Controlling Context Factors in Abstractive Summarization of Long Documents

Marcio Fonseca

Doctor of Philosophy

Institute for Language, Cognition and Computation
School of Informatics
University of Edinburgh

2024

Abstract

The massive influx of textual data poses a significant challenge in technical fields, fueling the research of text summarization systems. Through innovative approaches in representation learning and extensive data utilization, these systems have demonstrated remarkable advancements, particularly within domain-specific contexts. More recently, large language models (LLMs) such as *ChatGPT* demonstrated an impressive ability to generate abstractive summaries that are fluent and relevant according to human judgments, even without domain-specific training. While those models are regarded as strong general-purpose summarizers, technical documents require more nuanced control of *contextual factors* that depend on the target audience and task goals.

In this thesis, we argue that integrating contextual factors that are not easily distilled from reference summaries is crucial for advancing in summarization of long technical documents. We establish a conceptual framework separating *intrinsic factors* that can be determined from document-summary pairs (e.g., redundancy and relevance) and *extrinsic factors* (e.g., conciseness and rhetoric) that depend on the task context and subjective intentionality. Guided by this framework, we approach the summarization problem as a factorized energy-based model, in which we optimize for intrinsic and extrinsic factors separately. Our model, FACTORSUM, achieves significant improvements in terms of lexical alignment to reference summaries while requiring modest compute resources compared to baselines.

Furthermore, we delve into the application of large language models to three types of scientific summarization tasks: abstract generation, summarization for reviews, and lay summarization. Our results show that those LLMs excel at the controllability of stylistic features such as budget and narrative perspective. However, these models exhibit gaps in the understanding of domain concepts in scientific papers, which limits more fine-grained control. Finally, we also propose an approach to improve the lexical alignment of summaries guiding LLM summarizers with keywords derived from FACTORSUM, thus combining the strengths of both approaches.

In conclusion, our investigation confirms that large language models are powerful tools for summarization tasks, occasionally eclipsing human-authored summaries according to expert judgments. However, we find that LLMs struggle to match the richness of human perspectives in lay summarization, for instance. Our factorized modeling approach partially addresses these limitations, and hopefully, inspires future work focusing on context-aware summarization.

Acknowledgements

In some ways, a PhD is like sailing. There is something that pushes you to explore uncharted waters despite the mental and technical hurdles. It takes time to adapt to the restricted space and adjust the sails, but you eventually find the proper course to the destination. However, things are not always hunky-dory. The conditions may change drastically, and if you do not act fast, you capsize! And to put it mildly, the waters of the Firth of Forth are *freezing*! You feel scared, high and dry... but this is when fellow sailors extend their hands and help you set sail again. And now that I can see the first glimpse of distant land, I would like to acknowledge those people.

First, I extend my heartfelt gratitude to my fellow crew members of The Cohort. Especially to Shay Cohen for his unwavering support and invaluable insights, and for trusting my work. Also, I would like to thank Zheng Zhao and Ronald Cardenas for their camaraderie, our sessions over pizzas, and for helping me with the summary evaluation experiments. To Yftah Zizer, your engaging discussions and collaboration were instrumental in the realization of my first publication. During my academic journey, I also received feedback from Mirella Lapata and Ivan Titov, whose insights were important in shaping this thesis. Also, I would like to thank the thesis examiners Edoardo Ponti and Jackie Cheung for providing excellent suggestions to improve this work.

And to our sponsor Actelligent, thanks for making this project possible. Our discussions and experiments with the Actelligent team led to further development of the concepts and the model introduced in Chapter 3, and to the eventual successful adaptation to the summarization of financial documents. These novel developments executed by their technical team are powering exciting new financial services that provide summaries of financial news articles and corporate reports.

On my way back to Brazil, I have to moor in the port of Lisbon and thank my cousin Dênis Rubra and his wife Juliana (and the cats) for having me as a guest and for being such great friends. In the port of Rio de Janeiro, I have many special people to thank. First, my parents Mario and Luzenir, and my brother Eduardo, who are always there for me, and always have a smile on their faces despite all the challenges. After all, *o vento que venta lá, venta cá*. Also, to my friend Daniel Andrade who gave me support during difficult moments. As did my friend Bruno Vergueiro, who managed to visit me and endure the cold weather like a true Viking! In the (imaginary) port of Brasília, likely my final destination, I would like to thank Patricia Almeida, Michael Onishi, and Eduardo Andrade. They gave me a lot of encouragement to pursue this program and are always checking in to make sure I keep my sanity. Thank you!

To a special person, Stephanie, I am profoundly grateful for everything you did for me. Our many happy moments will be fondly remembered.

Land Ho!

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Marcio Fonseca*)

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

> Amidst the vast expanse of text lies the essence of knowledge, waiting to be distilled into clarity. In the pursuit of understanding, we navigate the labyrinth of language, seeking paths that lead to insight. Through innovation and ingenuity, we strive to unveil the treasures hidden within the boundless pages of discourse.
>
> ChatGPT

Recent developments in natural language processing (NLP) led to unprecedented and irreversible changes in the relationship between humans and machines. Surfing the trends of scaling laws, computational artifacts known as language models enlarged 3 orders of magnitude in 5 years (Zhao et al., 2023) and escaped the constraints of researchers' labs to become consumer-facing products. These commoditized large language models (LLMs) now help millions of people in creative tasks such as drafting articles[1], suggesting recipes for a vegan breakfast, or writing computer programs[2].

Among those language generation tasks, text summarization is arguably one of the most impacted by LLMs. For the first time, users have access to general-purpose models that output coherent and fluent summaries of news articles (Goyal et al., 2022), sci-

[1]The epigraph of this chapter was generated by OpenAI's `gpt-3.5-turbo-0613` on 22 February 2024, using the following prompt: *I am writing a PhD thesis that proposes new approaches to text summarization involving long documents. Please provide an inspirational epigraph to use at the beginning of the introduction chapter.*

[2]https://github.com/features/copilot

entific papers (Fonseca and Cohen, 2024), and books (Chang et al., 2024). According to recent benchmarks (Zhang et al., 2024), the quality of LLM-generated summaries is even superior to human-written reference summaries used in model evaluation, leading to claims that summarization might be "almost dead" (Pu et al., 2023).

If it is the case that summarization is mostly solved, what are the remaining challenges keeping it alive as a research area? To clarify this question, we revisit a central (and elusive) concept for summarization: *importance*. According to Jones (1998), a summary can be defined as "a reductive transformation of source text to summary text through content reduction by selection and/or generalisation on what is *important*" [emphasis added]. While intuitive, this definition is unsatisfyingly circular, leaving the crux of the problem — the criteria for important content — unspecified. Instead, Jones (1998) focus on relevant *context factors* such as purpose and audience, without specifying the role they play in the selection of important information.

Recently, Peyrard (2018) attempted to address this gap by proposing an information-theoretic framework that decomposes importance into three factors: *redundancy*, *relevance*, and *informativeness*. Redundancy is a property of the summary alone, and a summary should be constructed such that it maximizes the entropy over semantic units, or equivalently, minimizes the redundancy of information. In contrast, relevance refers to the coverage of salient semantic units in a way that approximates their distribution in the source document. Finally, informativeness measures the coverage of novel information given background knowledge, that is, a summarization context. Thus, importance can be viewed as an optimization objective that maximizes relevance and informativeness while minimizing redundancy[3].

Generating summaries with high importance in an information-theoretic sense is of course not sufficient. Good summaries must also be coherent, fluent, and grammatical, which is a goal of most natural language generation systems. Putting all those elements together, one can generalize Peyrard (2018) and (Jones, 1998) ideas by stating that the quality of a summary is perceived via 3 dimensions (organized in Figure 1.1): *linguistic*, *intrinsic importance*, and *extrinsic importance*. Linguistic quality relates to properties that are desirable to most language generation tasks, including redundancy, fluency, and coherence. Intrinsic importance refers to judgments of content salience that depend on the source document alone, independently of any external factors. In

[3]The importance theory proposed by Peyrard (2018) has limitations, especially in the choice of semantic units, which require better formalization. However, we find that the main importance components – *redundancy*, *relevance*, and *informativeness* – and their relationship with external factors serve as a useful framework to inspire and conceptualize the motivation of this thesis.

contrast, extrinsic importance takes into account contextual and subjective factors such as summary audience and goals.

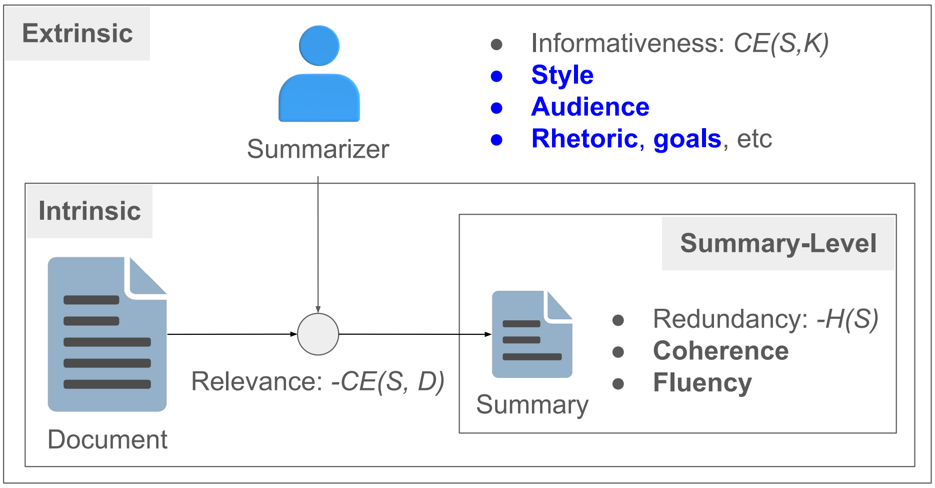

Figure 1.1: Illustration of summary quality factors classified into summary-level (linguistic), intrinsic, and extrinsic dimensions. Redundancy, Relevance, and Informativeness are concepts from the importance theory by Peyrard (2018). CE denotes the cross-entropy function, D is document input for summarization, and K is a external knowledge. This thesis focus on the controllability of properties related to **extrinsic (or contextual) factors**, including *style* and *concept coverage*.

Breaking down summary quality into those dimensions also helps to characterize the recent progress of text summarization systems, which can be divided into *pre-GPT* and *GPT phases*[4]. In the *pre-GPT phase*, the surge of deep-learning-based language models (See et al., 2017; Zhang et al., 2020) pushed summarization progress towards matching token-level statistics of reference summaries. Those systems moved away from explicit models of importance in favor of implicit, end-to-end optimization approaches that tackle the *relevance* factor proposed by Peyrard (2018). In this period, the increase in model, dataset, and input context sizes (Zhang et al. 2020; Zaheer et al. 2020; Beltagy et al. 2020, *inter alia*) resulted in better alignment with reference summaries as measured by metrics such as ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2019). However, this paradigm quickly showed its limits as collecting high-quality reference summaries is costly in most domains. Also, evaluation results suggested that high ROUGE scores do not correlate with human perception of quality (Stiennon et al., 2020).

[4]We consider that GPT-2 (Radford et al., 2019) marks the start of the *GPT phase*, when language models demonstrate capacity to solve some NLP tasks without supervised fine-tuning.

The *GPT phase* is characterized by a departure from the task-specific training paradigm. Large decoder-only models and training datasets continue to scale, leading to a remarkable finding: language models can engage in question answering, summarization, and translation (to some extent) without being explicitly trained on those tasks (Radford et al., 2019; Brown et al., 2020). Instead, these tasks can be specified via *prompts* like `TL;DR`, which hints the task is to summarize the content before the prompt. These findings imply that language models can learn about *relevance* solely from the unsupervised language modeling objective, although not as efficiently as the supervised counterparts. Importantly, GPT-like models can generate summaries that approach human-level linguistic quality (Zhang et al., 2024), thus addressing the *redundancy*, *fluency*, and *coherence* factors. Those factors combined with the high relevance of generated content make large language models strong general-purpose summarization systems, as they are competent in all linguistic and intrinsic factors shown in Figure 1.1.

While the competence of large language models as summarizers for wide-audience content is well-studied (Goyal et al., 2022), the summarization of long technical documents is not sufficiently explored in the literature. In contrast to most news articles, technical documents cover several aspects or rhetorical zones (Teufel and Moens, 2002), which can have different relevance according to the task context. For instance, when writing an abstract of a paper, authors will emphasize aspects of the methods and results that they judge would persuade readers about their research contributions. On the other hand, different paper reviewers can make arbitrary decisions about the coverage of concepts and the relationship to previous work, depending on how the paper content resonates with their technical background. In Figure 1.2, we provide an example of such diverse concept coverage between two human-written summaries.

The example above illustrates the effects of variability of communicative intentions in the summarization process. Achieving such rich text production would require that LLMs capture not only the concepts necessary for producing high-quality summaries but also infer *beliefs*, *intentions*, and other internal states of the agents that produced the summaries. Recently, Andreas (2022) argued that the observation of a *sentiment neuron* in product reviews evidence that language models can infer the latent positive/negative stance of the reviewer. Similarly, we pose the question of whether LLMs consistently capture *summmarization intentionality*, which includes decisions about conciseness, narrative style, concept coverage, and other *context factors*. To address this question, we propose methods to control and evaluate intention alignment in

Article title: Detecting Hallucinated Content in Conditional Neural Sequence Generation

SUMMARY A: **[MOTIVATION]** The paper addresses the problem of ”hallucinated” content in conditional neural generation for two specific tasks: machine translation and summarization. **[METHOD]** It proposes a new task for faithfulness assessment, which classifies each token as either hallucinated or not. **[METHOD]** The classifier uses a pre-trained LM (either XLM-R or ROBERTa) and is fine-tuned on synthetic classification data created using both ’noisified’ real data and a pretrained LM (BART). **[RESULT]** Experiments on either summarization and MT system outputs labeled for hallucinations show relatively encouraging classification results (e.g., F1 of 0.46 to 0.66 for MT, and 0.56 to 0.66 for summarization).

SUMMARY B: **[MOTIVATION]** This paper proposes hallucination detection at the token level, which predicts if each token in the generation output is hallucinated or faithful to the source input. **[BACKGROUND]** In contrast, previous studies usually work on the sentence level. **[METHOD]** To create synthetic training data, a denoising pre-trained LM is first used to generate (potentially) unfaithful counterparts T’ of the references T. **[METHOD]** Then, token-level labels are obtained by comparing T and T’ via edit distance. **[METHOD]** Finally, a standard classification model is trained on the token-level labels by concatenating the source S, true and unfaithful targets T (T’). **[RESULT]**

Figure 1.2: Two human-written reviews for the same article from the MuP dataset (Cohan et al., 2022), showing different coverage of scientific concepts (annotated in bold). While summarizer A omits background information, summarizer B chooses to contrast with previous work (blue) and to omit results (red). This example illustrates the challenges of subjectivity in summarization and motivates the modeling of context factors (Jones, 1998).

several scientific summarization settings, leading to competitive results compared to fine-tuned models.

