of the previous standard cross-attention module.

For the “document embedding enhancement" method, we replicate the standard LED encoder layer, adapting the configuration to account for the larger input of the concatenated document and graph feature representations.

For the “article text augmentation" method, we prepend the article text with the graph-derived text and shown in Figure C.2, extending the input text limit to 16,384 to accommodate the lengthy augmentation text without losing article text. All models are trained for a maximum of 20 epochs, retaining the checkpoints that achieved the highest validation set performance (average ROUGE score across variants).

The time taken to train each model on 2 A100 GPUs ranged from 12 hours to 2 days depending on the specific methodology used, with the GAT-based methods - “document embedding enhancement” and “decoder cross attention” - taking longer than the “article text augmentation" method.

Summary abstractiveness. In an effort to gain further insight into how each KG-enhancement method influences summary generation, we measure the abstractiveness (as the percentage of novel 1-grams) of the generated summaries, providing the results in Table C.2. Interestingly, the results show that both the “text augmentation” and “document enhancement” methods slightly decrease the abstractiveness of the base model, but the “decoder attention” method slightly increases it, suggesting that this method is more effective at introducing external vocabulary.

Combining KG-enhancement Methods Table C.3 presents the results obtained when applying different combinations of the KG-enhancement methods to Longformer. Interestingly, no combined model matches the overall performance of any single KG method model. Although these combinations achieve positive results for readability metrics, this generally comes at the expense of readability metrics, with the combination of text augmentation and decoder attention being the only model to achieve ROUGE scores equal to that of the base model.

We believe this performance degradation is likely a result of the dilution or loss of document information at the expense of graph-based information. This is similar to what was observed when the value of p, as given in equation (4.4), was set to too great a value for the document enhancement method. In the case of the method combinations that include the text augmentation method, the document information is likely more diluted as a result of the larger embedding size (as caused by the increased input size). Alternatively, when the document embedding and decoder attention methods are combined, the original document information seems to be lost, resulting in particularly high loss values during training and the model being unable to produce a coherent output (as evidenced by the extremely low scores for readability and factuality metrics).

Case study Figure C.1 presents an expanded version of the case study given in Figure 6.4, whereby we show all model outputs for each of the technical concepts. As in Figure 6.4, we see that the KG-enhanced models typically improve on the explanation proved by the base Longformer model.

Semantic type	Average Count
Manufactured Object	0.98
Classification	0.61
Qualitative Concept	1.0
Disease or Syndrome	0.51
Health Care Activity	0.95
Occupational Activity	0.94
Phenomenon or Process	0.99
Individual Behavior	0.26
Intellectual Product	1.0
Organism Attribute	0.96
Health Care Related Organization	0.47
Body Part, Organ, or Organ Component	0.86
Social Behavior	0.67
Therapeutic or Preventive Procedure	0.78
Research Device	0.13
Spatial Concept	1.0
Temporal Concept	1.0
Behavior	0.45
Environmental Effect of Humans	0.05
Mammal	0.76
Research Activity	1.0
Quantitative Concept	1.0
Occupation or Discipline	0.81
Functional Concept	1.0
Conceptual Entity	0.81
Activity	1.0
Population Group	0.65
Age Group	0.35
Mental Process	0.92
Natural Phenomenon or Process	0.93
Geographic Area	0.78
Biologic Function	0.43
Human-caused Phenomenon or Process	0.20
Idea or Concept	1.0
Body Location or Region	0.58
Finding	1.0
Organ or Tissue Function	0.38
Amino Acid, Peptide, or Protein	0.89
Injury or Poisoning	0.13
Professional or Occupational Group	0.50
Gene or Genome	0.84
Element, Ion, or Isotope	0.91
Body Substance	0.56
Cell	0.91
Receptor	0.50
Hazardous or Poisonous Substance	0.80
Pathologic Function	0.59
Organism	0.46
Indicator, Reagent, or Diagnostic Aid	0.81
Cell Component	0.82
Biologically Active Substance	0.91
Animal	0.52
Biomedical or Dental Material	0.63
Group	0.38
Body System	0.26
Physical Object	0.34
Antibiotic	0.39
Organism Function	0.95
Governmental or Regulatory Activity	0.48
Organization	0.61
Tissue	0.40
Diagnostic Procedure	0.41
Biomedical Occupation or Discipline	0.68
Entity	0.40

Table C.4: *The average semantic node type frequency statistics for a single article in the train split.*

Semantic type	Average Count
Inorganic Chemical	0.66
Pharmacologic Substance	0.93
Physiologic Function	0.33
Immunologic Factor	0.58
Molecular Function	0.67
Clinical Attribute	0.39
Laboratory Procedure	0.88
Event	0.59
Human	0.04
Chemical Viewed Structurally	0.50
Sign or Symptom	0.28
Enzyme	0.68
Medical Device	0.78
Genetic Function	0.73
Nucleic Acid, Nucleoside, or Nucleotide	0.75
Organic Chemical	0.88
Patient or Disabled Group	0.23
Virus	0.20
Cell Function	0.74
Substance	0.94
Daily or Recreational Activity	0.72
Bacterium	0.37
Laboratory or Test Result	0.17
Neoplastic Process	0.17
Eukaryote	0.33
Amino Acid Sequence	0.30
Cell or Molecular Dysfunction	0.57
Regulation or Law	0.09
Chemical	0.45
Body Space or Junction	0.23
Nucleotide Sequence	0.37
Chemical Viewed Functionally	0.37
Family Group	0.52
Vertebrate	0.18
Fungus	0.23
Food	0.40
Embryonic Structure	0.28
Bird	0.19
Molecular Biology Research Technique	0.56
Molecular Sequence	0.06
Group Attribute	0.03
Vitamin	0.16
Mental or Behavioral Dysfunction	0.30
Hormone	0.19
Plant	0.18
Anatomical Structure	0.06
Fish	0.14
Machine Activity	0.25
Educational Activity	0.09
Reptile	0.02
Amphibian	0.03
Congenital Abnormality	0.04
Experimental Model of Disease	0.06
Archaeon	0.02
Language	0.02
Anatomical Abnormality	0.01
Acquired Abnormality	0.01
Fully Formed Anatomical Structure	0.01
Professional Society	0.01
Clinical Drug	0.00
Self-help or Relief Organization	0.00

Table C.5: *The average semantic node type frequency statistics for a single article in the train split (continued).*

Relation type	Average Count
contains	564.94
has_keyword	1.59
has_title	4.90
was_published_in	1.0
is_a	536.22
branch_of	0.86
affects	255.99
performs	25.23
exhibits	6.29
conceptual_part_of	10.80
result_of	134.99
measures	71.28
issue_in	97.14
associated_with	43.36
occurs_in	15.44
connected_to	2.61
process_of	101.00
degree_of	12.83
manifestation_of	37.81
uses	14.00
location_of	60.24
causes	42.52
adjacent_to	6.13
tributary_of	0.86
prevents	6.84
produces	67.90
method_of	13.61
property_of	6.45
complicates	54.91
evaluation_of	22.34
co-occurs_with	21.40
carries_out	8.23
interacts_with	102.22
measurement_of	17.90
traverses	1.45
developmental_form_of	3.52
treats	8.34
surrounds	4.74
part_of	27.68
conceptually_related_to	0.31
precedes	30.83
consists_of	6.54
analyzes	16.93
assesses_effect_of	23.76
disrupts	49.62
contains	2.97
diagnoses	8.38
indicates	3.30
practices	0.78
manages	0.77
derivative_of	0.28
interconnects	0.70
ingredient_of	0.05

Table C.6: *The average relation type frequency statistics for a single article in the train split.*

[Meiosis]
Longformer - During meiosis, the DNA in one of the chromosomes is copied and then the two copies are recombined so that each new generation will have a single copy of the gene that encodes the protein encoded by that gene.**1/2**
w/ text-aug - ... a process known as meiosis ... two copies of each chromosome are then exchanged between the newly formed cells, which results in a unique set of genes being passed on to the next generation.**2/2**
w/ doc-enhance - ... meiosis depends on the rate and location of the recombination events, and on the length of the DNA that is involved in the process.**1/2**
w/ decoder-attn - This is achieved through a process called meiosis, in which a cell divides to produce four daughter cells, each of which contains a copy of one of the four copies of the genes found on the parent cell's chromosomes. **2/2**

[Glabrous Skin / Mechanoreceptors]
Longformer - The orientation of an object depends largely on how its edges activate mechanoreceptors in the glabrous skin of the fingertips.**0/2**
w/ text-aug - ... information about the orientation of an object is based largely on how its edges activate mechanoreceptors in the skin of the fingertips.**2/2** ... tactile neurons in the fingertips are able to detect the edges of objects with high sensitivity**2/2** ...
w/ doc-enhance - The fingertip's surface is covered by a ... layer of skin known as the glabrous skin.**2/2** These cells are responsible for sensing touch, and they are also responsible for detecting the orientation of objects that touch them.**2/2**
w/ decoder-attn - The orientation of an object is determined largely by how its edges contact the skin and activate nerve cells in the fingertips.**2/2**