1.1 Thesis Statement

In this thesis, we argue the next challenge in summarization lies in *context factors* (Jones, 1998), that is, elements that are *extrinsic* to the input document and require a deeper understanding of the task context. These factors are particularly relevant for summarizing long and technical documents that can be interpreted from different perspectives according to the summarizer’s background and goals. Firstly, we propose methods to introduce extrinsic context factors in summarization pipelines, leading to improvement in summaries of long technical documents. Furthermore, we provide evidence that current summarization systems based on large language models approach human-level performance on most linguistic and intrinsic quality factors such as fluency, redundancy, and relevance, turning existing evaluation protocols largely non-informative. Finally, we propose controllability metrics as an alternative way to measure summarization progress and present methods to guide summarization according to different task goals and contexts. The thesis contributions include:

- A formalization of summarization as an energy-based model that factorizes intrinsic and extrinsic importance factors into separate optimization problems. Based on our factorized formulation, we introduce FACTORSUM, a summarization model that generates summary snippets (summary views) that can be globally optimized according to extrinsic factors such as summary budget and content coverage. This model achieves competitive performance across different benchmarks for scientific (Cohan et al., 2018) and legal summarization (Malik et al., 2024), while using about 4 times fewer parameters compared to existing baselines (Zhang et al., 2020). Furthermore, the results suggest a promising versatility in domain adaptation settings, with lexical alignment scores comparable to models trained on in-domain data.

- A comprehensive evaluation of large language model summarizers in diverse scientific summarization tasks, including abstract generation (Cohan et al., 2022), peer review summarization (Cohan et al., 2022), and lay summarization (Goldsack et al., 2022). we propose control metrics and prompts covering key stylistic factors (length and narrative perspective) and keyword coverage. For abstract generation and review summarization, the results reveal lexical alignment to reference summaries (measured by ROUGE) comparable to strong supervised baselines, while achieving remarkable controllability of stylistic features. Finally, the human evaluation results indicate a strong preference for machine-generated summaries, mainly due to their more comprehensive coverage of scientific concepts. These results confirm and expand previous studies based on news articles (Goyal et al., 2022; Zhang et al., 2024).

- A method to improve the lexical alignment of LLM summarizers via keyword guidance. Summaries from strong supervised models are leveraged to generate keywords for summarization prompts. Furthermore, we show how to improve the recall of keywords in summaries using classifier-free guidance (CFG; Ho and

Salimans 2022; Sanchez et al. 2023). The approach combines the high linguistic quality of LLM generations with the high lexical alignment of FactorSum.

- A *annotate-plan-edit* (APE) pipeline that allows per-concept control of scientific concepts in summaries. Using large language models and annotators/editors, we demonstrate the effectiveness of the method on scientific lay summarization (Goldsack et al., 2022), by emphasizing background content in summaries. The investigation also highlights differences in concept coverage between human and model summaries.

- An extensive evaluation of concept understanding in the scientific and financial domains. We design different types of factual and nonfactual (perturbed) concept definitions that integrate the context of a sentence-level concept classification task. Besides confirming the intuition that model scale is crucial (but not always sufficient) for concept understanding, the empirical results reveal relevant differences in the behavior of open-source and proprietary models, when subject to nonfactual concepts.

1.2 Outline

This thesis is organized as follows:

- *Chapter 2* reviews the recent summarization research, focusing on how scaling and improvements in language modeling are the main drivers for summarization progress, outpacing domain-specific engineering — the *bitter summarization lesson*. Also, an energy-based model view of summarization is outlined, motivating the experiments in subsequent chapters.

- *Chapter 3* introduces a factorized energy-based model, FactorSum, to separate the optimizations with respect to intrinsic and extrinsic importance factors. It also presents the experiments on long document summarization benchmarks, including scientific papers and government reports.

- *Chapter 4* expands the controllability investigation to large language models. By controlling stylistic and content coverage factors, we evaluate the performance of LLMs in 3 variants of scientific summarization: abstract generation, peer review summarization, and lay summarization.

- *Chapter 5* explores the controllability of scientific concepts in summaries for lay audiences. Also, it includes an evaluation of scientific and financial concept understanding for large language models of diverse scales.
- *Chapter 6* summarizes the findings and discusses limitations and possible future directions.

Chapter 2

Background: From Engineered Importance to Context Factors

This chapter provides an overview of recent summarization progress, focusing on milestones that changed the modeling of content importance: the transition from human-engineered importance to neural sequence-to-sequence models, and the surge of causal large language models. Underlying this presentation is the argument that most of the summarization-specific inductive biases proposed over the years were superseded by larger models, datasets, and input contexts. This *bitter summarization lesson* (inspired by Sutton 2019) motivates the formalization of summarization desiderata as an energy-based model introducing *context factors* (Jones, 1998) and serves as a framework for the development and evaluation of novel summarizers in the subsequent chapters.

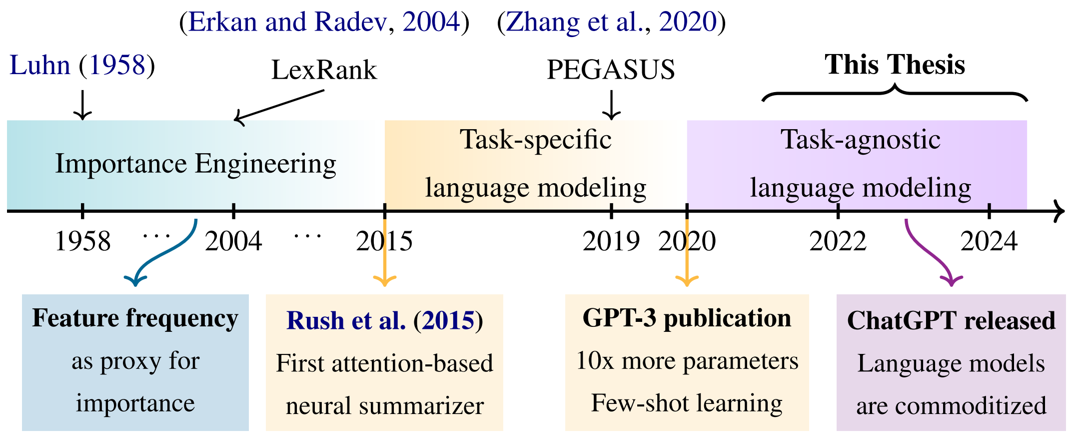

Figure 2.1: Illustrative milestones dividing automatic text summarization research into three dominant paradigms: importance engineering, task-specific supervised language modeling, and task-agnostic self-supervised language modeling.

2.1 Summarization Concepts and Scope

Since the seminal work by Luhn (1958), summarization has evolved as an important field to address the issue of consuming large volumes of textual information. Besides its utility as a discipline, summarization also serves as an important testbed for artificial intelligence systems, since it involves complex language understanding and generation steps. According to Jones (1998), to produce a good summary, a system has to 1) *interpret* the input, that is, to build an intermediate *representation* of the source text; 2) *transform* this representation to keep only relevant information; 3) and finally, *generate* a final textual summary. The different elements and design decisions in each of those steps characterize the different approaches to text summarization.

According to the inputs, the summarization tasks are commonly classified as *single-document* or *multi-document*, depending on the number of distinct sources being processed. The domain and characteristics of the documents usually demand specialized models. For instance, news articles (See et al., 2017) are known to be relatively short and concentrate relevant information at the beginning of the documents. In contrast, summarizing scientific articles (Cohan et al., 2018; Beltagy et al., 2020) requires processing longer documents with important more dispersed across sections.

The second step refers to the manipulation of input representations according to some content importance model (Peyrard, 2018). A plethora of approaches were proposed to capture important content, leading to different phases of summarization research depicted in Figure 2.1. These methods include word frequencies (Sparck Jones, 1972), topic models, sentence centrality metrics (Erkan and Radev, 2004), and neural network text encoders (Liu and Lapata, 2019).

Finally, the third step determines the type of summary to be generated. In *extractive* summarization, an optimization procedure is applied to select sentences based on the importance model from the second step. In contrast, abstractive summaries require the generation of a new text from the summary representation. This process can involve sentence compression/fusion and more recently, large neural language models (Zhang et al., 2020; Goyal et al., 2022).

In this thesis, we focus on *abstractive summarization of long technical documents*, including scientific articles (Cohan et al., 2018) and legal judgments (Malik et al., 2024). To situate our work in the broader summarization literature, we briefly present the relevant methods and milestones in summarization research, which we divide into three distinct phases (see Figure 2.1): importance via feature engineering, task-specific supervised sequence modeling, and task-agnostic unsupervised language modeling. For a more comprehensive literature review on summarization, refer to Peyrard (2019) and Nenkova and McKeown (2012).

2.2 Importance via Feature Engineering

The first phase of summarization research is characterized by the careful design of textual features that correlate to content importance. These features can based on the intuition of researchers or learned from data using machine learning techniques. Those systems are typically sentence-level extractive summarizers that combine a relevance scoring function and an optimization procedure to select sentences that maximize the relevance scores, subject to a given budget (Peyrard, 2019).

Frequency-driven approaches The simplest form of intuitive feature is word frequency, which serves as a signal for relevant topics in a document. However, choosing words merely by their frequency can lead to the choice of non-informative *stopwords* such as articles and conjunctions. To address this issue, more elaborated text mining approaches such as TF-IDF (Sparck Jones, 1972) take into consideration the term frequency on a background corpus. Terms that are more frequent in a given source document than in the rest of the corpus are considered as *topic signatures* (Lin and Hovy, 2000). Other topic modeling techniques applied to summarization include Latent Semantic Analysis (Deerwester et al., 1990; Yeh et al., 2005) and Latent Dirichlet Allocation (Blei et al., 2003; Wang et al., 2009).

Graph-based approaches Another class of summarization systems derives importance from structured representations of the documents. LexRank (Erkan and Radev, 2004) is a notable example that constructs a graph in which sentences are vertices and edges represent pairwise similarities between sentences. Then, a measure of centrality obtained via the PageRank algorithm (Brin and Page, 1998) is used to select the most central/relevant sentences. Later work proposes variants of this concept by adding sentence position information in news articles Zheng and Lapata (2019) and discourse-aware edge weighting based scientific rhetorical structure (Dong et al., 2021).

Discourse-driven and semantic approaches Rhetorical Structure Theory (Mann and Thompson, 1987) defines a method to structure texts by establishing discourse relations between *nucleus* and *satellite* text spans. The nuclei contain the most relevant content and hence are good candidates for summary sentences (Marcu, 1997, 1998). Alternatively, Liu et al. (2015) propose the use of Abstract Meaning Representation (Banarescu et al., 2013) to parse sentences into graphs of concepts and relations. Then, the sentence graphs are transformed into a summary graph, which is then converted to text using an AMR-to-text generator, leading to a data-driven, domain-agnostic summarization system.

Cognitively-inspired approaches A different line of work uses knowledge from cognitive science to inform summarization system design. By leveraging a cognitive model for narrative text comprehension, Zhang et al. (2016) propose a system that consolidates propositions from a source document in a *episodic memory*. These propositions and corresponding sentences are then selected based on their activation scores computed during reading. In a similar vein, Cardenas et al. (2024) addresses extractive summarization using the Micro-Macro Structure theory (Kintsch and Van Dijk, 1978) of cognitive processes involved in text comprehension. A key aspect of this formulation is that working memory during reading can be modeled as a *memory tree* with propositions as nodes. In this case, propositions that are closer to the root of the tree are selected as the most relevant content, while less important information in the leaves tends to be pruned as the sentences are processed.

2.3 Importance via Supervised Sequence Modeling

The next trend in summarization research capitalized on the representation learning breakthroughs started in the computer vision community (Deng et al. 2009; Krizhevsky et al. 2012, *inter alia*) and later adapted to natural language processing problems (Kalchbrenner et al., 2014; Bahdanau et al., 2014). In this phase, summarization systems moved away from explicit importance modeling to end-to-end optimization based on large datasets and neural network models. Instead of extractive sentence selection via discrete optimization procedures (Carbonell and Goldstein, 1998), these systems can generate novel abstractive summaries, conditioned on the source text (Rush et al., 2015).

Besides computing resources, one crucial enabler of neural sequence-to-sequence summarizers was the availability of new large-scale summarization datasets such as CNN/DailyMail (Hermann et al., 2015) and PubMed (Cohan et al., 2018). In Fig-

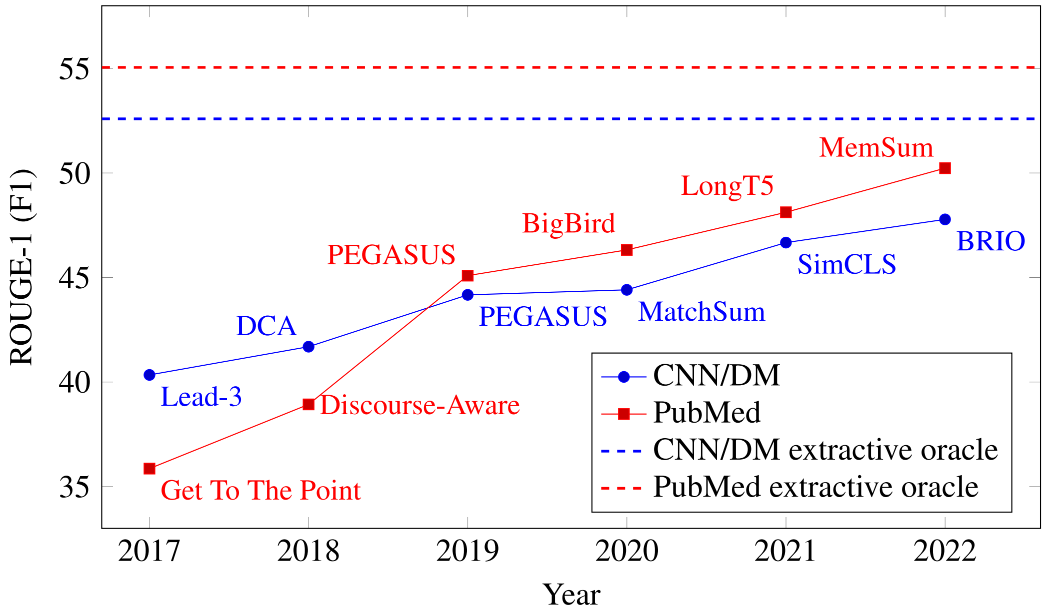

Figure 2.2: Summarization performance per year, as measured by ROUGE-1, on the CNN/DailyMail (Hermann et al., 2015) and PubMed (Cohan et al., 2018) benchmarks.

ure 2.2, we present a sample of the best-performing system across the years, showing how advances in language modeling helped the generated summaries to achieve higher lexical similarity to reference summaries, as measured by ROUGE (Lin, 2004). The first model, the pointer-generator network (See et al., 2017), was designed to address factuality issues in early language generation models. It promotes a balance between generating new words from the vocabulary and copying words from the source text (pointing), which encourages more accurate reproduction of content in summaries. This method was further augmented with hierarchical discourse information by Cohan et al. (2018), leading to significant performance improvements.

The transition from attention-based recurrent networks (Bahdanau et al., 2014) to transformer-based architectures (Vaswani et al., 2017) resulted in marked performance improvements for summarizers. PEGASUS (Zhang et al., 2020), a large-scale model (by the standards of that time), introduced innovations in the pre-training strategy by adding a masked sentence prediction objective, which resulted in performance gains across several domains.

Later work addressed the problem of processing long documents with more than 1,000 sub-word tokens, which is a known problem of the transformer architecture. This limitation derives from the self-attention mechanism used in those models, which is defined as follows (Vaswani et al., 2017):

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{2.1}$$

where Q, K, and V are matrices that project the input embeddings into queries, keys, and values, respectively. The scalar d_k is the dimension of the queries and keys. Hence the multiplication QK^T is a dot product between all pair queries and keys, which scales quadratically with the number of input tokens. Zaheer et al. (2020) explore different sparsification strategies for this attention mechanism, scaling the input context up to 8 times compared to the original attention implementation. Similar approaches were also introduced by Longformer (Beltagy et al., 2020) and LongT5 (Guo et al., 2022), which are particularly suitable for summarizing long documents such as PubMed, as shown in Figure 2.2.

In addition to general-purpose architectures such as LongT5, more specialized models leverage summarization-specific innovations. BRIO (Liu et al., 2022b) and SimCLS (Liu and Liu, 2021) replace the regular training based on a single reference summary with a contrastive objective that scores multiple candidate summaries according to their quality. MemSum (Gu et al., 2021) approaches extractive summarization with reinforcement learning and an iterative sentence selection algorithm based on the sentence context and extraction history.

Although Figure 2.2 show clear trends of improvement in terms of similarity to reference summaries, there are known limitations of this paradigm, especially related to the quality of existing benchmarks and evaluation protocols (Kryściński et al., 2019). For instance, recent work by Stiennon et al. (2020) shows that when optimizing summaries directly from human feedback, ROUGE scores are informative for low-quality summaries but do not correlate with human preferences in general.

2.4 Importance via Unsupervised Language Modeling

A more recent paradigm shift in summarization (and language generation in general) is the usage of large language models (LLMs) based on the GPT architecture (Brown et al. 2020; Touvron et al. 2023; Achiam et al. 2023, *inter alia*). Those systems not only abandon the idea of explicit importance modeling but also are not specifically trained for summarization. Instead, the models are exposed to summarization (and other tasks) via the pre-training data and during instruction tuning (Ouyang et al., 2022).

Recent work demonstrates that LLMs achieve strong performance in text summarization when evaluated on human preferences. For instance, Goyal et al. (2022) found that human evaluators prefer summaries produced by GPT-3 over strong supervised baselines such as BRIO (Liu et al., 2022b). Furthermore, Zhang et al. (2024) concluded that instruction-tuned models performed on par with human freelance writers and, agreeing with Kryściński et al. (2019), found that the evaluation of news summarization is limited by the low quality of reference summaries. These results reinforce the need for more comprehensive and robust evaluation protocols. In Chapter 4, we provide an extensive evaluation of LLMs on summarization of long documents, complementing the existing literature focusing on news articles.

2.5 Summarization Context Factors

Most work presented in previous sections focuses on features of the documents or implicit features that should be captured by a large collection of document-summary pairs. However, several aspects that determine content relevance and the shape of a summary might not be inferred from the source but depend on context factors (Jones, 1998), which can be classified into *input*, *purpose*, and *output* factors.

Input factors refer to the source format and subject. For instance, scientific papers have specific structures with headings that provide information about the location of relevant content. Papers also use a particular kind of technical language and visual presentation tools (e.g., tables and diagrams) that synthesize experimental methods and results that should be reflected in scientific summaries.

The second (and most relevant) context factor type is *purpose*, which relates to how the summary audience and goals. In scientific peer review, it is common to provide summaries that emphasize the main strengths and weaknesses of papers, which determines the focus of summaries targeting a highly technical audience (authors and other reviewers). In contrast, paper abstracts tend to highlight the technical merits and main results, since the author has the intention to persuade the reviewer of the impact of the contributions. Other kinds of summaries aim to inform lay audiences (Goldsack et al., 2022), adding more background information and commonsense analogies about the topic, as well as simplifying technical jargon. We discuss experiments on scientific lay summarization in Chapter 4.

Lastly, *output* factors capture aspects of the style and format of the summary. Typically, a summary can be presented as running text, use bullet points (as in some news articles), or use a pre-defined structure as in summaries for papers reporting randomized controlled trials (RCTs; Dernoncourt and Lee 2017). In Chapter 4, we explore particular prompts to change the narrative style of a summary and match the first-person perspective found in paper abstracts.

Recent research is focusing on alternative evaluation protocols based on context factors. Ter Hoeve et al. (2022) conduct a survey among university students and suggest possible research directions to improve summary alignment to students' preferences. Another survey by August et al. (2024) investigates the effects of using simplified language in scientific summaries as perceived by readers with varying levels of expertise. In the future, we expect more research on evaluation using context factors as more powerful summarizers based on large language models are developed.

2.5.1 A Energy-Based View of Context Factors

While the context factors presented in the previous section provide a useful conceptual direction for summarization research, they lack sufficient formalization to inform novel modeling approaches. In this section, we formulate an energy-based model (EBM; LeCun et al. 2006) intuition that serves as a framework for our context-aware summarization experiments.

EBMs capture dependencies between variables by assigning a scalar value (an *energy*) to each combination of variable values. Then, one can perform *inference* by finding the value of a variable that minimizes the energy while keeping the other variables fixed. During *learning*, the parameters of the energy function are optimized so that compatible variable configurations have lower energies and bad configurations are assigned high energies.

Let $E_\theta(D,S,F)$ be an energy function for a summarization model parameterized by θ. Then, given a source document $D = d_i$ and context factors $F = f_i$, the best summary S^* is inferred via the following optimization procedure:

$$S^* = \arg\min_{S \in \mathcal{S}} E_\theta(D = d_i, F = f_i, S). \tag{2.2}$$

During training, the parameter θ in Eq. 2.2 is optimized by minimizing the *loss functional* $\mathcal{L}(E_\theta, \mathcal{T})$ that measures the quality of the energy function E_θ given a dataset of $\mathcal{T} = \{(D_i, R_i) \colon i = 1 \dots K\}$ consisting of documents D_i paired with reference summaries R_i:

$$\theta^* = \arg\min_{\theta} \mathcal{L}(E_\theta, \mathcal{T}). \tag{2.3}$$

This energy-based formulation is convenient to incorporate context factors due to its flexibility. First, there is a clear distinction between the two optimization steps — learning and inference — so that context factors can be addressed in a modular way without impacting the training step. And importantly, these context factor objectives are more versatile since they do not need to be expressed as probabilistic models depending on intractable partition functions over all possible summaries.

In practice, for the supervised sequence-to-sequence models such as PEGASUS (Zhang et al., 2020), the optimization objective in Eq. 2.3 corresponds to the maximum likelihood estimation:

$$\theta^* = \arg\max_{\theta} \sum_{i}^{K} \log p_{g_\theta}(R_i|D_i), \tag{2.4}$$

where p_{g_θ} is a probability distribution derived from neural network model $g_\theta : D \longrightarrow S$. In those cases, the context factors F are latent and it is not possible to fix specific factor values during inference. One possible solution is to augment the training dataset by adding extra control tokens (Fan et al., 2018) or prompts (He et al., 2022) to the inputs so that controllability factors can be specified at inference time. The obvious disadvantage of this approach is that any change in the space of control factors requires retraining the entire model. In Chapter 3, we propose an alternative solution that factorizes the energy in Eq. 2.2 so that summaries are optimized for context factors via a separate *extrinsic* energy, avoiding expensive retraining.

The large language models described in Section 2.4 are trained on massive datasets and are exposed to diverse kinds of summarization tasks and context factors F during pre-training and instruction fine-tuning. Consequently, these models offer extreme flexibility for specifying context factors via prompts. These prompts range from simple stylistic *output* factors such as *summarize the document in 5 sentences* or more abstract *purpose* factors such as *summarize the document for a lay audience*. In Chapter 4, we investigate different types of such prompt strategies to adapt summaries to diverse communication goals.

Chapter 3

Factorizing Intrinsic and Extrinsic Importance

Typically, supervised summarization models are trained end-to-end to replicate the style of reference summaries. While effective, this approach makes it harder to adapt to context factors that require summary styles outside the training distribution. One such example is generating summaries that are longer (or shorter) than the reference summaries observed during training. In this chapter, we propose to introduce context factors by disentangling content selection from the final summary budget optimization, making abstractive summarization models more versatile and cheaper to train.

The proposed method, FACTORSUM, does this disentanglement by factorizing summarization into two steps through an energy function: (1) generation of *abstractive summary views* covering salient information in subsets of the input document (*document views*); (2) combination of these views into a final summary, following a budget and content guidance. This guidance may come from different sources, including from an *advisor* model such as BART or BigBird, or in oracle mode – from the reference.

This factorization achieves significantly higher ROUGE scores on multiple benchmarks for long document summarization, namely PubMed, arXiv, and GovReport. Notably, our model is effective for domain adaptation. When trained only on PubMed, it achieves a 46.29 ROUGE-1 score on arXiv, outperforming PEGASUS trained in-domain by a large margin. Furthermore, we report results of FACTORSUM on the summarization of court judgments, where it achieves competitive results but cannot match the performance of transformers designed for long documents such as Longformer.

3.1 Introduction

Casting summarization as a language transduction problem is convenient given the existence of powerful neural sequence-to-sequence models that produce high-quality textual outputs (Zhang et al., 2020; Lewis et al., 2020). However, this framework conflates multiple steps of the summarization process into a single feedforward step without taking into account the contextual factors involved (Jones, 1998).

One such decision depending on context factors is the quantity of information to be included in a summary, reflecting on the generated outputs' length. This factor is particularly relevant for long documents that cover many aspects of interest for which different summaries may be suitable. For instance, in samples from summarization datasets such as PubMed and arXiv (Cohan et al., 2018), there are many abstracts including terse passages about the background or methods of the research. In contrast, others will add more details about those aspects. Often, those choices are due to the author's preferences and do not represent an ideal summary for the document.

Furthermore, current evaluation protocols based on n-gram overlap are sensitive to summary lengths (Sun et al., 2019). For instance, the results in Section 3.3.2.1 provide evidence that matching ground-truth lengths increases performance significantly. Thus, recent progress in summarization may be the effect of better length prediction and not the actual summarization desideratum: a reductive transformation of the source text that keeps the important information (Jones, 1998).

To address this issue, we propose to avoid budget information as a confounding factor as much as possible in sequence-to-sequence training. Instead, we treat budget decisions as *extrinsic guidance* during summary generation, that is, an objective that is unrelated to the content of the documents. In this setting, the neural abstractive model is responsible for the generation of short passages (*summary views*) capturing relevant topics of the input document (*intrinsic importance* objective), while the *extrinsic importance* objective will encourage the adherence of generated summaries to context factors such as budgets or aspect coverage.

Specifically, we formulate FACTORSUM, a factorized energy-based model (LeCun et al., 2006) aiming to find a summary that maximizes the total importance given a source document, a reference dataset, and contextual factors such as budget and content guidance. A key piece of our model is the sampling of *random document views* that allows the abstractive model to focus on shorter summarization tasks with less influence of varying summary lengths. Document views are random subsets of the input documents, sampled at the sentence level (refer to Section 3.2.1 for details), which provide multiple shorter versions of the same input. Then, for each document view, the model generates a separate summary (a *summary view*), resulting in multiple summaries for each original document. This approach allows the processing of long documents without truncation, which is a recurring problem in summarization (Beltagy et al., 2020; Zaheer et al., 2020).

Our model comprises two optimization procedures: learning and inference. In the learning phase, the model parameters are optimized so that summary views with important content (as informed by reference summaries) will have lower energies. In practice, this is implemented by training a neural sequence-to-sequence model to predict summary views. During inference, a greedy optimization algorithm is used to find the combination of summary views that maximize the compatibility with the target budget and other types of guidance. This process is illustrated in Figure 3.1.

Our experimental results on the PubMed, arXiv, and GovReport summarization benchmarks (Cohan et al., 2018; Huang et al., 2021) show that our approach using budget guidance alone is competitive with resource-intensive baselines such as PEGASUS (Zhang et al., 2020). The results confirm that matching reference summary lengths significantly impacts ROUGE scores, often more than different modeling approaches. We also investigate the use of existing baselines as additional guidance during summary generation. In contrast to teacher models in knowledge distillation literature (Hinton et al., 2015), we leverage existing model prediction during inference only, and thus, we adopt the term *advisor* model to refer to our summarization guidance approach. When guided by BigBird or BART, our model obtains state-of-the-art results on PubMed, arXiv, and GovReport.

Additionally, we perform domain adaptation experiments in which models trained on PubMed, arXiv, and GovReport have no access to samples from the evaluation dataset during training. Our results indicate that FACTORSUM can adapt better to out-of-domain data, outperforming strong baselines trained in domain on both PubMed and arXiv. This finding suggests a good generalization capacity and is evidence that we achieved our objective to disentangle content selection from budget decisions.