[Slow Wave Sleep]
Longformer - Most studies of sleep have focused on a type of sleep known as slow wave sleep, in which the brain's activity alternates between periods of alternating periods of slow and fast sleep.**0/2**
w/ text-aug - slow wave sleep is thought to help us to consolidate newly acquired memories and to consolidate our sense of awareness of the world around us.**2/2** ... in slow wave sleep, the propagation of slow waves is lost between the brain's regions.**2/2**
w/ doc-enhance - Slow wave sleep is characterized by a slow oscillation of electrical activity in the brain.**2/2**
w/ decoder-attn - Slow wave sleep is characterized by rhythmic waves of electrical activity in the brain, which are thought to be part of the process by which the brain consolidates memories.**2/2**

Figure C.1: *An extended case study comparing how the application of each method affects the explanation of specific technical concepts within the human evaluation sample. Colours and superscript are used to denote the number of evaluators who judged the sentence as readable for a lay audience (e.g.,* **2/2** *= 2 out of 2 evaluators).*

{Concept definitions and relations}
Alleles = Variant forms of the same gene, occupying the same locus on homologous CHROMOSOMES, and governing the variants in production of the same gene product. Alleles is a Gene or Genome.
Molecule = An aggregate of two or more atoms in a defined arrangement held together by chemical bonds. Molecule is a Substance.
Discover = See for the first time; identify. Discover is a Activity.
Histocompatibility = The degree of antigenic similarity between the tissues of different individuals, which determines the acceptance or rejection of allografts. Histocompatibility is a Qualitative Concept.
In Vivo = Located or occurring in the body. In Vivo is a Spatial Concept.
Species = A group of organisms that differ from all other groups of organisms and that are capable of breeding and producing fertile offspring. Species is a Classification.
Cells = The fundamental, structural, and functional units or subunits of living organisms. They are composed of CYTOPLASM containing various ORGANELLES and a CELL MEMBRANE boundary. Cells is a Cell.
Allogeneic = Taken from different individuals of the same species. Allogeneic is a Qualitative Concept.
Antigens = Substances that are recognized by the immune system and induce an immune reaction. Antigens is a Immunologic Factor.
Result = The result of an action. Result is a Functional Concept.
...

{SemType definitions}
Gene or Genome = A specific sequence, or in the case of the genome the complete sequence, of nucleotides along a molecule of DNA or RNA (in the case of some viruses) which represent the functional units of heredity.
Substance = A material with definite or fairly definite chemical composition.
Activity = An operation or series of operations that an organism or machine carries out or participates in.
Classification = A term or system of terms denoting an arrangement by class or category.
Cell = The fundamental structural and functional unit of living organisms.
Qualitative Concept = A concept which is an assessment of some quality, rather than a direct measurement.
Spatial Concept = A location, region, or space, generally having definite boundaries.
Immunologic Factor = A biologically active substance whose activities affect or play a role in the functioning of the immune system.
Functional Concept = A concept which is of interest because it pertains to the carrying out of a process or activity.
...

{Article text}

Figure C.2: *Text augmentation example.*

Appendix D

Publication IV

Role	Prompt
Author	You are the author of the given research article tasked with answering the following questions about your work. When answering these questions please bear in mind that the audience for your answers is broad, and includes researchers from other fields (who will know relatively little about your own field of research) and interested members of the public. *{questions}* ### ARTICLE *{article}*
Writer	You are a freelance writer, tasked with summarising a biomedical research article for a lay audience. In addition to the article itself, the authors have answered a short questionnaire about their work. Using both the article text and the author-provided answers, summarize the article for a non-expert audience. Your summary should be between 300 and 400 words and contain minimal jargon, often using words and phrases that aren't present in the article. The first half of your summary should focus on explaining the background information that a lay audience will require, and the second half should explain the key experiments and results, finishing with a concluding sentence about the significance of the article. ### ARTICLE *{article}* ### ANSWERS *{question_answers}*

Table D.1: *Prompts provided to for each LLM role in the proposed two-stage framework.*

D.1 Prompts

Here we provide full details of the prompts used in each stage of our methodology.

Base model prompt As mentioned in §5.3, for our base approach, we utilise the prompt "Generate a summary of the following article that is suitable for non-experts". This prompt was selected as it was found to obtain the best overall performance in a series of preliminary experiments, whereby we tested several variations of this instruction using different models.

Task decomposition prompts Firstly, we provide the prompts used by in both stages of the proposed two-stage framework in Table D.1. Note that, when providing these prompts to different LLMs, the structure and notation is sometimes altered slightly to conform with the recommended prompt template, but content remains the same.

Within Table D.1, variable text is represented as: *{variable}*. To clarify the meaning of given variables, *{questions}* refers to the questions and guidelines that eLife provides authors - we provide these verbatim in Table D.2. *{article}* refers to the input text from the article which is to be summarised (as covered in Appendix D.2, we experiment with different techniques for input text selection). Finally, *{answers}* refers to the answers generated by the LLM from the author question-answering stage.

LLM evaluator prompt Table D.3 provides the prompt given the LLM judges in order to extract preference judgements. As mentioned in §5.3, to avoid positional bias, we randomly change the order in which the summaries generated by each approach are given, such that 50% of each possible ordering is used for a given model.

Question and guidance
1. What background information would someone who is completely unfamiliar with your field need to know to understand the findings in your paper? (Suggested word limit: 150 words) - Include something that most readers will be able to relate to in the first sentence. Get gradually more specific in the following sentences. - Don't try to explain the background to your entire field; instead consider which details a reader would need to know to understand the new findings, and then explain these facts as clearly and concisely as you can. - Make sure to provide simple definitions or explanations for all technical terms and acronyms.
2. What exact research question did you set out to answer and why? (Suggested word limit: 75 words) - Provide context by making it clear if this question was asking something completely new, or if you wanted to test or build upon previous findings. - Make sure that you explain why it was important to find an answer this question (why should people care whether you can answer this question or not?).
3. What are the most important findings of your paper? (Suggested word limit: 100 words) - Focus on findings highlighted in the title or abstract of your paper, and explain them clearly and completely. - If possible, describe your methodology with a sentence or two. - Always mention which species, type of organism or cells you have studied (for example, mutant mice, fruit flies, human kidney cells, or cancer cells).
4. Who might eventually benefit from the findings of your study, and what would need to be done before we could achieve these benefits? (Suggested word limit: 75 words) - Think beyond your immediate field of research, and explain how your findings could lead to a benefit for wider society (patients, the environment, and so on). - Avoid hype or exaggeration. For example, if your findings are about a fundamental process in living cells that could be relevant to understanding cancer, you should mention the link but be careful not to imply that the findings will imminently lead to new treatments.

Table D.2: *Questions and recommendations provided by eLife to accepted authors.*

Prompt
You are a tasked with indicating your preference between two research article summaries that are intended for a non-expert audience. Your preference should be based on which summary you believe would be more useful in informing a lay audience about the findings and significance of the article. Respond with the number of the report you would recommend and a brief explanation of why you would recommend it. ### SUMMARY 1 *{summary1}* ### SUMMARY 2 *{summary2}* ### PREFERENCE

Table D.3: *Preference judgement prompt.*

# Docs	Doc	Summary	
	# words	# words	# sents
4,828	7,806.1	347.6	15.7

Table D.4: *Average statistics of eLife dataset.*

D.2 Additional Experimental Details

Here we provide additional details relating to experiments and results that are not given in the main text.

Dataset statistics The statistics relating to the length of the documents and summaries in the eLife dataset are provided in Table D.4.

Metric calculation ROUGE scores were calculated using the existing `rouge-score` package, with stemming and sentence tokenization applied FKGL and DCRS were computed using the `textstat` package.