Finally, we evaluate FACTORSUM on a different technical domain: court judgment cases. Specifically, we use CIVILSUM (Malik et al., 2024), a novel dataset of Indian court decisions paired with human-written summaries. These legal documents exhibit different features compared to scientific articles such as the prevalence of relevant information towards the end of documents. While we find that FACTORSUM perform

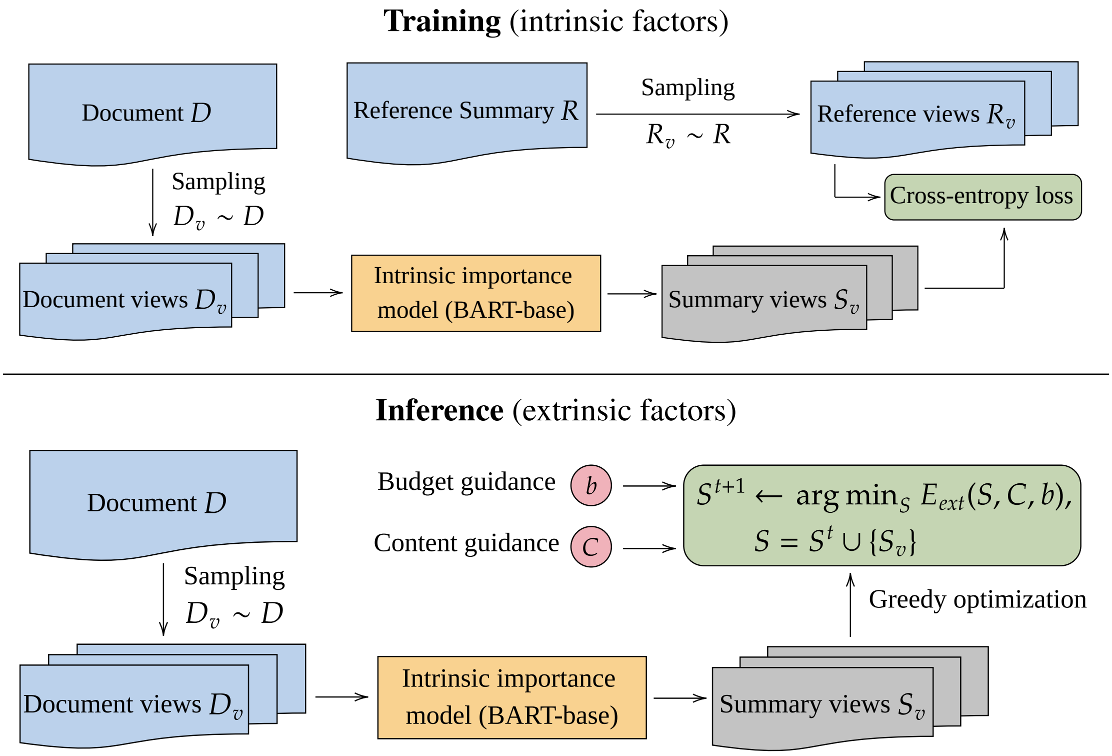

Figure 3.1: An overview of the summarization model. During training **(top)**, an encoder-decoder model (BART) is trained to predict summary views S_v from document views D_v, which are subsets of sentences randomly sampled from the original documents D. At inference time **(bottom)**, the encoder-decoder model generates multiple summary views for each document. The sentences in those views are combined into a final summary following a greedy optimization algorithm that maximizes adherence to a budget and content guidance.

well on this task, our qualitative human evaluation reveal significant limitations in summary fluency compared to baselines such as Longformer (Beltagy et al., 2020).

3.2 Intrinsic and Extrinsic Importance

Since our objective is to explicitly model budget decisions, we need a definition of importance that accounts for context factors. Inspired by the information-theoretic notion of importance developed by Peyrard (2018), we introduce the notion of importance with respect to *intrinsic* and *extrinsic* semantic units.

Intrinsic semantic units are those specific to the document (e.g., salient topics) whereas extrinsic units are related to a priori external preferences or require domain knowledge and grounding that is hard to capture from the textual corpora alone. We argue that the usual setting of end-to-end supervised summarization evaluated by ROUGE (Lin, 2004) optimizes for intrinsic importance. In this work, budget and content guidance provided by advisor model summaries (Section 3.2.1.2) play the role of extrinsic information.

Formally, we define the best summary S^* for a document D as the summary that minimizes the following factorized energy (LeCun et al., 2006):

$$E(\theta, D, S, C, b) = E_{int}(\theta, D, S) + E_{ext}(S, C, b), \quad (3.1)$$

where we call E_{int} and E_{ext} intrinsic and extrinsic energies respectively, while b denote the summary budget guidance and C is a guidance content provided by an advisor model as explained in Section 3.2.1.2. This energy point of view allows us to unify the notion of extrinsic and intrinsic semantic units, and present the duality of the energy functions with respect to learning and inference.

Furthermore, the factorization of the total energy function makes the problem more tractable and leads to the following advantages:

- Model components can be changed or replaced more cost-effectively. For example, adding more components to the extrinsic objective would not require retraining the intrinsic importance model.

- Issues with differentiability of the extrinsic guideline loss with respect to the summary views generator parameters are avoided.

- More complex inference procedures than just feed-forward computation are possible.

An overview of the model components and the summary inference procedure is represented in Figure 3.1. Given a document D, n_d document views are generated, each covering a random subset of sentences from the original document (Section 3.2.1). Then, the intrinsic importance model generates *summary views* S_v, each partially covering salient content from the original document D (Section 3.2.1.1). Finally, the extrinsic importance model will optimize the final summary S so that it maximizes the alignment of the content to a target budget and content guidance. In the following sections, we detail the model components as well as the training and inference procedures.

3.2.1 Sampling Document Views

Our model requires multiple summary proposals (or views) to allow the optimization of the extrinsic energy (Eq. 3.1). One further motivation for using document views is that it allows the intrinsic encoder-decoder model to focus on shorter sequences, which makes the less affected by truncation issues. To generate multiple *views* for the same document, we implement the following steps (refer to Figure 3.2 for an illustration):

- From a document D, we generate a random sample of sentences, which we call a *document view* D_v. The number of sentences in D_v is controlled by the sampling factor parameter $s_f \in [0,1]$, so that $\text{n\_sents}(D_v) \approx s_f \cdot \text{n\_sents}(D)$.

- Also from D, we extract oracle sentences o_i corresponding to each sentence r_i in the reference summary R. We choose as oracle the sentences that maximize the sum of ROUGE-1 and ROUGE-2 F1 scores: $o_i = \text{argmax}_{s \in D} \text{ROUGE\_1}(s, r_i) + \text{ROUGE\_2}(s, r_i)$. In our experiments, we found that ROUGE performs on par with metrics such as BERTScore (Zhang et al., 2019), while demanding less computation.

- For each oracle sentence $o_i \in D_v$ we collect the corresponding sentence r_i from the reference summary R. These sentences r_i form the reference summary R_v for the document view D_v. If there is no oracle sentence in the document view, the reference summary view R_v is empty[1].

For each document D from a training dataset $\mathcal{T}$, we repeat the sampling procedure described above n_d times, yielding a new dataset $\mathcal{T}' = \{(D_v^{(i)}, R_v^{(i)}) \; : i = 1, \ldots, |\mathcal{T}| \times n_d\}$ with n_d times more samples than the original data. The number of samples n_d and the sample fraction s_f are hyperparameters that to be tuned for each dataset. For PubMed, by sampling $n_d = 20$ views per document, each with 20% of the document sentences, we obtain document views with 17.2 sentences on average, while covering 99.1% of the original oracle sentences. In Section 3.3.2.5, we provide statistics for different n_d and s_f and the heuristics we use for choosing appropriate values.

The intuition behind this sampling method is that if a sentence is relevant for the entire document, it should also be relevant in different contexts. Besides allowing the decoupled energy minimization objective, this approach also makes the input documents and corresponding summaries much shorter than the original data. Thus, our

[1] Except during training, when we enforce that each document view has at least one oracle sentence.

method scales to long documents without requiring specialized architectures for modeling long sequences (Beltagy et al., 2020; Zaheer et al., 2020). Also, in contrast to previous work, this summary sampling is domain-agnostic as it does not make any assumption about the discourse structure of the document (Dong et al., 2021; Gidiotis and Tsoumakas, 2020).

3.2.1.1 Intrinsic Importance Model

Powerful sequence-to-sequence models such as PEGASUS (Zhang et al., 2020) and BART (Lewis et al., 2020) are trained to estimate the probability of a sequence of tokens given a document by minimizing cross-entropy with respect to the data distribution. We hypothesize that these models are good candidates to fulfill the intrinsic importance objective, as described below.

Learning Given the training dataset $\mathcal{T}'$ consisting of document views $D_v^{(i)}$ and reference summary views $R_v^{(i)}$ (Section 3.2.1), we define the intrinsic loss as a negative log-likelihood functional:

$$L(E_{int}, \mathcal{T}') = \frac{1}{|\mathcal{T}'|} \sum_{i=1}^{|\mathcal{T}'|} \underbrace{L(R_v^{(i)}, E_{int}(\theta, D_v^{(i)}, \mathcal{S}))}_{-\log p_\theta(R_v^{(i)}|D_v^{(i)})}$$

where $p_\theta(R_v^{(i)}|D_v^{(i)})$ is a distribution over the possible summaries $\mathcal{S}$, specifically a sequence-to-sequence neural network model (Lewis et al., 2020). During learning, we find the parameters θ^* that minimize the loss above.

Inference The summary generation is performed as usual in sequence-to-sequence models via beam search decoding (Sutskever et al., 2014). We sample summary views by generating a summary conditioned to the document views:

$$S_v^{(i)} \sim p_{\theta^*}(\,\cdot\,|D_v^{(i)}). \tag{3.2}$$

We assume these summary views are samples from low-energy regions of $E_{int} = -\log p_{\theta^*}(S_v^{(i)}|D_v^{(i)})$, thus contributing to the minimization of the factorized energy (Eq. 3.1).

3.2.1.2 Extrinsic Importance Model

The extrinsic importance energy function E_{ext} measures the compatibility between the summary, the guidance budget b, and the guidance content C. Thus, the optimal sum-

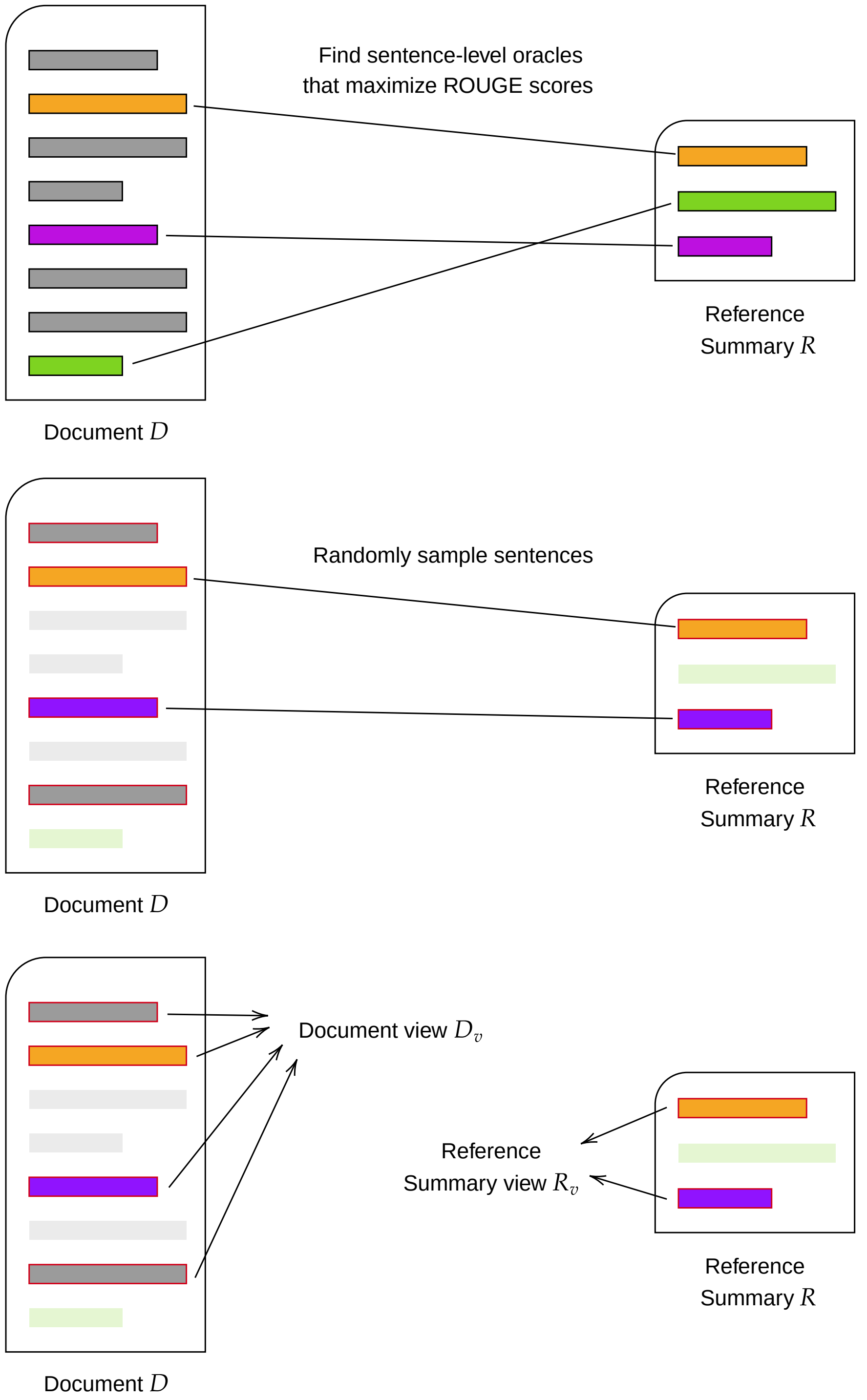

Figure 3.2: Steps for the generation of document and reference summary views. From top to bottom: 1) oracle sentences from the document D are selected based on similarity to sentences in the reference summary R; 2) sentences from the document are uniformly sampled, forming a *document view* D_v; 3) the *reference summary view* R_v is composed by sentences that correspond to oracles in the document view D_v.

mary S^* is defined as:

$$S^* = \arg\min_S E_{ext}(S, C, b). \tag{3.3}$$

The extrinsic energy is defined in terms of the squared deviation with respect to the guidance budget and ROUGE-1 score between the generated summary and the guidance content:

$$E_{ext}(S, C, b) = \alpha \, (|S|/b - 1)^2 - \beta \text{ ROUGE\_1}(S, C), \tag{3.4}$$

where $|S|$ denotes the length of the summary in words, the content C is a summary provided by an advisor model. The hyperparameters α and β weight the contribution of each guidance signal. In our experiments, we use $\alpha = \beta = 1.0$ and, as advisor models, PEGASUS (Zhang et al., 2020) and BigBird (Zaheer et al., 2020).

In our implementation, there is no learning step for the extrinsic importance model. For inference, we design a greedy algorithm to minimize the energy as detailed in Algorithm 1. Let V_D be the set of n_d summary views for the document D. Starting from the initial condition $S = \emptyset$, the procedure selects the summary view $S_v \in V_D$ that minimizes the energy $E_{ext}(S \cup \{S_v\}, C, b)$. The view S_v is added to the summary if it satisfies the following additional conditions:

- **Non-redundancy**: the view S_v cannot be redundant with respect to the current summary S. We consider as redundant a summary view that has a (word-level) normalized Levenshtein distance[2] (Levenshtein et al., 1966) to any sentence in S lower than a threshold $t = 0.4$[3] (`is_redundant` function in Algorithm 1).

- **Energy reduction**: $S \cup \{S_v\}$ must have a lower energy than the current best summary S^*. After p iterations without improvement, the algorithm returns the current best summary S^*. Unless otherwise stated, this patience parameter is set to $p = n_d$, which means the algorithm iterates over all available views.

When $\beta = 0$ in Eq. 3.4 (no content guidance), each step of the greedy algorithm adds the longer summary view that satisfies the non-redundancy condition above, except for the last step, when a shorter view may better match the budget guidance. We provide further details on pre- and postprocessing summary views in Appendix 3.2.2.1.