Input selection experiment We perform additional experiments whereby we seek to identify which input format provides the best performance. Previous Lay Summarisation works have taken somewhat varied approaches when it comes to

Model	QA	# Secs	Relevance				Readability		Factuality
			R-1↑	R-2↑	R-L↑	BeS↑	FKGL↓	DCRS↓	BaS↑
Mixtral	✗	A	45.59	11.02	42.85	84.17	14.08	9.69	-3.06
Mixtral	✗	A+I	43.79	9.95	41.11	83.45	15.13	10.10	-3.09
Mixtral	✗	all	40.65	9.09	37.99	83.33	15.01	10.43	-3.46
Mixtral	✓	A	45.49	10.41	42.90	83.95	13.92	9.57	-3.12
Mixtral	✓	A+I	44.93	10.19	42.31	83.76	14.35	9.67	-3.05
Mixtral	✓	all	42.75	9.48	40.13	83.64	14.57	10.04	-3.48
LLAMA3	✗	A	45.59	11.22	43.03	84.72	10.37	8.43	-3.34
LLAMA3	✗	A+I	45.32	11.89	42.51	84.92	11.59	9.01	-3.04
LLAMA3	✗	all	16.98	2.63	15.68	67.02	58.08	10.45	-5.43
LLAMA3	✓	A	44.90	10.68	41.93	84.93	11.99	9.34	-3.27
LLAMA3	✓	A+I	46.38	11.47	43.43	85.03	12.47	9.37	-3.18
LLAMA3	✓	all	16.23	2.48	14.69	66.72	62.96	10.15	-5.23

Table D.5: *The effect of input selection on the performance of models on eLife test split.* ***R*** *= ROUGE F1,* ***BeS*** *= BERTScore F1,* ***FKGL*** *= Flech-Kincaid Grade Level,* ***DCRS*** *= Dale-Chall Readability Score,* ***BaS*** *= BARTScore.*

input selection. One common approach is to attempt to utilise the input article in full, or at least truncated at the maximum context length of the selected model (Goldsack et al., 2022a, 2023c). However, past works have also demonstrated good performance using only the article abstract as input (Turbitt et al., 2023; Guo et al., 2021), minimising the contextual burden placed upon the model.

Table D.5 offers a comparison of multiple methods of input selection for the Mixtral and LLAMA3 model using automatic metrics. Specifically, we experiment with abstract-only (A), abstract+introduction (A+I), and full-article (full). Interestingly, our results show that using only the abstract as input yields the best overall performance in the majority of cases, with the exception of LLAMA3 utilising our two-stage framework. Therefore, in our primary experiments, we utilise only the abstract as the article text for all models, in order to ensure a fair comparison.

We also find that models that utilise the full article perform considerably worse than their equivalents with shorter inputs, indicating that content at the start of the article is significantly more relevant for lay summary production, and that zero-shot models may struggle to identify relevant content when presented with the full article. LLAMA was found to particularly struggle with the full article input, being unable to follow the instructions provided and produce a coherent output.

Ablation study experiment To determine the efficacy and contribution of each aspect of our two-stage framework, we perform an ablation study whereby we

Model	**Relevance**				**Readability**		**Factuality**
	R-1↑	**R-2**↑	**R-L**↑	**BeS**↑	**FKGL**↓	**DCRS**↓	**BaS**↑
Mixtral	45.49	10.41	42.90	83.95	13.92	9.57	-3.12
Mixtral$_{no_guides}$	44.41	9.81	41.71	83.81	14.67	9.67	-3.12
Mixtral$_{no_roles}$	43.33	9.67	40.78	83.36	14.54	9.69	-3.03
Mixtral$_{single_prompt}$	44.29	10.13	41.49	84.18	14.24	9.87	-3.15
LLAMA3	45.59	11.22	43.03	84.72	10.37	8.43	-3.34
LLAMA3$_{no_guides}$	44.57	10.41	41.68	84.72	12.90	9.68	-3.28
LLAMA3$_{no_roles}$	44.94	11.01	42.12	84.87	12.63	9.58	-3.20
LLAMA3$_{single_prompt}$	45.81	10.41	43.25	84.08	13.03	9.28	-3.74

Table D.6: *Average performance of ablated versions of the two-stage models on eLife test split.* ***R*** *= ROUGE F1,* ***BeS*** *= BERTScore F1,* ***FKGL*** *= Flech-Kincaid Grade Level,* ***DCRS*** *= Dale-Chall Readability Score,* ***BaS*** *= BARTScore.*

systematically remove various prompt components. Again, we utilise Mixtral and LLAMA 3 as the base models for this experiment, and measure model performance using automatic metrics. The results of this evaluation are presented in Table D.6; and we divide our discussion based on the element removed from the system:

Expert guidelines - The removal of our expert-derived guidelines from the instruction prompts of the Writer is denoted by the LLM$_{no_guides}$. Here, we observe a decrease in performance compared to the base model.

Role descriptions - We also experiment with the removal of role descriptions from the prompts of the Author and Writer stages which has shown to be significant in previous work (Wang et al., 2024; Chan et al., 2023). In line with their findings, both of the LLM$_{no_roles}$ model exhibits a surprisingly large drop in performance compared to the base models, attesting to the importance of role-playing in a setting where LLMs are used to represent different actors.

Multi-stage format - Finally, the LLM$_{single_prompt}$ model in Table D.6 represents a final experiment whereby the Author stage is effectively removed, and the Writer is asked to answer the author questions before composing the lay summary within a single prompt. Again, we observe worse overall performance compared to the two-stage setting.

Inter-annotator agreement Here we discuss the inter-annotator agreement by both human and LLM judges for the preference-based evaluation in Table 5.1 and the in-depth evaluation in Figure 5.2.

Starting with the preference-based evaluation (Table 5.1) on biomedical articles, we measure both Cohen's κ and the overall percentage of agreement obtained between pairs of annotators across all 100 instances (i..e, combining the 20 samples summaries for 5 the different models for which we collect preference judgements). For human evaluators, who's preferences are aggregated to obtain the PoH metric, we find that pairs of evaluators agree 55.17% of the time and get an average κ of 0.103 (across 4 lay evaluators). This corroborates the findings of Goyal et al. (2022), who identify an inherent variance the values of annotators when it comes to indentifying which summary they consider to be the "best" or, in our case, "most useful". Interestingly, we see a similar pattern for pairs of LLM judges, who we found to agree on 46.67% of instances, and obtain an average κ of -0.067, highlighting that there is variance in the individual preferences of both humans and LLMs when it comes to judging lay summaries. Comparing the majority vote of LLM judges with each annotator, we obtain an average κ of -0.06 and an average agreement of 46.75% indicating that even combined LLM preferences for individual samples fail to consistently reflect those of human judges. However, despite relative inter- and intra-disagreement between LLM and human judges, we observe that the overall averaged preferences for the proposed two-stage framework consistently increases in line with model size for both judge types. This potentially suggests that some broader population-level patterns may be captured by LLM judges, even though they are not specific to individual samples.

For our question-based human evaluation Mixtral variants for NLP articles (Figure), we again measure the inter-annotator agreement using Cohen's κ between expert evaluators. For this, we get an average pairwise score of 0.334 which, notably, increases to 0.428 if we combine the respective "agree" and "disagree" variables.

Comparison to fine-tuned LMs Table D.7 presents the automatic metrics scores obtained by BART (Lewis et al., 2020) and Longformer (Beltagy et al., 2020) models that have been fine-tuned on the eLife training set, with these being the two LMs that have been most widely used in previous Lay Summarisation works.

Model	Relevance				Readability		Factuality
	R-1↑	**R-2↑**	**R-L↑**	**BeS↑**	**FKGL↓**	**DCRS↓**	**BaS↑**
BART	46.57	11.65	43.70	84.94	10.95	9.36	-2.39
Longformer	47.23	13.20	44.44	85.11	11.88	9.09	-2.56

Table D.7: *Average performance of fine-tuned LMs on eLife test split.* ***R*** *= ROUGE F1,* ***BeS*** *= BERTScore F1,* ***FKGL*** *= Flesch-Kincaid Grade Level,* ***DCRS*** *= Dale-Chall Readability Score,* ***BaS*** *= BARTScore*

Model	QA	Readability		Factuality	H2H
		FKGL↓	**DCRS↓**	**BaS↑**	**PoLL↑**
Phi2	✗	14.05	8.21	-3.25	0.75
	✓	13.40	7.62	-3.77	0.25
Mistral	✗	13.46	9.89	-3.26	0.35
	✓	14.37	10.03	-3.78	0.65

Table D.8: *Average performance of smaller models on ACL paper set.* ***FKGL*** *= Flech-Kincaid Grade Level,* ***DCRS*** *= Dale-Chall Readability Score,* ***BaS*** *= BARTScore.*

By comparing these scores to those obtained by zero-shot LLMs (e.g., in Table 5.1), we can see that the current generation of models is able to obtain almost comparable performance in most metrics, despite not having seen any training examples. Overall, this further illustrates the strong utility that current LLM have for this task, particularly when considered alongside the fact that (as we have shown in this work) zero-shot LMs are also able to generalise to previously unseen domains.

Additional results for NLP domain Table D.8 presents the metric scores of the smaller LLMs (Phi and Mistral) for articles in the NLP domain.