[2] We use the `textdistance` library: https://github.com/life4/textdistance

[3] The redundancy threshold was manually tuned by inspecting sample outputs from the validation set.

Algorithm 1: Greedy summary generation. Input parameters are the set of summary views V_D for document D, content guidance C, budget guidance b, redundancy threshold t, and patience p. See the "Non-redundancy" paragraph in Section 3.2.1.2 for a discussion about the `is_redundant` function.

Input: V_D, C, b, t, p

Output: S^*

$S \leftarrow \emptyset, S^* \leftarrow \emptyset$;

$i \leftarrow 0$;

while $V_D \neq \emptyset$ **and** $i \leq p$ **do**

  $S_v^* \leftarrow \arg\min_{S_v \in V_D} E_{ext}(S \cup \{S_v\}, C, b)$

  if $E_{ext}(S \cup \{S_v^*\}, C, b) > E_{ext}(S^*, C, b)$ **then**

    $i \leftarrow i + 1$;

    $S \leftarrow S \cup \{S_v^*\}$;

  else if not $is\_redundant(S, S_v^*, t)$ **then**

    $i \leftarrow 0$;

    $S \leftarrow S \cup \{S_v^*\}$;

    $S^* \leftarrow S$;

  $V_D \leftarrow V_D \setminus \{S_v^*\}$;

end

3.2.2 Training the Intrinsic Importance Model

The intrinsic importance model $p_\theta(R_v^{(i)} | D_v^{(i)})$ described in Section 3.2.1.1 is implemented using the BART sequence-to-sequence model (Lewis et al., 2020). We fine-tune the `bart-base` checkpoint from huggingface[4] on the datasets of document and summary views $\mathcal{T}'$ presented in Section 3.2.1. Unless otherwise stated, we use $n_d = 20$ samples per document and a sampling factor $s_f = 0.2$, as explained in Appendix 3.3.2.5. To ensure replicability, we use a random seed for document views sampling.

For the training process, we use 4 GeForce GTX 1080 Ti GPUs each with 12GB of memory. Table 3.1 details the training set size (number of documents and summary views), number of training steps, and time to train the intrinsic importance models for each dataset. The main training hyperparameters are presented in Table 3.2.

[4] https://huggingface.co/facebook/bart-base

Dataset	Docs	Summary Views	Steps	Hours
PubMed	115k	2.3M	100k	91
arXiv	200k	4M	200k	180
GovReport	17.5k	340k	50k	51

Table 3.1: Statistics for number of documents, summary views, training steps, and total training time for each dataset (intrinsic model based on BART-base).

Parameter	BART base	BART large
Checkpoint	base	large
Number of parameters	139.4M	406.3M
Training time	Table 3.1	224 hours
Batch size per GPU	4	1
Gradient accumulation	4	32
Effective batch size	64	32
Learning rate	5×10^{-5}	2×10^{-5}
Learning rate scheduler	linear	
Optimizer	AdamW	
Adam β_1	0.9	
Adam β_2	0.999	
Adam ε	1×10^{-8}	
Metric best model	ROUGE-1 F1	
Floating point precision	FP16	
Max source length	1024	
Max target length	128	768
Generation beams	4	8
Length penalty	1.0	0.6
DeepSpeed ZeRO[5]	-	Stage 1

Table 3.2: Training details and hyperparameters for the intrinsic model (BART-base) and the end-to-end baseline for GovReport (BART-large).

3.2.2.1 Inference Details

For reproducibility purposes, we provide the generation details for the end-to-end baseline models in Table 3.3. The intrinsic importance model (BART-base) generation uses the same maximum source length, maximum target length, beam size, and length penalty defined for training in Table 3.2.

Extrinsic importance model For summary length control, we adjust the budget guidance so that the average summary lengths is close to the average length of the first 1,000 summaries from the validation set (or the entire validation set for GovReport). The budget guidance for each model/guidance type is shown in Table 3.4. For the domain adaptation results in Table 3.8, the budget guidance corresponds to the target dataset, i.e., it is the in-domain budget listed in Table 3.4.

In addition to the inference procedure described in Algorithm 1, we apply a preprocessing step that divides each summary view $S_v \in V_D$ into sentences. The resulting sentence-tokenized set of summary views V_D is the union of all sentences. We use the sentence tokenizer provided by `NLTK`[6].

For FACTORSUM versions using content guidance, the best summary S^* returned by Algorithm 1 is reordered according to the following procedure: (1) for each summary view in S^*, we collect the index of the oracle sentence in the content guidance text[7]; (2) The summary views are sorted according to the list of corresponding oracle indexes, using the Python `sorted` function[8].

3.3 Experiments with Scientific Papers and Reports

Our first experimental setup focus on scientific papers and long government reports. In this section, we specify details about datasets, baselines, and the evaluation protocol.

3.3.1 Experimental Setup

Datasets Our experiments are performed on the PubMed and arXiv datasets, consisting of documents extracted from the homonymous scientific repositories (Cohan et al., 2018). To further test the generalization capacity of the model, we also perform

[6] `nltk.org`
[7] Oracle sentences determined as described in Section 3.2.1.
[8] `https://docs.python.org/3/library/functions.html#sorted`

PEGASUS	
Checkpoint (arXiv)	pegasus-arxiv
Checkpoint (PubMed)	pegasus-pubmed
Number of parameters	570.8M
Max source length	1024
Generation beams	8
Length penalty	0.8
BigBird	
Checkpoint (arXiv)	bigbird-pegasus-large-arxiv
Checkpoint (PubMed)	bigbird-pegasus-large-pubmed
Number of parameters	576.9M
Max source length	3072
Generation beams	5
Length penalty	0.8
BART-large	
Checkpoint (GovReport)	See Section 3.2.2
Number of parameters	406.3M
Max source length	1024
Generation beams	8
Length penalty	1.0
All models	
Max target length (arXiv, PubMed)	256
Max target length (GovReport)	768

Table 3.3: Summary generation details and parameters for the end-to-end baselines.

Guidance Budget	Guidance Content	Pubmed	arXiv	GovReport
FactorSum - no content guidance				
Oracle	-	216	169	656
Fixed	-	213	167	656
Model	-	217	170	698
FactorSum - content guidance				
Oracle	Lead	221	169	632
Fixed	Lead	217	167	624
Fixed	Model	232	175	658
Model	Model	227	177	658

Table 3.4: Budget guidance used for FactorSum models in Table 3.6. Model guidance is provided by BART-large for GovReport and BigBird for PubMed and arXiv.

the experiments on GovReport, a dataset containing long reports published by U.S. Government Accountability Office (GAO; Huang et al. 2021). The only preprocessing applied is to filter out documents with empty articles or summaries, which lead to the training, validation, and test splits shown in Table 3.5. We do not truncate the articles or their abstracts.

Evaluation Metrics We evaluate our models using the ROUGE F-measure metric (Lin, 2004), with the implementation used by Zhang et al. (2020)[9].

[9] https://github.com/google-research/pegasus

Dataset	Samples			Summaries	
	Train	**Validation**	**Test**	**Sentences**	**Words**
PubMed	119,920	6,631	6,658	6.8	204.8
arXiv	202,917	6,436	6,440	12.6	292.6
GovReport	17,517	973	973	17.6	546.0

Table 3.5: Key statistics for the summarization datasets. ”Sentences” and ”Words” denote the average number of words and sentences in the summaries (training split).

Baseline models We use the following summarization baselines, for which implementations and pre-trained models are publicly available:

- **PEGASUS** (Zhang et al., 2020), an encoder-decoder transformer-based model that uses a specialized pre-training task of predicting entire masked sentences and achieves strong performance across several datasets.

- **BigBird** (Zaheer et al., 2020), a model based on a sparse attention mechanism that allows transformer-based models to process up to 8 times longer sequences efficiently.

- **BART** (Lewis et al., 2020), a transformer-based denoising autoencoder that has strong performance on text generation tasks. We train our own version of BART-large on GovReport with a longer maximum target length of 768 tokens.

Also, we add results for the following abstractive systems: DANCER (Gidiotis and Tsoumakas, 2020), HEPOS (Huang et al., 2021), DYLE (Mao et al., 2022), and SUMM^N (Zhang et al., 2021).

As stated above, we train our own end-to-end BART-large baseline on the GovReport dataset. Since the target summaries are long, we set the maximum summary generation length to 768 tokens, which makes the memory requirements to exceed most single-GPU capacities (even with batch size equal to one). To address this problem, we resort to model parallelism techniques provided by the DeepSpeed library (Rajbhandari et al., 2020), allowing the efficient distribution of the model across 4 GeForce GTX 1080 Ti GPUs, each with 12GB of memory. We use gradient accumulation to achieve a effective batch size of 32. Table 3.2 details the training hyperparameters.

3.3.2 Results and Discussion

In this section, we analyze the contribution of different guidance factors on summarization performance by controlling summary budget and content guidance. Furthermore, we conduct an ablation study, examining the summary views and the greedy summary generation contributions to the overall performance. Finally, we discuss domain adaptation results. Sample summaries are provided in Appendix A.4.

Model		PubMed				arXiv				GovReport			
		R-1	R-2	R-L	Len	R-1	R-2	R-L	Len	R-1	R-2	R-L	Len
Previous work													
DANCER†		46.34	19.97	42.42	-	45.01	17.60	40.56	-	-	-	-	-
HEPOS†		**48.12**	**21.06**	42.72	-	48.24	20.26	41.78	-	56.86	22.62	53.82	-
DYLE†		-	-	-	-	46.41	17.95	41.54	-	**61.01**	**28.83**	**57.82**	-
SUMM^N†		-	-	-	-	-	-	-	-	56.77	23.25	53.90	-
PEGASUS		43.83	18.72	40.29	180	43.06	16.39	38.65	168	-	-	-	-
BigBird		45.48	19.92	41.81	185	46.15	18.60	41.46	164	-	-	-	-
BART-large		-	-	-	-	-	-	-	-	52.82	19.12	49.99	596
Guidance													
Budget	**Content**	**FACTORSUM - no content guidance**											
Oracle	-	47.37	19.10	43.27	208	48.87	18.83	43.96	167	59.80	24.13	56.12	651
Fixed	-	45.41	18.66	41.63	206	47.22	18.60	42.61	165	58.77	23.99	55.19	650
Model	-	44.64	17.98	40.76	185	46.40	18.21	41.85	164	57.18	23.34	53.66	638
FACTORSUM - content guidance													
Oracle	Lead	48.31	19.99	44.35	208	49.69	19.32	44.85	166	60.73	25.24	57.20	650
Fixed	Lead	46.27	19.29	42.57	205	48.05	19.05	43.49	165	59.67	25.02	56.22	649
Fixed	Model	47.50	20.33	<u>**43.76**</u>	205	<u>**49.32**</u>	<u>**20.27**</u>	<u>**44.76**</u>	165	60.10	25.28	56.65	648
Model	Model	47.34	20.31	<u>43.52</u>	185	48.74	<u>20.12</u>	44.19	164	58.78	24.87	55.37	638

Table 3.6: ROUGE F1 scores and average words per summary on the test sets for different types of guidance during inference. *Lead* guidance is the first k sentences from the source document (Section 3.3.2.2). Model guidance is provided by BART-large for GovReport and BigBird for PubMed and arXiv. The choice of budget guidance values is described in Appendix 3.2.2.1 and validation scores are provided in Appendix A.1. Results for models marked with † are taken from the original publications. <u>Underlined results</u> are statistically equivalent to the best methods ($p < 0.05$).

3.3.2.1 Effects of Budget Guidance

To test the impact of budgets on the summarization performance we provide three types of guidance:

- **Fixed**: the budget guidance is set at a fixed value for all summaries. The budget is 205, 165, and 648 words, which are the average summary lengths in the validation sets of PubMed, arXiv, and GovReport respectively.

- **Oracle**: the model uses the reference summary length as the budget guidance.
- **Model-based**: the model uses the length of the summary produced by an advisor model (BART for GovReport and BigBird for PubMed/arXiv) as budget guidance.

To fairly compare the different models, we use an additive budget correction so that the average number of tokens in system summaries is close to the average length of reference summaries from the validation set (Table 3.4). The average summary lengths for each model is presented in Tables 3.6 and 3.8. Also, we provide ROUGE scores for varying budget guidance values in Section 3.3.2.6.

Fixed budget The simpler version of FACTORSUM is only guided by a fixed budget and its performance is competitive with most baselines, including PEGASUS and BigBird (Table 3.6, ”no content guidance”). It is important to note that the intrinsic importance model is based on a `bart-base` model with 139M parameters, which is 4 times smaller than PEGASUS and BigBird.

Oracle budgets The second section of Table 3.6 shows that for FACTORSUM with no content guidance, having access to the oracle lengths improves the scores by about 2, 1.6, and 1 ROUGE-1 on PubMed, arXiv, and GovReport respectively. We observe a similar effect with content guidance (third section of Table 3.6). These results agree with our hypothesis that the impact of budgets on ROUGE is significant and often larger than the differences between different modeling approaches.

Model-based budgets One may argue that inferring summary lengths is part of the task. Thus, we also test how summary lengths provided by BART and BigBird affect the summarization performance. For all datasets, we observe that using model budget guidance is detrimental to the scores compared to fixed budget guidance (second and third section of Table 3.6). These results suggest that summary lengths are hard to predict from the source documents, and highlights the potential benefits of divorcing content selection and budget optimization.

3.3.2.2 Effects of Content Guidance

We also examine the impact of different types of content guidance. The first sort of guidance, *Lead*, takes the first k sentences from the source article. We choose the lowest k so that the guidance text has at least the number of words as the fixed target budget for each dataset (see Section 3.3.2.1). This content guidance improves the scores by ∼0.8 (PubMed and arXiv) and ∼0.9 (GovReport) ROUGE-1 points over the model without guidance. Notably, FACTORSUM with *Lead* guidance achieves 48.05 ROUGE-1 on arXiv and 59.67 ROUGE-1 on GovReport *without relying on predictions from strong baselines*. We can further improve performance by providing content guidance from BigBird and BART, leading to strong performance on all datasets (Table 3.6). From these empirical results, we conclude that content guidance is a simple and effective method to turn strong sequence-to-sequence baselines into more flexible summarization systems. It should be possible to add more types of guidance to adapt the summaries to specific needs such as topic coverage.

3.3.2.3 Ablation Study

Our method comprises two main components, a document and summary views and a ranker extracting the most salient information from those summaries using budget guidance. To shed some light on the contribution of each component to the overall performance, we conduct two ablation analyses. First, to better understand the inherent potential of the summary views, we feed them to the FACTORSUM ranker with reference (oracle) content guidance. Having reference as content guidance serves as an upper bound for what we can achieve using the summary views. We do the same to an ensemble of PEGASUS and BigBird, which is a concatenation of their summaries, for comparison.