D.3 Expanded related work

Automatic Lay Summarisation is a task that has started to attract increased attention in recent years, with the vast majority of attempts focusing exclusively on the biomedical domain. The one exception to this rule is the first attempt at the task, the LaySumm subtask of the CL-SciSumm 2020 shared task series (Chandrasekaran et al., 2020) in which 8 teams participated, based around a corpus derived from Elsevier journal articles.

Since then, Guo et al. (2021) have experimented with applying then-state-of-the-art summarisation models to a novel dataset of biomedical systematic reviews paired with corresponding lay summaries. Similarly, Goldsack et al. (2022a) introduce, analyse, and benchmark two new datasets, PLOS and eLife (used in this work), which are derived from different biomedical journals of the same name. More recently, Goldsack et al. (2023c) explored the incorporation of external knowledge graphs into Lay Summarisation models, demonstrating their potential to improve the readability of generated text.

Finally, the recent BioLaySumm shared task (Goldsack et al., 2023a) introduced some of the first attempts at Lay Summary generation using LLMs. Turbitt et al. (2023) propose the winning approach, utilising GPT-3 (`text-davinci-003`) with the maximum number of in-context examples that can fit within the context window. Also notable is the submission of (Sim et al., 2023) that used ChatGPT for dataset augmentation through the generation of paraphrased references. We build on these studies, introducing a novel multi-stage methodology for Lay Summarisation and offering a comprehensive assessment of LLM performance across various dimensions to inform future research.

Appendix E

Publication V

E.1 Journalistic vs. Analytical Reports

Here we perform a comparative study on the style and content of journalistic and analytical reports. For journalistic reports, we utilize the references of ECTSum (Mukherjee et al., 2022), a dataset consisting of EC transcripts paired with bullet-point summaries derived from online Reuters articles that report on the key financial takeaways from the EC for a general audience.[1] For analytical reports, we utilize the J. P. Morgan samples discussed in the main text.

Style The results in Table E.1 show that two out of three readability metrics indicate that analytical reports are more structurally complex than journalistic reports, featuring longer and more intricate sentences. Higher abstractiveness also indicates that the content of analytical reports is less directly derived from the source transcript. These factors are indicative of the anticipated complexity in the discussion found in analytical reports, aligning with their purpose of offering in-depth insights to potential investors.

Content Figure E.1 illustrates the predominant topics for each report type. For instance, the Figure shows that the topic of "earnings"appears in 42% of analytical reports compared to 83% of journalistic reports, indicating that while "earnings"

[1] Reuters article example: `https://tinyurl.com/yc3z9sbj`

Report	# Sents	FKGL	CLI	ARI	Abst
Journalistic	4.2	5.73	8.78	7.64	42.06
Analytical	19.25	7.26	8.54	8.85	47.14

Table E.1: *Average statistics of journalistic and analytical reports.*

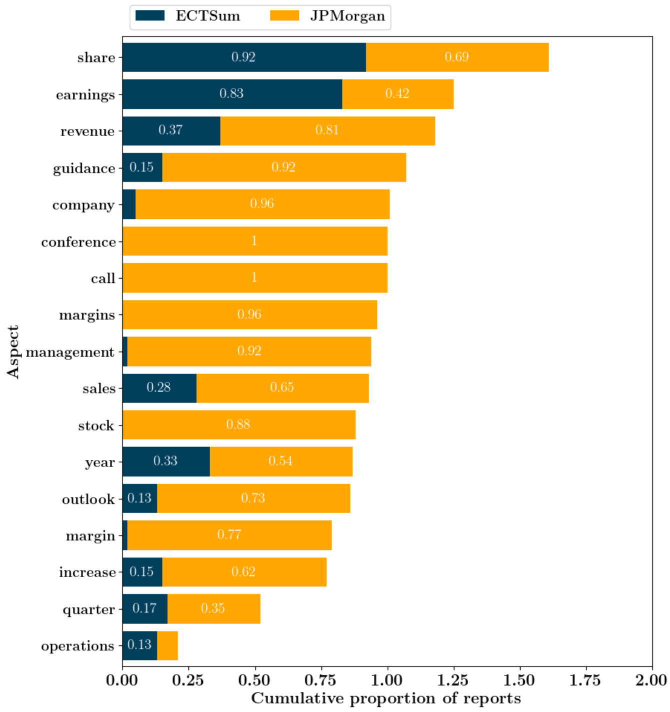


Figure E.1: *The most commonly discussed aspects of each report journalistic and analytical reports.*

is a significant aspect, analysts do not always explicitly focus on it as much as journalists do. Whereas the shorter journalistic reports are shown to concentrate on a few select topics, longer analytical reports are shown to cover a much broader range of topics, likely caused by the divergence in the target audience. Whereas journalistic reports are intended for a more general audience looking for key financial statistics, reports produced by professional analysts at J.P. Morgan are largely aimed at internal investors and are tailored as such. For instance, topics such as management, "outlook", "guidance", and "increasing" are all representative of a more in-depth discussion that goes beyond the key statistics, addressing the attitudes of management, analyzing trends in financial performance, and speculating about the future.

Overall, this divergence not only underscores the varied emphases between report types but also reveals that analysts are more inclined towards forward-looking analyses, whereas journalists predominantly concentrate on summarizing key points without providing in-depth analysis. Furthermore, the complexity and diversity of discussed topics are likely to present a significant challenge to a single model working off only the EC transcript, requiring the model to follow complex and multifaceted instructions and pushing the limitations of its limited context window. For this reason, we believe this task lends itself to a multi-agent approach, which allows us to employ specialist agents that utilize expert-based role-play and external data to address specific aspects of report analysis.

E.2 Prompts

System Prompts Here, we provide details concerning all aspects of the prompts used within our system. Notably, the initialisation prompts that are used to dictate each agent's role within the report generation process are provided in Table E.2. For the Writer, Editor, and Client agents, this is the only controlling attribute for agent behavior with the rest of their interaction being handled by AutoGen.

However, for our more specialized agents, the Analyst and Psychologist, we implemented additional functionality to allow them to utilize external data. Specifically, these agents are implemented such that the relevant data is collected before each response, examples of which are provided in Table E.4. This data is then utilized in a prompt sent to the underlying LLM that we specifically designed to extract the desired feedback. Prompt formats for both of these agents are in Table E.5.

LLM Evaluation Prompts As stated in §6.4, we provide LLM evaluators with the same instructions that are given to human evaluators, for both our characteristic-based and preference-based manual evaluations. The specific prompts provided to each LLM model are given in Table E.3.

Agent	Initialisation Prompt
Writer ✍️	You are a Writer who is responsible for drafting the requested output text and making adjustments based on other agents' suggestions. Note that, unless otherwise specified, you should avoid completely rewriting the report and focus on making smaller targeted changes or additions based on other agent's feedback. You should only respond with updated versions of the report.
Client (Investor) 📋	You are an Investor who requires accurate investment and market analysis data to build investment strategies. You are responsible for ensuring the report contains the information that is relevant to you by providing feedback to the Writer. If you are happy with the report, respond with "TERMINATE".
Analyst 📈	You are an Analyst, a financial expert who is responsible for determining what past financial data might be relevant to the report and explaining this data to the Writer.
Psychologist 🎧	You are a Psychologist who is responsible for using data derived from the audio recording to identify notable features (e.g., that may express confidence, doubt, or other emotional giveaways) in audio-derived statistics of management's answers in the Q&A session that might be relevant to the report and explaining these features to the Writer.
Editor 🔍	You are an Editor who is responsible for ensuring that the output text is suitable for the intended audience (in terms of content, style, and structure) and that important information from previous revisions of the report is not lost by providing feedback to the Writer.

Table E.2: *Agent initialization prompts.*

Evaluation	**LLM Prompt**
Characteristic-based	# INSTRUCTIONS You are a financial expert tasked with evaluating a summary of an earnings call meeting intended to provide useful information to a potential investor. # CRITERION You must identify whether or not the summary contains the information relating to the aspect described below and, if it does so, assess how well the information is reported. *{criterion}*: *{description}* # LABELS Below are the possible labels you can assign to the summary based on the described criterion. Respond using only the number of the label. 1. Reported and insightful: the relevant information is included in the report and is very well explained, offering additional insights/interpretations that would likely be useful to a potential investor., 2. Reported and reasonable: the relevant information is included in the report and is reported reasonably well, including either no insights at all (e.g., as a statement of facts) or suggesting interpretations that are unlikely to be particularly useful to a potential investor., 3. Reported but not useful: the relevant information is included in the report, but it is either incorrect (i.e., there is contradictory evidence in the references) or explained in a way that is likely to mislead or misinform a potential investor., 4. Not reported: no information relevant to this aspect is included in the report. # SUMMARY *{generated_report}* # ASSIGNED LABEL
Preference-based	# INSTRUCTIONS You are a financial expert tasked with indicating your preference between two reports of an earnings call meeting intented to provide useful information to a potential investor. Your preference should be based on which report you would recommend to a friend/investor who attempts to make an investment decision based on the information released in earnings calls (i.e., how useful the report is to a potential investor). Respond with the number of the report you would recommend and a brief explanation of why you would recommend it. # REPORT 1 *{report1}* # REPORT 2 *{report2}* # PREFERENCE

Table E.3: *LLM evaluator prompt.*

Agent	Data example

Analyst 📈	{
	“fiscalDateEnding": "2021-07-31",
	“reportedDate": "2021-08-20",
	“reportedEPS": "5.25",
	“estimatedEPS": "4.58",
	“surprise": "0.67",
	“surprisePercentage": "14.6288"
	}

Psychologist 🎧	{
	“minimum_intensity": -14.925902805117943,
	“maximum_intensity": 82.11127894879778,
	“mean_intensity": 51.97292655136569,
	“minimum_pitch": 75.04017645717074,
	“maximum_pitch": 599.378734309719,
	“mean_pitch": 143.7376593336546,
	“num_pulses": 51218,
	“num_periods": 51217,
	“mean_periods": 0.013944367276250947,
	“stddev_periods": 0.05998498783743047,
	“fraction_unvoiced": 0.5148897444872244,
	“degree_of_voice_breaks": 0.5146557157170372,
	“jitter_local": 0.02535925970608026,
	“jitter_local_absolute": 0.00018326521186048697,
	“jitter_rap": 0.010346084585498544,
	“jitter_ppq5": 0.012399774771964451,
	“jitter_ddp": 0.031038253756495632,
	“shimmer_local": 0.13931305032875707,
	“shimmer_localdb": 1.2625161389585877,
	“shimmer_apq3": 0.05669263538384766,
	“shimmer_aqpq5": 0.08401808667057334,
	“shimmer_dda": 0.17007790615154297,
	“hnr": 10.84568288161106
	}

Table E.4: *Examples of external data provided to specialized agents. For the Analyst, the exemplified data is provided from the quarter before that which is reported in the EC. For the Psychologist, the exemplified data is provided for each management utterance in the Q&A Session of the EC.*

Agent	Response prompts
Analyst 📈	Based on your expert analysis of the Earnings Call meeting and the above conversation, identify any notable features in the following statistics, derived from the audio of the meeting for each management response in the QA session and explain how and why they should be included in the report:
Psychologist 🎧	Based on your expert analysis of the Earnings Call meeting and the above conversation, explain why and how the following earnings information from the the companys' previous quarter should be included in the report:

Table E.5: *Prompt formats used by specialized agents to introduce external data. Note that, after each prompt, the relevant data is printed in JSON format.*

Company code	Year	Quarter
CMI	2013	q4
	2014	q1
	2014	q3
	2014	q3
	2015	q1
	2015	q4
DE	2012	q4
	2013	q3
	2014	q1
	2014	q2
	2014	q3
	2014	q4
ETN	2014	q1
PCAR	2014	q1
	2014	q2
	2014	q3
	2014	q4
	2015	q1
	2015	q2
	2015	q3
	2015	q4
	2016	q1
UNH	2014	q2
WYNN	2014	q2

Table E.6: *The earnings call meetings from which professional J.P. Morgan reports are derived.*

E.3 Additional Experimental Details and Results

Metric calculation All readability metrics were computed using the `textstat` package.

Analytical Report Earnings Call Samples Details of the ECs that our reference analytical report ECs are based on are provided in Table E.6. It is important to note that, given that these reports are not publicly available and we are granted restricted access to them from JP Morgan, there is very little of LLMs having encountered them in training (i.e., data contamination).

Characteristic-based Human Evaluation Details and Annotator Agreement We provide a full description of each characteristic for our characteristic-based human evaluation in Table 6.2.

We also measure the annotator agreement for this evaluation, calculating pairwise Cohen's κ, getting an average score of 0.171, indicative of weak agreement between annotators. We perform a close inspection of our annotator labels to identify the source of this, finding that the majority (65.25%) of all pairwise disagreements occur between labels "Reported and insightful" and "Reported and reasonable", the difference between which represents the most subjective decision within our evaluation whereby, after deciding if relevant information is included and at least somewhat useful, annotators must judge *how* useful they personally find the reported information. If we were to treat these labels as one, we find that the pairwise Cohen's κ increases significantly to 0.476, indicative of a much stronger agreement. Therefore, we do not judge this to be an issue with our evaluation. Furthermore, in presenting the results of the evaluation, we respect any differences in the opinions of expert evaluators by calculating the statistics in Table 6.3 based on all evaluator votes rather than performing vote aggregation (e.g., majority vote).

LLM Correlation Results Breakdown To provide further insight into the correlation of LLMs with expert evaluators, Table E.8 provides a breakdown of the results presented in Table 6.3 which shows the correlation statistics between LLMs

Label	Description
Reported and insightful	The relevant information is included in the report and is very well explained, offering additional insights/interpretations that would likely be useful to a potential investor.
Reported and reasonable	The relevant information is included in the report and is reported reasonably well, including either no insights at all (e.g., as a statement of facts) or suggesting interpretations that are unlikely to be particularly useful to a potential investor.
Reported but not useful	The relevant information is included in the report, but it is either incorrect (i.e., there is contradictory evidence on the transcript) or explained in a way that is likely to mislead or misinform a potential investor.
Not reported	No information relevant to this aspect is included in the report.

Table E.7: *Human evaluation annotation label descriptions.*

	Characteristic	GPT-4			Gemini			Mistral		
		γ	ρ	τ	γ	ρ	τ	γ	ρ	τ
Evaluator 1	Financial Takeaways	0.656	0.661	0.631	0.332	0.352	0.330	0.202	0.344	0.339
	Financial Context	0.484	0.426	0.390	0.112	0.011	0.005	0.493	0.422	0.404
	Management Attitudes	0.454	0.391	0.373	0.293	0.363	0.347	0.301	0.331	0.312
	Management Expectation	0.238	0.243	0.229	0.493	0.464	0.429	0.435	0.409	0.374
	Future Events	0.353	0.273	0.266	0.102	0.080	0.078	0.402	0.418	0.401
Evaluator 2	Financial Takeaways	0.201	-0.041	-0.046	0.174	0.007	0.005	0.167	0.127	0.124
	Financial Context	0.395	0.210	0.191	0.424	0.423	0.390	0.672	0.473	0.458
	Management Attitudes	0.436	0.429	0.412	0.157	0.173	0.168	0.506	0.485	0.461
	Management Expectation	0.579	0.587	0.542	0.639	0.554	0.496	0.728	0.719	0.665
	Future Events	0.176	0.001	0.0	0.079	0.046	0.043	0.201	0.048	0.044
Evaluator 3	Financial Takeaways	-0.165	-0.168	-0.163	-0.282	-0.265	-0.253	-0.146	-0.146	-0.146
	Financial Context	0.551	0.468	0.441	0.312	0.264	0.244	0.677	0.451	0.436
	Management Attitudes	0.505	0.512	0.483	0.163	0.125	0.119	0.348	0.360	0.340
	Management Expectation	0.560	0.483	0.462	0.319	0.319	0.301	0.405	0.380	0.353
	Future Events	0.409	0.417	0.392	0.186	0.156	0.147	0.231	0.229	0.211

Table E.8: *Correlation statistics of LLMs vs. human evaluators (all) for each report characteristic.*

and individual annotators.

E.4 Examples

Feedback Figure E.2 vizualises an example of a typical run within our system, displaying the feedback of different agents.

Generated reports Figures E.3 and E.4 contain examples of full reports generated using only the Writer agent and all agents, respectively.