Second, to understand the importance of the FACTORSUM ranker (see Algorithm 1), we replace it with TextRank (Mihalcea and Tarau, 2004), a prominent algorithm for extractive summarization. Having two input variants (summary views and PEGASUS and BigBird ensemble) and three ranker variants (TextRank, FACTORSUM - no content guidance, and FACTORSUM - reference content guidance) results in six models. We use fixed-length guidance for all of the models.

Table 3.7 shows our results on the PubMed and arXiv datasets. We observe significantly higher results for FACTORSUM - reference content guidance) when applied on summary views, compared to PEGASUS and BigBird Ensemble Summaries. This

Ranker	PubMed			arXiv		
	R-1	R-2	R-L	R-1	R-2	R-L
PEGASUS + BigBird Ensemble Summaries						
TextRank	43.93	18.33	38.40	44.15	16.89	37.32
FactorSum	45.38	**19.43**	41.49	45.30	17.60	40.38
FactorSum-Oracle	48.90	21.81	44.76	49.34	20.13	43.99
Summary Views						
TextRank	42.10	16.71	37.54	42.66	16.41	37.70
FactorSum	**45.41**	18.66	**41.63**	**47.22**	**18.60**	**42.61**
FactorSum-Oracle	51.75	23.31	47.53	53.51	22.94	48.29

Table 3.7: ROUGE F1 scores on the test sets for the ensemble experiments. We compare summary predictions given by the concatenation of PEGASUS and BigBird summaries against summaries derived from FactorSum summary views. We use two sentence rankers: an unsupervised TextRank baseline and FactorSum extrinsic importance ranker. FactorSum and FactorSum-Oracle use no content guidance and reference summary guidance, respectively. All models use *fixed budget* as described in Section 3.3.2.1. Best non-oracle results are **bold-faced**. Underlined results are statistically equivalent to the best scores ($p < 0.05$).

shows that when both inputs reach their full potential under ideal guidance, summary views are superior to an ensemble of the two strong baselines, thus containing more salient information. We observe a similar pattern when using FactorSum - no content guidance as a ranker, showing that summary views serve as a better input to a more realistic ranker. In addition, we notice that for both types of input, FactorSum - no content guidance significantly outperforms TextRank, thus contributing to FactorSum overall performance.

3.3.2.4 Domain Adaptation

Our last experiment, unusual in the summarization literature, aims to test if a model trained on PubMed/arXiv/GovReport performs well when applied to out-of-domain (OOD) samples. Our intuition is that FactorSum should adapt well to OOD budget distributions, whereas the intrinsic model captures domain-specific patterns with less influence from length and content position noise.

The adaptation performance for similar domains (PubMed and arXiv) is much higher than summarizing GovReport documents when trained on scientific articles. FACTORSUM outperforms end-to-end baselines in all cases, especially when there is a large gap in average summary lengths between the domains. However, it can still achieve significant improvements on arXiv, for which all models output summaries with similar average lengths.

Training \ Evaluation	PubMed				arXiv				GovReport			
	R-1	**R-2**	**R-L**	**Len**	**R-1**	**R-2**	**R-L**	**Len**	**R-1**	**R-2**	**R-L**	**Len**
End-to-end baseline (BigBird for PubMed and arXiv; BART-large for GovReport)												
PubMed	45.48	19.92	41.81	185	42.33	15.16	38.09	161	19.35	3.57	18.10	222
arXiv	39.47	14.95	35.77	177	46.15	18.60	41.46	164	16.61	2.25	15.09	352
GovReport	37.18	11.10	33.96	203	35.11	8.94	31.67	203	52.82	19.12	49.99	596
FACTORSUM - fixed budget, no content guidance												
PubMed	45.41	18.66	41.63	206	44.61	15.88	40.16	165	42.49	15.07	39.92	350
arXiv	44.40	16.87	40.51	209	47.22	18.60	42.61	165	**48.75**	**18.07**	**45.94**	414
GovReport	39.67	12.63	35.37	213	38.34	10.74	33.72	167	58.77	23.99	55.19	650
FACTORSUM - fixed budget and content guidance (BigBird guidance for PubMed and arXiv; BART-large guidance for GovReport)												
PubMed	47.50	20.33	43.76	205	**46.29**	**17.13**	**41.86**	166	42.24	15.03	39.68	344
arXiv	**45.87**	**18.10**	**42.02**	210	49.32	20.27	44.76	165	48.65	18.03	45.85	410
GovReport	41.27	14.01	37.10	211	40.00	11.85	35.47	176	60.10	25.28	56.65	648

Table 3.8: ROUGE F1 scores and average words per summary for the domain adaptation experiments. Models trained on PubMed, arXiv, and GovReport samples (rows) are used to summarize articles from the other dataset test splits (columns). The choice of budget guidance values is described in Appendix 3.2.2.1. Shaded scores are in-domain results from Table 3.6. Underlined results are statistically equivalent to the best cross-domain scores ($p < 0.05$).

Our most important finding is that the ROUGE scores are not as severely affected as expected in this cross-domain setting. Notably, FACTORSUM trained on PubMed without content guidance achieves 44.61 ROUGE-1 on arXiv, outperforming PEGA-

SUS trained in-domain. When guided by BigBird summaries (also OOD), FACTORSUM scores 46.29 ROUGE-1 on arXiv, also outperforming BigBird trained in-domain. Similar results are observed for models trained on arXiv and evaluated on PubMed. On GovReport, FACTORSUM can produce much longer summaries that cover more relevant content, which explains the substantial improvements in ROUGE scores over the end-to-end baselines. However, summaries generated by FACTORSUM trained on arXiv/PubMed cannot match the average length of 650 words produced by the in-domain version. The reason for this gap is that OOD models generate summary views with a lower variety of content, which are eliminated by the redundancy control described in Section 3.2.1.2.

3.3.2.5 Document Sampling Experiments

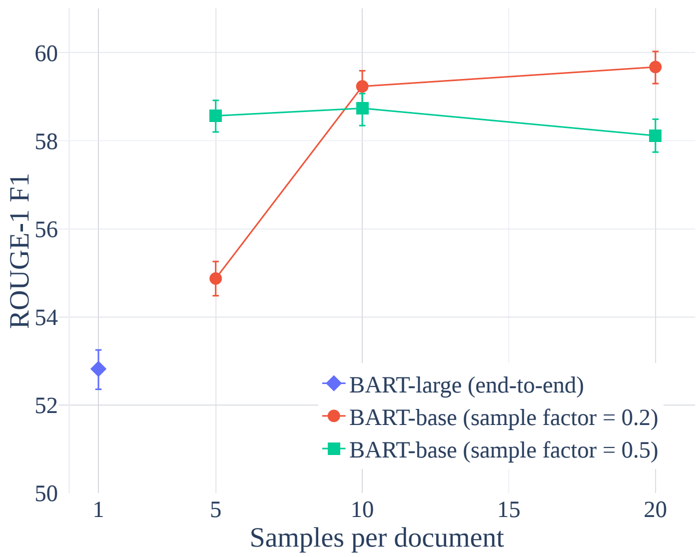

Figure 3.3: ROUGE-1 (F1) scores for different values of sampling factor (s_f) and number of samples per document (n_d), evaluated on the GovReport test set. BART-large is an end-to-end baseline, which is equivalent to $n_d = 1$ and $s_f = 1$.

The sampling factor (s_f) and number of samples per document (n_d) are important hyperparameters that affect the summarization performance and computation costs of FACTORSUM. By increasing the number of samples per document, the coverage of oracle sentences is also increased. Also, a smaller sampling factor will make each document view shorter and less prone to input truncation, at the cost of oracle sentence coverage.

For our experimental setup in Sections 3.3 and 3.3.2, we pick the sampling factor so that the number of sentences/tokens fit BART input limit (1024 tokens) with minimal truncation. The number of samples per document is chosen so that the coverage of oracle sentences in the original article is close to 100%, while keeping the resulting dataset size and training costs manageable. According to Table 3.9, a sampling factor $s_f = 0.2$ and number of samples $n_d = 20$ fullfil the requirements above for all datasets.

Sampling factor (s_f)	**Oracle coverage (%)** $n_d = 5$	$n_d = 10$	$n_d = 20$	**Average sentences**
		PubMed		
0.1	38.9	65.1	87.6	8.4
0.2	66.7	90.2	99.1	17.2
0.25	77.0	95.6	100	21.6
0.33	87.9	99.7	100	29.0
0.5	98.3	100	100	43.8
		arXiv		
0.1	37.9	63.8	88.1	27.6
0.2	64.6	88.4	99.0	55.7
0.25	75.6	93.3	99.7	69.8
0.33	86.8	97.7	100	93.2
0.5	96.5	99.9	100	140.0
		GovReport		
0.1	39.6	65.2	88.5	29.8
0.2	65.4	88.7	99.0	60.2
0.25	76.0	93.4	99.7	75.4
0.33	86.4	97.5	100	100.7
0.5	96.6	100	100	151.3

Table 3.9: Oracle sentence coverage and average number of sentences in sampled documents (validation sets) for different configurations of sampling factor s_f and samples per document n_d.

To further investigate the effects of different sampling strategies, we train FACTORSUM versions with sampling factor $s_f \in \{0.5, 0.2\}$ and samples per document $n_d \in \{5, 10, 20\}$. In Figure 3.3, we report evaluation results on the GovReport test set, which confirm the importance of oracle sentence coverage for the final summarization performance. Specifically, for a fixed $n_d = 5$, the model with $s_f = 0.2$ (65.4% oracle coverage) achieves 54.87 ROUGE-1 versus 58.57 ROUGE-1 for the version $s_f = 0.5$ (96.6% oracle coverage). A model with the same sampling factor $s_f = 0.2$ but using more samples per document achieves up to 59.67 ROUGE-1 ($n_d = 5$, 99.1% oracle coverage).

Finally, we observe that lower sampling factors achieve higher ROUGE scores, specially compared to the BART-large end-to-end baseline ($s_f = 1$, $n_d = 1$). These results suggest that working with shorter inputs is beneficial for summarization. We believe that this difference in performance is due to less truncation of the inputs.

3.3.2.6 Evaluation Results for Varying Budgets

In Figure 3.4, we provide results for varying budget guidance values on PubMed, arXiv, and GovReport test sets. For all datasets, there is a consistent improvement for content-guided summaries versus FACTORSUM without content guidance. We also note that BigBird content guidance leads to significant improvements on PubMed and arXiv but BART-large is statistically equivalent to Lead guidance on GovReport, which means there is still room for improvement in our end-to-end BART-large baseline.

3.4 Experiments with Legal Documents

In a second set of experiments, we evaluate FACTORSUM on a summarization benchmark for legal documents. We use CIVILSUM dataset (Malik et al., 2024), which is derived from civil cases heard by the Supreme Court of India and Indian High Courts from the country's independence (1947) up until the 2010–2011 calendar year. In comparison to related work such as IN-Abs (Shukla et al., 2022), CIVILSUM has significantly more samples and offers summaries with higher compression rates. A summary of the main dataset statistics compared to existing legal summarization datasets is provided in Table 3.10.

The judgment summaries exhibit clear patterns in the organization of content. First, a summary usually starts with a legal document reference that applied to the case. Another salient stylistic feature is that most of the paragraphs in the summaries include textual references to the relevant paragraphs in the judgments, which we hypothesize is an important signal for summarization modeling. To leverage this information, we

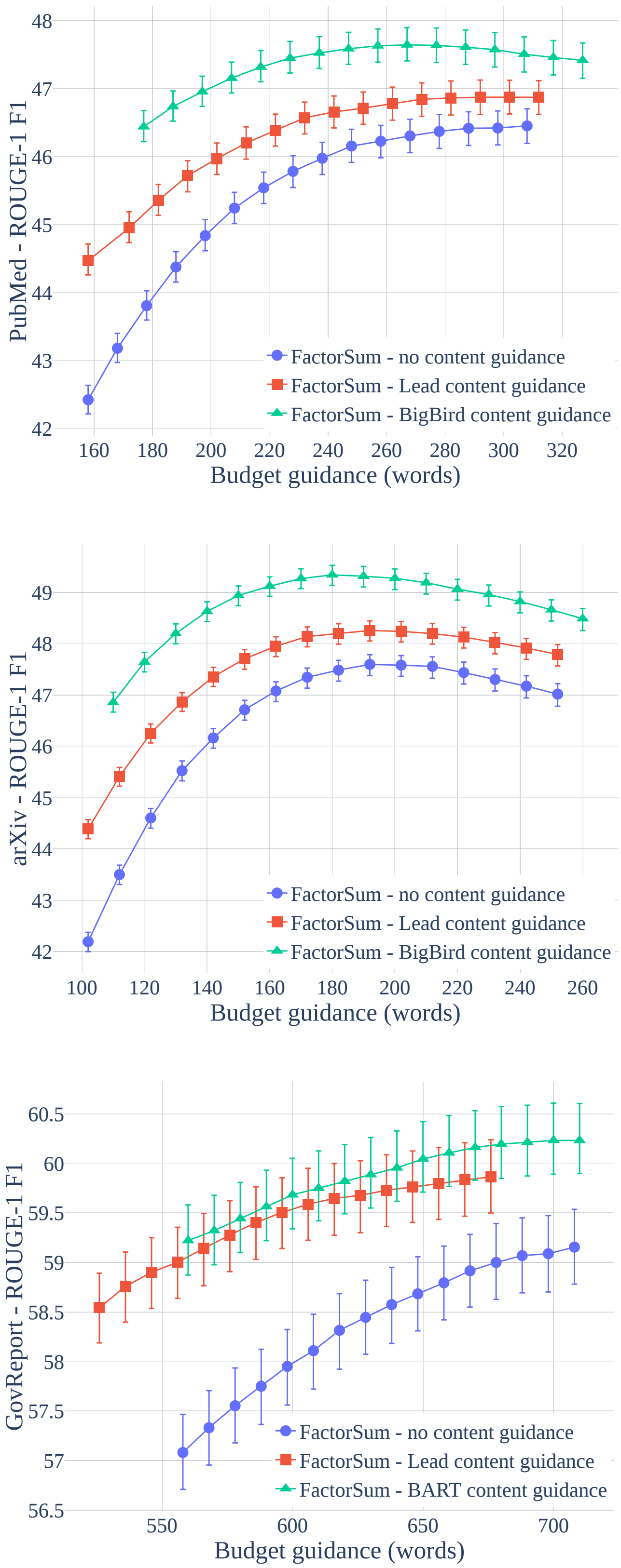

Figure 3.4: ROUGE-1 (F1) scores for different summary budget and content guidance computed on PubMed (top), arXiv (middle), and GovReport (bottom) test sets. Error bars indicate 95% confidence interval.

Dataset	# documents			Document		Summary	
	train	valid	test	words	sentences	words	sentences
IN-Abs	7,030	-	100	4,378	-	1,051	-
EUR-LexSum	3,447	689	459	11,864	340	1,011	32
CivilSum	21,015	1,168	1,167	2,123	90	104	4.5

Table 3.10: Statistics for legal summarization datasets, including IN-Abs (Shukla et al., 2022) and EUR-LexSum (Klaus et al., 2022). We report the number of documents per split and average length in words/sentences for the input documents and reference summaries.