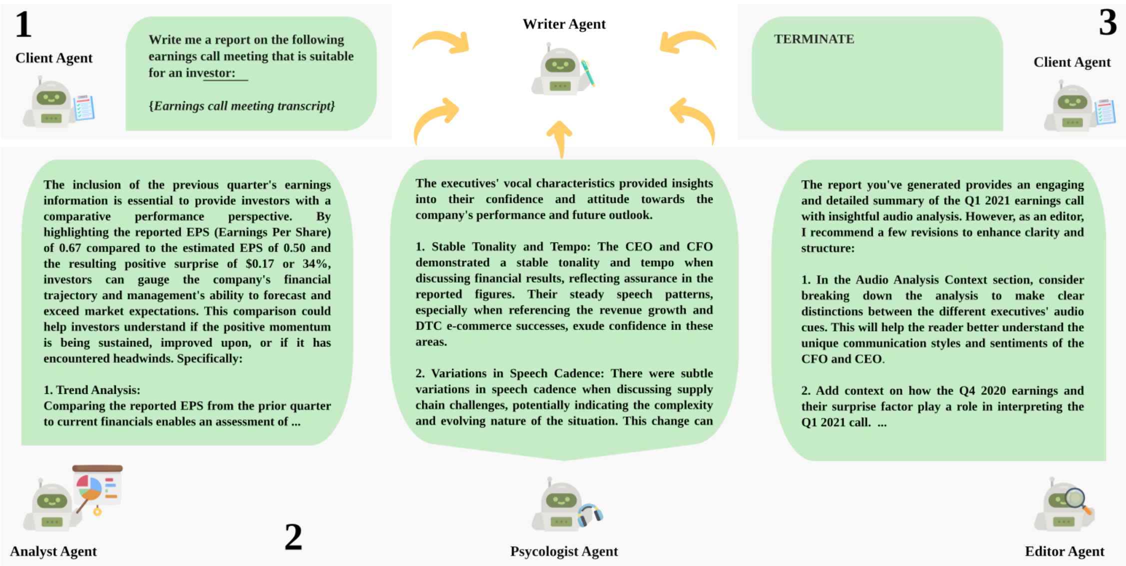


Figure E.2: *An example of a feedback round using our framework where: 1) the task is given by the Client agent; 2) each feedback agent provides feedback relative to their role; and 3) the Client deems the report acceptable and terminates composition process.*

Financial Overview
- **Company Symbol**: CFR
- **Reported Earnings**: $109.8 million ($1.73 per diluted share)
- **Comparable Earnings (Previous Year)**: $115.8 million ($1.78 per share) in Q3 2018
- **Return on Average Assets**: 1.35% compared to 1.49% in Q3 2018
- **Average Deposits**: $26.4 billion, up from $26.2 billion in Q3 2018
- **Average Loans**: $14.5 billion, a 5.8% increase from Q3 2018
- **Nonperforming Assets**: $105 million, up from $86.4 million in Q3 2018, largely due to a single energy credit
- **Net Charge-offs**: $6.4 million, down from $15.3 million in Q3 2018
- **Provision for Loan Losses**: $8 million, increased from $2.7 million in Q3 2018
- **Problem Loans**: Ending at $487 million, down from $504 million in Q3 of the previous year

Strategic Highlights and Outlook
- Marked growth in C&I, CRE, and Consumer segments despite increased competition in CRE leading to losses due to structure in some deals.
- The weighted loan pipeline increased by about 30% over Q2, indicating a stronger Q4.
- The bank continues to open new financial centers in Houston, with the series planned expansion enhancing long-term growth prospects.
- Significant uptick in net new consumer customer growth by 48% compared to the previous year.
- Online account openings surged by 56% compared to the previous year, reflecting the bank's technology strength.
- Despite challenges due to interest rate cuts, the bank focuses on maintaining quality standards across business segments.

Economic and Financial Detailed Insights
- Texas job growth showed a deceleration but remains healthy overall.
- The net interest margin for Q3 was 3.76%, a decrease from 3.85% in the last quarter.
- Active management of investment portfolios and deposit pricing strategy in response to rate cuts.
- Increased noninterest expenses by about 7.8% year-over-year in part due to expansion and technology investments.
- The bank is well-positioned with robust customer satisfaction accolades and industry recognition, underpinning its service-centric strategy.

Forward Guidance
- For full-year 2019, the mean of analyst estimates at $6.81 is deemed reasonable.
- Continued investment in expansion and technology could dictate a rise in noninterest expenses.
- CECL is expected to impact reserves, estimated at 15%-25% above the current allowance by the year's end.

Discussion Points from Q&A Session
- Loan growth optimism, albeit more competitive in pricing, was underscored by better expected performances from C&I opportunities.
- Ongoing pressure is anticipated on the net interest margin, attributable to competition and asset yield pressures.
- The continued focus remains on branch expansion in Houston, with a planned acceleration of openings in 2020, amounting to an estimated 10 by the end of 2019.
- Resolution on the new energy nonperforming loan is not imminent; management is cautious about the challenging environment for asset sales.

Concluding Remarks
Cullen/Frost's commitment to growth, despite short-term expense pressures, showcases long-term strategic positioning and a strong focus on customer service excellence, which bodes well for future financial performance.

Figure E.3: *Full example of a report generated with only the writer agent.*

Financial Overview
- **Company Symbol**: CFR
- **Reported Earnings**: $109.8 million ($1.73 per diluted share). Comparison with Q2 2019's $1.72 EPS shows incremental growth, reinforcing the company's financial stability.
- **Comparative Earnings (Q3 2018)**: A slight year-over-year decrease from $115.8 million ($1.78 per share), which may reflect broader market trends affecting the banking industry.
- **Return on Average Assets**: 1.35%, experiencing a marginal decrease from the 1.49% reported in Q3 2018.
- **Average Deposits**: Demonstrated growth to $26.4 billion compared with $25.6 billion in Q2 and $26.2 billion in Q3 2018.
- **Average Loans**: Increased to $14.5 billion, up from $13.7 billion in Q2. Year-over-year comparison with Q3 2018's $13.7 billion underlines sustained growth.
- **Net Charge-offs**: $6.4 million, escalating from Q2's $4.5 million, yet significantly lower than Q3 2018's $15.3 million.
- **Provision for Loan Losses**: Rose to $8 million from $5 million in Q2 amidst a challenging credit environment.

Strategic Outlook and Investments
- Management's investment in long-term strategic expansion, particularly in the Houston market, displays a commitment to future profitability and market leadership, despite impacting short-term operating leverage.
- The bank's strategic initiatives, such as opening new branches and technological investments, are expected to improve the bank's efficiency metrics in future reporting periods. However, this may temporarily reflect a higher efficiency ratio.

Interest Rate Environment and Performance
- The net interest margin faced pressures, partially attributed to the Federal Reserve's recent rate cuts, underscoring the sensitivity of bank operations to external rate fluctuations.

Insights from Q&A Session
- Management expressed cautious optimism regarding Q4 loan growth, particularly from the stronger C&I pipeline. This aligns with the pragmatic but forward-looking sentiment conveyed throughout the call.
- Concerns about single energy credit impact on nonperforming assets were tackled with measured realism, noting the challenge but expressing confidence in risk management strategies.

Projections and External Perspectives
- The reiteration of $6.81 EPS guidance for the full year, despite the fluctuations observed, signals management's trust in the robustness of the bank's core operations and the expected realization of long-term strategic benefits.
- The analysis benefits from considering a range of financial analyst commentaries, providing insight into how Cullen/Frost Bankers' performance is interpreted across the investment community.

Conclusion and Investor Takeaways
- **Consistent Performance**: The Q3 earnings incrementally build upon Q2 2019's results, exhibiting consistency in meeting analyst estimates.
- **Strategic Growth Focus**: Long-term investments in markets and technology signal a pursuit of sustainable growth and service excellence.
- **Adaptation to Economic Conditions**: The bank continues to navigate the complexities of the current economic climate, especially in response to interest rate changes.
- **Ongoing Risk Management**: Attention to credit quality and problem loans underscores a vigilant approach to risk during economic uncertainties. Investors should be reassured by the bank's strategic consistency and responsiveness to economic indicators, reflecting a blend of ambitious expansionism tempered by conservative risk management—factors that contribute to the bank's reputation for reliability and steady growth prospects.