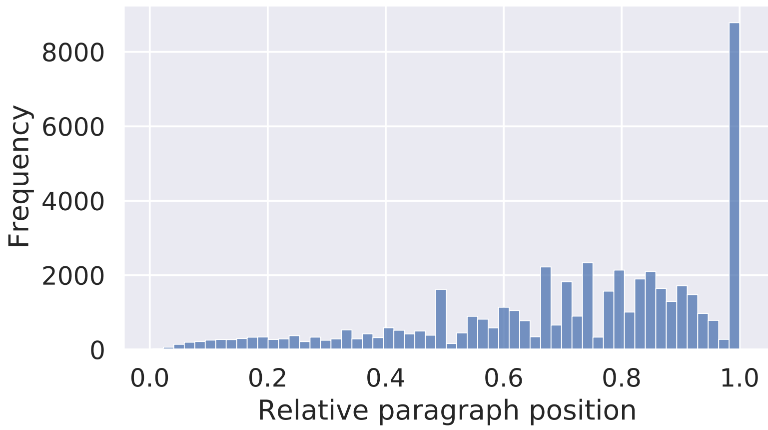

Figure 3.5: Distribution of relevant paragraph positions in the documents (training split) exhibiting tail bias.

use regular expressions to extract paragraph references of the form `[Paras 10, 15, 17]` (refer to Table 3.11 for an example). By applying this heuristic, we create a dataset where each paragraph in a judgment is labeled as 1 if mentioned in the summary, and 0 otherwise. Out of 23,350 documents in the dataset, 22,682 ($\approx 97\%$) contain at least one referenced paragraph in the reference summaries.

This paragraph reference information reveals an interesting insight about the information distribution in the dataset: most of the relevant content is located towards the end of the documents, a characteristic we refer as to *tail bias* (Figure 3.5). A consequence of this finding is that summarization systems that are biased towards leading information, as commonly seen in news summarization (Grenander et al., 2019; Zhu et al., 2021), should not perform well on our benchmark. We explore this tail bias in various settings in Section 3.4.1.

For Respondent No. 3. :- R.K. Malik, Advocate. A. Haryana Labour Department (Group A and Group B) Rules, 1987, Rules 9 and 7 - It is noted that the existing rules have been repealed and the Draft Service Rules framed and approved by Public Service Commission, but the draft rules have not been notified in Gazette and thus, cannot be considered as executive instructions. 1985(1) SLR 41, relied upon. **[Paras 7 and 8]**
B. Haryana Labour Department (Group A and Group B) Rules, 1987, Rules 7 and 9 - In relation to the constitutional validity of Article 16, seniority and acting promotion granted to the petitioner, it was established that the petitioner's promotion was regularised from 6.10.1986, but with no back salary. However, the respondent was appointed to the post with effect from 24.2.1984 and appointment regularised by Public Service Commission with effect from 11.1.1986, thereby proving that the respondent was senior to the petitioner. **[Paras 7 and 8]**

Table 3.11: Sample abstract from the CIVILSUM test set (ID = 716).

3.4.1 Experimental Setup

In this section, we describe summarization experiments with various types of architectures designed to process long documents. The objective is to provide a baseline performance assessment and to measure how the distribution of relevant information in the documents affects summarization performance. The baseline models are detailed as follows:

- **Random extractive baseline.** To get an estimate for the task difficulty, we randomly sample paragraphs from the documents up to 7% of the total words[10], subject to a minimum of 100 words. If the document has 100 or fewer words, the entire document is used as the summary.
- **Extractive oracle paragraphs.** We also obtain oracle extractive summaries that include only paragraphs mentioned in the reference summaries. The budget constraints are the same as the random extractive baseline described above. In contrast to the random baseline, there is no random sampling and the oracle paragraphs are added in the same order as they appear in the input document.

[10]We adjust to percentage of words so that the average number of words is similar to the average number of words in the reference summaries.

- **Longformer** (Beltagy et al., 2020), a transformer-based model that implements an attention mechanism that scales linearly with the input length, which makes it suited for the processing of long documents. We experiment with various input configurations, including documents truncated to the first (*lead*) 4,096 input tokens and truncated to the first (*lead*) and last (*tail*) 1,024 tokens. Additionally, we evaluate the model using only oracle paragraphs as inputs.

- **LLAMA-2** (Touvron et al., 2023), a transformer-based large language model with sizes ranging from 7 to 70 billion parameters. We leverage the fine-tuned "chat" version of LLAMA-2 and provide it with 4,096 tokens as input. In this work, LLAMA-2 or LLAMA-2-CHAT refers to the chat version.

Data pre-processing For FACTORSUM, we augment the document-summary pairs by creating pairs of *document views* and *summary views* as described in Section 3.2.1. To this end, we first perform sentence tokenization on both documents and summaries. Then, we uniformly sample 20% of the sentences in the documents to serve as document views for each one of the 21,013 documents in the training set, resulting in 420,260 shorter training samples. Using the same approach, we obtain 23,360 and 23,340 document-summary view pairs for the validation and test sets respectively. Apart from the usual input truncation in transformer models, no further preprocessing is performed for Longformer and LLAMA-2.

Training and inference details We use a BART-base (Lewis et al., 2020) checkpoint from HuggingFace[11] as a starting point to train FACTORSUM summary views generator, following the same training protocol and hyperparameters detailed in Section 3.2.2. The training is performed for 50,000 steps and we choose the checkpoint with the highest ROUGE-1 F1 score on the validation split. We employ a pre-trained LED-base checkpoint from HuggingFace[12] for Longformer and finetuned the model using a learning rate of 1×10^{-4} on 4 NVIDIA A100 GPUs with 128 effective batch size. The maximum length for summary is set to 256 tokens. All other training details align with those used for FACTORSUM. During inference, FACTORSUM performs the greedy optimization described in Section 3.2.1.2 using the same sampling hyperparameters as the training phase (20 document views per sample, each with 20% of the original sentences), with a budget constraint of 190 words per summary.

[11] https://huggingface.co/facebook/bart-base
[12] https://huggingface.co/allenai/led-base-16384

Model	**Paragraphs**	**Input Tokens**	**R-1**	**R-2**	**R-L**
Extractive	random	-	31.81	8.38	20.92
Extractive	oracle	-	33.12	10.11	22.01
LLAMA-2-CHAT-7B	lead	4096	37.12	12.55	25.43
LLAMA-2-CHAT-13B	lead	4096	36.73	11.63	25.61
LLAMA-2-CHAT-70B	lead	4096	37.39	12.61	25.74
FACTORSUM	lead	1024	40.33	15.74	31.98
FACTORSUM	tail	1024	41.80	16.53	33.30
FACTORSUM	oracle	1024	**46.51**	**20.67**	**37.07**
Longformer	lead	4096	44.80	18.37	36.85
Longformer	lead	1024	41.77	16.15	34.37
Longformer	tail	1024	44.35	17.65	36.32
Longformer	oracle	1024	44.10	18.53	36.77

Table 3.12: ROUGE F-1 scores for the legal summarization task. *lead* and *tail* refer to summaries focusing on the start and end of documents, respectively. The *oracle* variants leverage oracle paragraph information as described in Section 3.4.1.

Longformer uses a beam size of 3. For LLAMA-2-CHAT, we query the model with the prompt template: "`${document}.\n Write a summary of the text above in 4 sentences.`", and parse the model's completion as the candidate summary. For sampling hyperparameters, we use a value of 0.6 for temperature, and 0.9 for top-p filtering (Holtzman et al., 2020).

Evaluation metrics In all experiments, we report performance measured by ROUGE-1/2/L F1 score (Lin, 2004), following previous work in the summarization literature. ROUGE metrics measure the word overlap, bigram overlap, and longest common sequence between system-generated and reference summaries. Also, we report the results of human evaluation in Section 3.4.2.2.

3.4.2 Results and Discussion

In this section, we present our main findings based on the automatic evaluation metrics and a more qualitative evaluation by legal experts.

3.4.2.1 Automatic Evaluation

The results in Table 3.12 show a large gap in performance between a paragraph-based extractive summarizer and the abstractive approaches. This result suggests that summarizing more fine-grained, intra-paragraph abstractive processing is required to generate high-quality summaries. Still, we can verify that paragraph references are highly informative, improving the scores by $\approx$ 14 R-1 over the random extractive summarizer, and $\approx$ 6 R-1 over FACTORSUM with lead guidance. Moreover, Longformer's performance with only oracle paragraphs as input is approximately $\approx$ 2.4 R-1 points higher than when using the first 1,024 tokens (lead). It is comparable to the performance achieved with 4,096 tokens, providing additional evidence for the informativeness of the referenced paragraphs.

Another salient pattern is the higher informativeness towards the end of the documents, which can be verified by comparing the results of FACTORSUM with lead and tail guidance. Similarly, we observe a strong loss in Longformer performance by truncating the documents to the first 1,024 tokens (lead) compared to using 4,096 tokens, but the loss in performance is much smaller when using the last 1,024 tokens (tail). Finally, we observe that LLAMA-2 exhibits superior performance to extractive summarization approaches, yet remains inferior to other abstractive methods. We posit this stems from evaluating LLAMA-2 in a zero-shot setting without fine-tuning. As LLMs show promise in legal summarization, we leave finetuning LLAMA-2 with in-domain data for future work. Additionally, we observe that scaling LLAMA-2 parameters does not further improve performance. Nonetheless, the zero-shot results demonstrate LLAMA-2's capability to generate reasonable abstractive summaries without training. Further tuning could likely adapt the model to the target summaries' style and content.

3.4.2.2 Human Evaluation

In addition to automated measures like ROUGE, we designed a human evaluation to collect preference annotations. Our evaluators comprised two trained Indian lawyers familiar with the cases, who examined 25 randomly selected samples. For each given document, annotators were presented with summaries from three summarization systems: FACTORSUM, Longformer, and LLAMA-2-70B (the source for each summary was not revealed to the evaluators). They were first instructed to select their most preferred summary to replace a technical judgment abstract. Subsequently, they were prompted to choose the best summary according to the following criteria:

1. **Relevance:** assesses the extent to which the summary accurately captures and reflects the key points of the judgment. A relevant summary should focus on important aspects of the text while excluding irrelevant or extraneous details.

2. **Consistency:** refers to the uniformity and logical flow of information within the summary. It evaluates whether the summary maintains coherence and does not contradict itself or the facts in the original judgment.

3. **Fluency:** measures the readability and naturalness of the summary's language. It considers factors such as grammatical correctness, syntactic structure, and overall smoothness of expression.

4. **Coherence:** assesses the clarity and organization of ideas within the summary and evaluates how well it connects and relates different pieces of information. A coherent summary should present information in a logical sequence with clear transitions between ideas.

5. **Informativeness:** evaluates the extent to which the summary provides the essential details from the judgment. It considers whether the summary includes relevant content while avoiding redundancy or trivial information.

Regarding the results for the first question, the first annotator preferred the FACTORSUM, Longformer, and LLAMA-2 summaries 10, 9, and 6 times, respectively. The second annotator preferred them 8, 13, and 4 times. The inter-annotator agreement as measured by Cohen's kappa was 0.44, indicating moderate agreement. For the second question, the preferences were 6, 6, 13, and 6, 7, 12, respectively. The inter-annotator agreement was 0.52, again suggesting moderate alignment[13]. These results imply that for technical adequacy, summaries from supervised models like FACTORSUM and Longformer were preferred. However, considering overall summary quality, the LLAMA-2 summaries were favored. The mixed results confirm our expectations that supervised models are better at following the technical style and jargon of reference summaries, while LLAMA-2 produces more fluent, high-quality summaries in a linguistic sense.

[13]The relatively low agreement is due to the highly ambiguous nature of the task. This aspect could be improved by providing a more structured calibration rounds to the annotators, which would increase the experiment costs.

Reference abstract: Under the Uttar Pradesh Public Services (Reservation for Scheduled Castes, Scheduled Tribes and other Backward Classes) Act, 1994, Rule 3(4) for Recruitment Eligibility, the State Engineering Services (Special Recruitment) Examination is conducted for the recruitment of SC and ST candidates.', 'In this process, it is reasonable to fix a minimum aggregate of 40% marks as the standard for evaluating candidate suitability.', 'It was established that the petitioner did not obtain the required minimum marks and was subsequently not called for an interview, therefore, there is no justification for intervention. **[Paras 7, 8, and 9]**
Longformer summary: constitution of india, article 16 - u.p. public services (reservation for scheduled castes, scheduled tribes and other backward classes) act, 1994, section 3(4) - interview - suitability - petitioners who appeared in state engineering services (special recruitment) examination, 1995 conducted by respondent no. 2 and were not called for interview - it was held that the petitioners were unsuitable for interview as they could not secure the 40% minimum marks required for judging their suitability. however, the criterion was not deemed arbitrary and did not violate any laws. hence, the petition was dismissed. **[paras 7, 8 and 9]**
LLAMA-2 summary: The petitioners, who were candidates in a special recruitment examination for S.C. and S.T. candidates, filed a writ petition against the respondents, claiming that they were not called for an interview due to the respondents' illegal fixing of 40% minimum marks in aggregate. The Court found that the requirement of securing 40% minimum marks in aggregate was not arbitrary and did not suffer from any legal infirmity. Therefore, the Court dismissed the writ petition, finding that the respondents did not commit any error in not calling the petitioners for an interview. The petitioners did not obtain the required 40% minimum marks in aggregate, making them ineligible for the interview.
FACTORSUM - Lead content guidance: Constitution of India, Articles 14, 16 and 226 - Selection - Interview - Selection for the post of S . C. and S .T. candidates - Petitioners appeared in State Engineering Services (Special Recruitment) Examination, 1995 conducted by respondent No. 2 for recruitment of SC and ST candidates. T. candidates - The requirement of 40% marks in aggregate for assessing the suitability of a candidate for being called for interview is not illegal . **[Para 5]** P. Public Services (Reservation for Scheduled Castes, Scheduled Tribes and other Backward Classes) Act, 1994, Section 4 - Recruitment - Interview - Post of Lecturer - Petitioners, who were appointed as Lecturers, challenged the appointment of Respondent No. 2 as Lecturer after obtaining 40% marks in aggregate. Constitution of India, Articles 14 and 16 - U .

Table 3.13: Sample abstract and generated summaries from the CIVILSUM test set (ID = 8). References to paragraphs in the source document are shown in blue color.