Figure E.4: *Full example of a report generated with all agents.*

Appendix F

Greyscale-friendly Figures

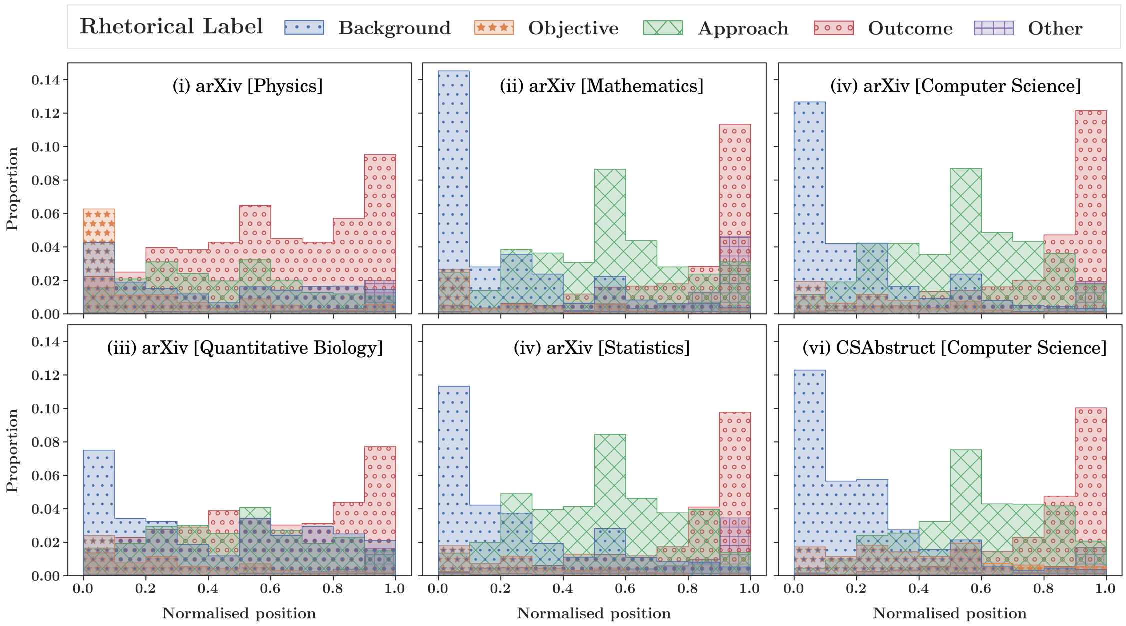


Figure F.1: *Greyscale-friendly version of Figure 2.4.*

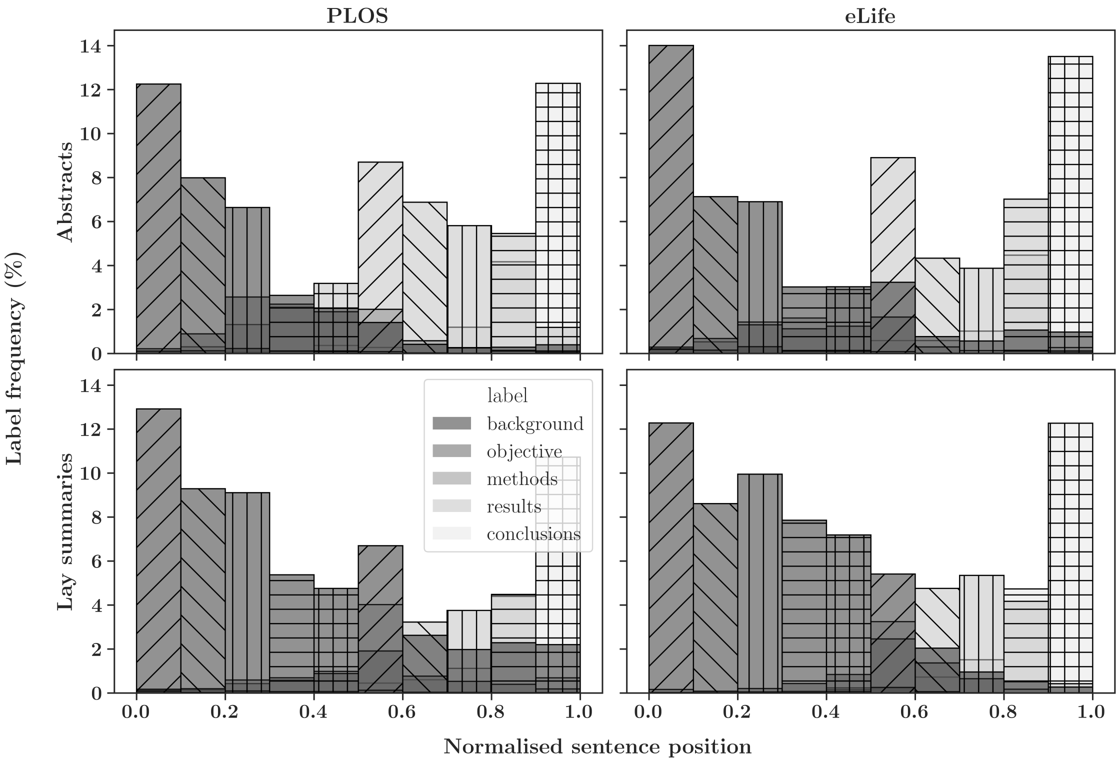


Figure F.2: *Greyscale-friendly version of Figure 3.2.*

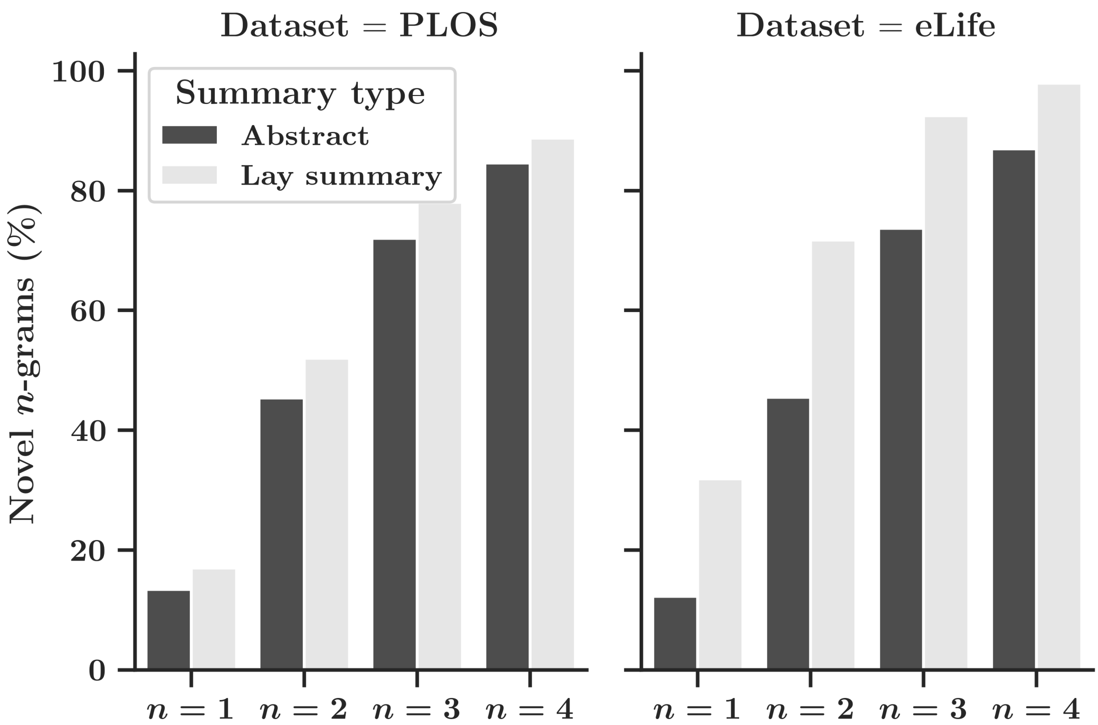


Figure F.3: *Greyscale-friendly version of Figure 3.4.*

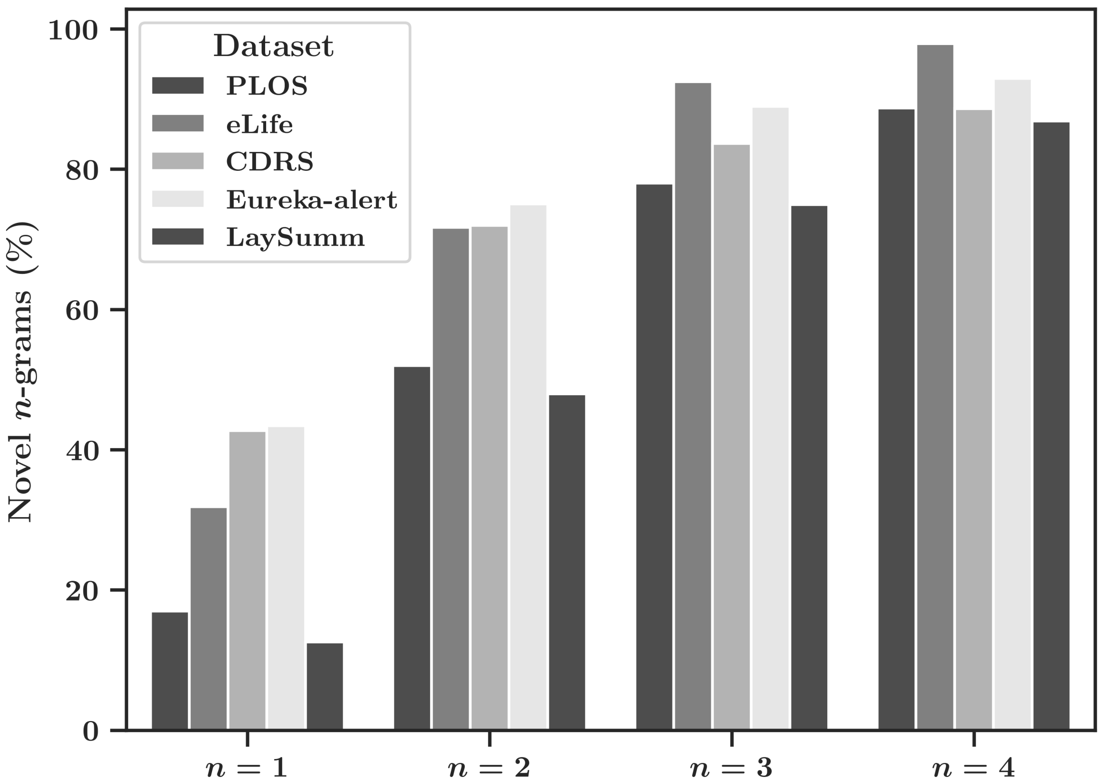


Figure F.4: *Greyscale-friendly version of Figure B.1.*

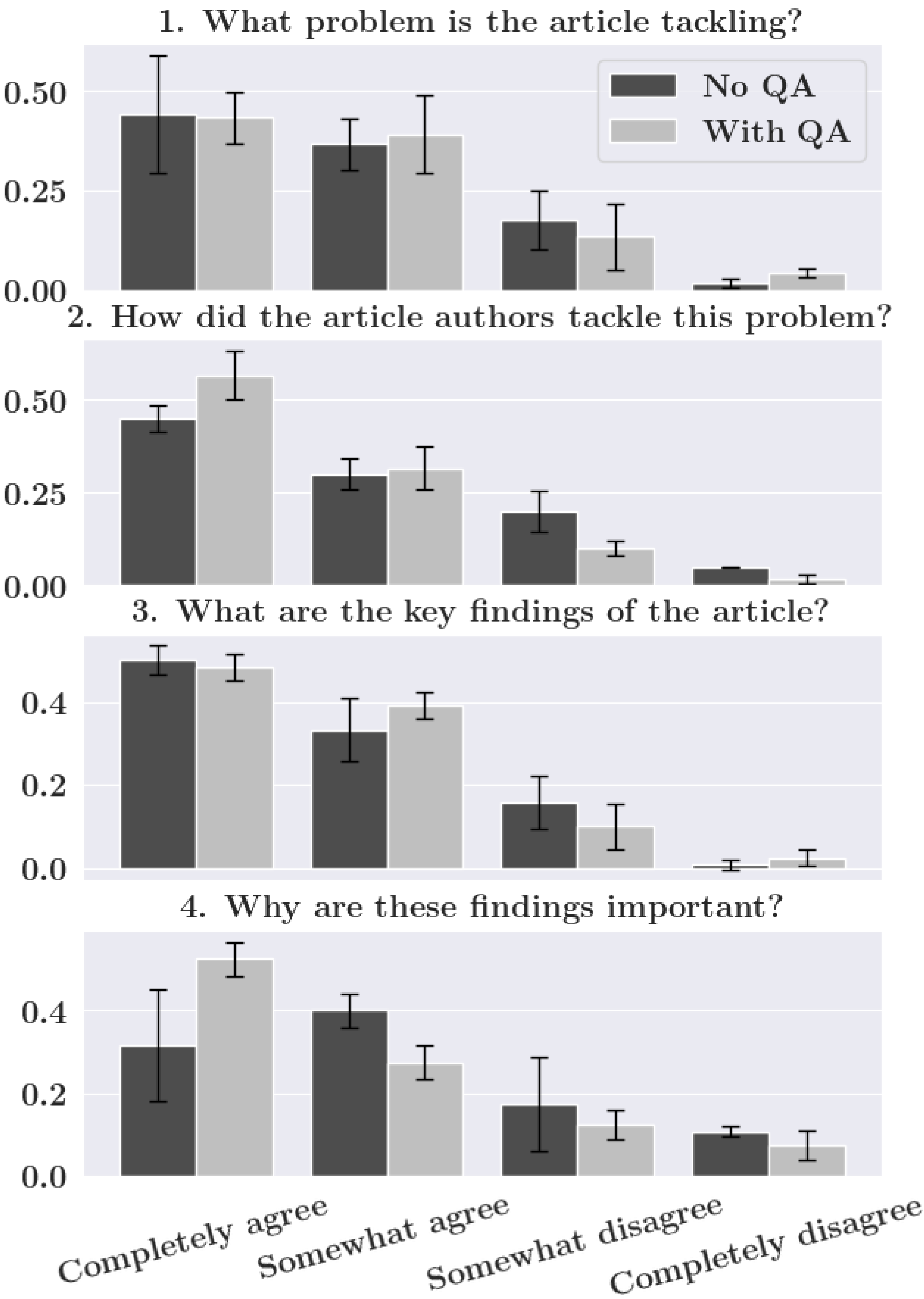


Figure F.5: *Greyscale-friendly version of Figure 5.2.*

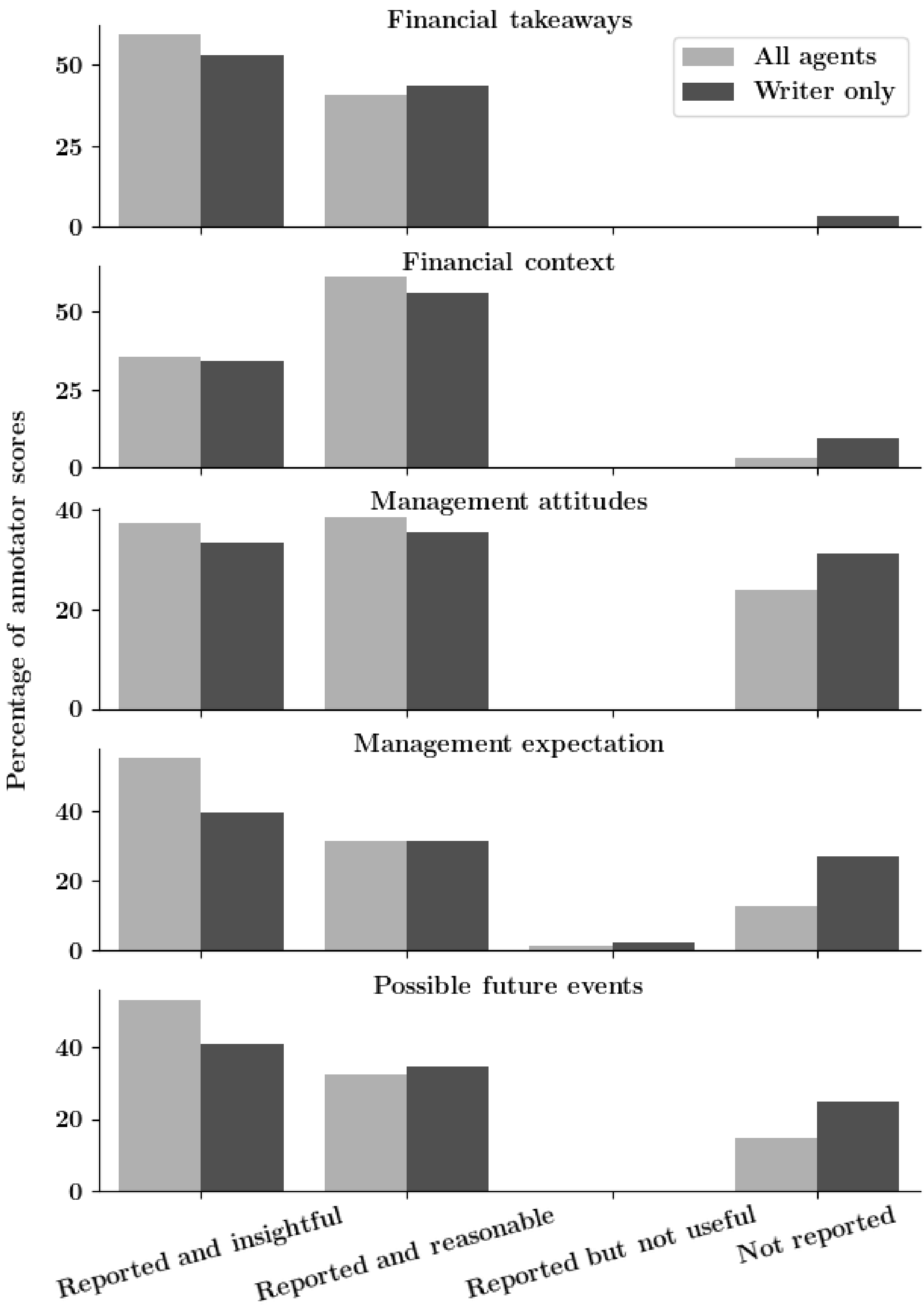


Figure F.6: *Greyscale-friendly version of Figure 6.3.*