In addition, the annotators observed that although the summaries were generally adequate and captured key points successfully, there were deficiencies in sentence construction ambiguity, erroneous interpretations of interest payment, and sporadic incompleteness. Furthermore, concerns were raised about the lack of conciseness and occasional omission of conclusions, which are crucial elements in summarizing legal judgments. In the sample summaries (see Table 3.13), evaluators highlighted specific issues, including the overuse of articles 14 and 16 of the Constitution of India without proper contextual relevance, a tendency to refer to party names instead of directly addressing the real issue, and the use of personal pronouns instead of maintaining an objective tone. Moreover, there was a lack of sufficient attention to the legal aspects of the issue, resulting in an incomplete and inadequate portrayal of the real issue from a legal standpoint. Overall, the evaluation suggests that the legal benchmark is challenging, and current summarization systems struggle to produce satisfactory summaries.

3.5 Related Work

Ranking and aggregating candidate summaries Recent work explores the idea of sampling summary candidates that are scored to form a final summary. The SimCLS (Liu and Liu, 2021) model generates several summary candidates using diverse beam search (Vijayakumar et al., 2016), which are ranked according to a learned evaluation function. Also, the Perturb-and-Select summarizer (Oved and Levy, 2021) uses similar ideas in a multi-document opinion summarization task. Instead of a diverse beam search, it performs random perturbations in the model inputs to generate candidates ranked according to a coherence model. Iso et al. (2021) presented COOP, a framework that improves the aggregation methods for multi-document opinion summarization representations by maximizing their input-output word overlap. Our approach differs from these models as our sampled summary views are not full summary candidates but partial views with important semantic content from the original document. Also, the SimCLS ranking uses intrinsic importance only (similarily with respect to the original document).

Divide-and-conquer approaches for long document summarization In the scientific domain, the DANCER model (Gidiotis and Tsoumakas, 2020) breaks a summarization task into multiple smaller sub-tasks, which share similar motivation to our intrinsic importance model. However, they use several heuristics to select specific sections of the papers (introduction, methods, results, and conclusion) while our sampling approach is domain agnostic. Recently, Mao et al. (2022) proposed DYLE, an *extract-then-generate* method that extract text snippets from chunks of the input document, obtaining state-of-the-art results on GovReport. Zhang et al. (2021) presented a multi-stage summarization method that generates coarse summaries for each document segment, which are then used to obtain a fine-grained summary. Cao and Wang (2022) add a learnable hierarchical bias term to the transformer attention mechanism, which allows the model to capture information about the structure of long documents.

Controllable summarization Controlling length in summaries has been addressed by leveraging positional encodings (Takase and Okazaki, 2019), a length-aware attention mechanism (Liu et al., 2022a), and optimization objectives that include a length constraint (Makino et al., 2019). Kikuchi et al. (2016) explore different length control techniques at the learning and decoding stages. Control of other attributes such as entity coverage and summary style were achieved with control tokens (Fan et al., 2018; He et al., 2022) and constrained Markov Decision Processes (Chan et al., 2021).

Similar to this work, GSum (Dou et al., 2021) uses content guidance to improve summary quality. However, we note that GSum's guidance is used as *input* of its sequence-to-sequence model, and shifts in guidance distribution would require further training. In contrast, FACTORSUM allows one to change the budget or content guidance without expensive retraining. Regarding evaluation, the best GSum variant achieves 45.09 ROUGE-1 on PubMed, whereas FACTORSUM achieves 45.41 R-1 without content guidance and 47.5 R-1 with content guidance. Finally, our sequence-to-sequence architecture is based on BART-base, thus requiring significantly fewer training parameters than GSum's dual BART-large encoders.

Legal summarization In the legal domain, previous work investigated the summarization of legislative texts, with a focus on US Congressional and California state bills (Kornilova and Eidelman, 2019) and topic modeling applied to multi-document summarization in the Brazilian lawmaking process (Silva et al., 2021). Further research on legal documents spanned diverse areas such as legal acts (Aumiller et al., 2022), legal cases (Bhattacharya et al., 2021; Elaraby and Litman, 2022; Ghosh et al., 2022; Shukla et al., 2022), debate dialogue (Duan et al., 2019), and European regulatory documents (Klaus et al., 2022).

3.6 Limitations

While our model requires fewer compute resources for training, the inference step is more expensive. For each document, $n_d = 20$ document views are sampled and n_d feedforward computations are performed for the intrinsic model (BART-base) before the greedy summary generation algorithm is applied. In our experiments, BART-base averages $0.27 \times n_d$ seconds per sample. The greedy generation adds in the worst case an average of 1.5 seconds per sample using a naive single-threaded implementation, which gives a total of 6.9 seconds per document. Fortunately, these computations are highly parallelizable, and more careful tuning of the number of views per document n_d would make the runtime similar to a single large neural model. For comparison, PEGASUS and BigBird-PEGASUS take on average 3.13 and 3.85 seconds per sample on a single GeForce GTX 1080 Ti GPU (batch size = 4).

Furthermore, our model exhibits factuality issues, which is a common challenge in abstractive summarization. However, by designing the model with multiple summary views and easily customizable constraints, with careful application of domain knowledge, the hallucination problems can be significantly mitigated. For instance, in the application of FACTORSUM in the financial domain, multiple constraints to retain factuality and reduce ambiguity were developed jointly with the Actelligent team by utilizing their domain knowledge. This resulted in the successful mitigation of nonfactual, ambiguous, or inconsistent content in the summaries.

Finally, the greedy optimization based on independent summary views leads to summaries with low discourse quality. This issue is easily observed in sample paper summaries (e.g., Table A.8), where there is a repetition of sentences starting with *In this paper*, and an unnatural ordering of content. One possible way to address this issue is by using a strong summarizer (e.g., a large language model) that is able to synthesize information from the entire document while using the summary views from FACTORSUM as context. This approach is investigated in Chapter 4.

3.7 Conclusion

In this work, we embrace the idea that general-purpose summary is an elusive goal and that contextual factors such as preferences for summary lengths are essential in the design of summarization systems (Jones, 1998). We propose a framework to separate budget decisions from selecting important content in the document using neural sequence-to-sequence models as building blocks. Our results suggest improved performance in both in-domain and cross-domain summarization of long documents. However, our human evaluation experiments in the legal domain reveal the limitations of our method in terms of summary fluency. In Chapter 4, we address the issue of fluency using large language model summarizers guided by keywords derived from FACTORSUM summaries. In Chapter 5, we investigate the effects of domain-specific extrinsic guidance that would encourage the summaries to cover aspects of interest, e.g., methods, results, and conclusions in scientific papers.

Chapter 4

Controlling Extrinsic Factors with Large Language Models

Large language models (LLMs) are an emerging technology that shows promising results across a wide range of language understanding problems. In contrast to the summarization method we presented in the previous chapter, FACTORSUM, LLMs often are more costly to fine-tune and deploy for task-specific inference. However, these models can generate text with remarkable quality across multiple domains, which can address the issues with lack of fluency of summaries produced by FACTORSUM. In this chapter, we investigate the application of large language models on scientific summarization tasks. We identify key stylistic and content coverage factors that characterize different types of summaries such as paper reviews, abstracts, and lay summaries. By controlling stylistic features, we find that non-fine-tuned LLMs outperform humans in the MuP review generation task, both in terms of similarity to reference summaries and human preferences. Also, we show that we can improve the controllability of LLMs with keyword-based classifier-free guidance (CFG) while achieving lexical overlap comparable to strong fine-tuned baselines on arXiv and PubMed. However, our results also indicate that LLMs cannot consistently generate long summaries with more than 8 sentences. Furthermore, these models exhibit limited capacity to produce highly abstractive lay summaries. Although LLMs demonstrate strong generic summarization competency, sophisticated content control without costly fine-tuning remains an open problem for domain-specific applications.

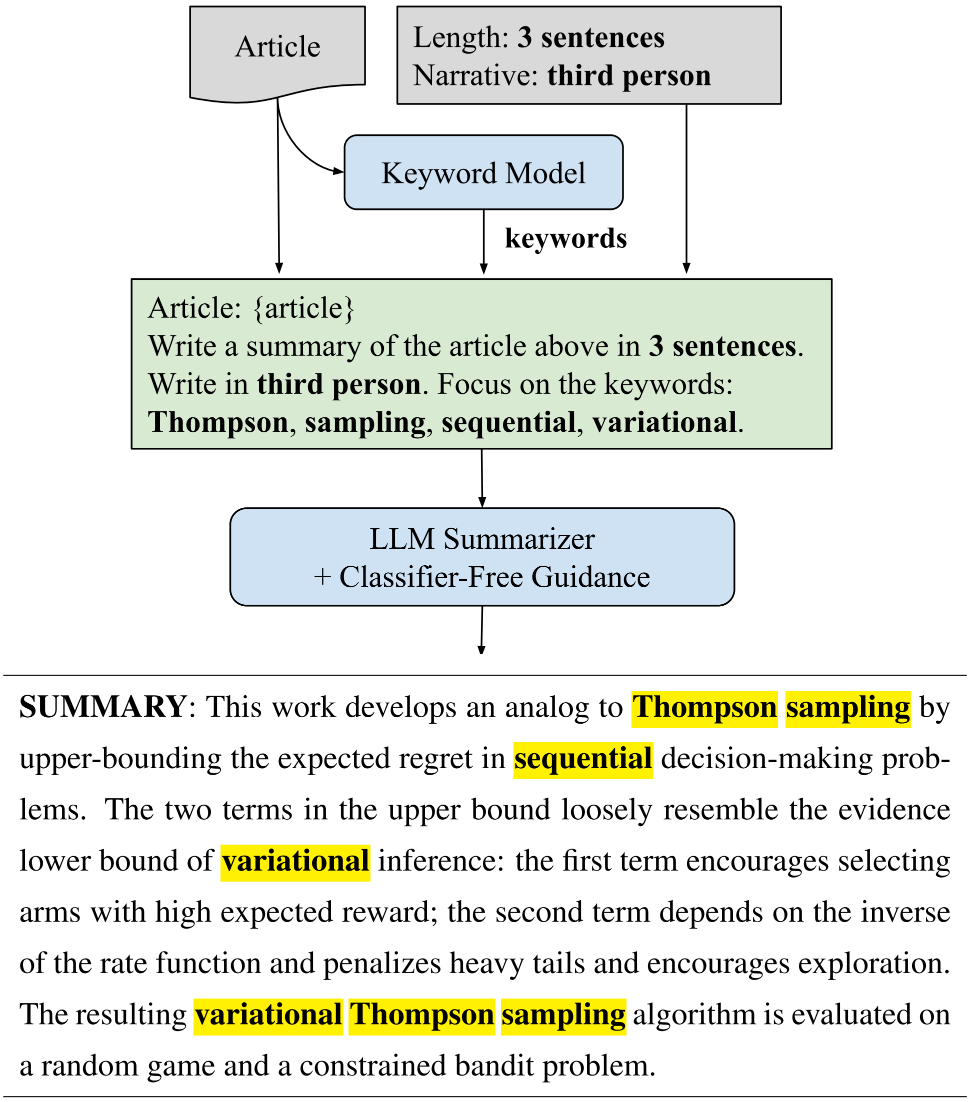

Figure 4.1: An overview of our controllability experiments. We expose LLM summarizers to prompts conveying communicative intentions related to *conciseness*, *narrative perspective*, and *keywords* inferred by a *keyword model*. Then, we measure how generated summaries adhere to those intentional targets.

4.1 Introduction

Recent work on the evaluation of large language models (LLMs) has shown unprecedented performance on diverse language generation tasks, even in zero-shot settings (Clark et al., 2021). Specifically in text summarization, Goyal et al. (2022) found that human evaluators prefer summaries produced by GPT-3 over strong supervised baselines. In a similar experiment testing ten large language models, Zhang et al. (2024) concluded that instruction-tuned models performed on par with human freelance writers. Additionally, their results suggest that the evaluation of news summarization is hindered by the low quality of reference summaries. Hence, the conventional summarization paradigm based on fine-tuning and evaluating on reference datasets is questionable in the face of increasingly competent language models.

In this chapter, we aim to fill this gap by posing the following research questions: 1) How do LLM summaries compare to human-generated paper summaries? 2) To what extent can LLM summarizers be controlled to fulfill different goals of scientific communication? In both cases, we focus on the zero-shot setting, since the costs of using such large models are often prohibitive compared to small-scale models such as FACTORSUM introduced in Chapter 3. First, there is a higher cost involved in fine-tuning large language models. Even when using parameter-efficient techniques (Hu et al., 2022; Lester et al., 2021), the memory requirements for fine-tuning the smallest state-of-the-art LLMs (Touvron et al., 2023) are vastly superior compared to the BART-base model used in FACTORSUM. The same problem arises during inference, making the infrastructure requirements for deploying such fine-tuned LLMs impractical in many cases. Most importantly, curating high-quality reference summaries is expensive in most real-world settings, which makes the zero-shot setting quite appealing.

To address the first question, we design an experiment to compare human and machine summarizers, in which human summaries are judged in terms of their similarity with respect to another set of human-written reference summaries. By evaluating the lexical overlap with reference summaries from the multi-perspective scientific summarization dataset - MuP (Cohan et al., 2022), we find that human reviewers achieve lower ROUGE scores (Lin, 2004) compared to LLMs. The high quality of LLM summaries is confirmed by a human judgment experiment, where machine-generated summaries are preferred in 83% of the instances.

The strong preference for LLM summaries hints that the variability in communicative intentions (Giulianelli et al., 2023; Andreas, 2022) outweighs the usual quality criteria such as coherence, fluency, and relevance (Fabbri et al., 2021). In fact, an inspection of human summaries for the same article reveals arbitrary decisions related to conciseness and coverage of scientific aspects. We argue that those subjective decisions affect the perceived quality of summaries and that evaluation protocols should consider the *adaptability* of summarizers to diverse contexts (Jones, 1998).

Motivated by the experiment on the MuP dataset and previous literature on controllable summarization (He et al., 2022), we elect three intentional aspects: conciseness, narrative perspective (first or third person), and keyword coverage. For each of those aspects, we define two elements: *intention prompts*, and *intention control metrics*.

Intentional prompts are designed to specify the task and to elicit a given intentional behavior from LLMs (e.g., *summarize this article in **5 sentences***). Finally, the control metrics assess the intentional alignment of prompts and the generated summaries.

To investigate the second research question, we evaluate our intentional prompts on the tasks of abstract generation (Cohan et al., 2018) and lay summarization (Goldsack et al., 2022), using GPT-3.5 and LLAMA-2 (Touvron et al., 2023). We find that both models can follow conciseness and narrative perspective intentions accurately to generate short abstracts for arXiv and PubMed. Furthermore, by using keyword-based intentional prompts and classifier-free guidance (Sanchez et al., 2023), we can direct to the lexical content of summaries leading to ROUGE scores comparable to strong supervised baselines.

However, the eLife lay summarization benchmark presents a harder challenge for LLMs. Besides being highly abstractive, these lay summaries are much longer (18 sentences on average), and have a particular concept distribution emphasizing research background. Our controllability results show that intentional prompts achieve limited success in replicating these characteristics.

Overall, our experiments spanning four benchmarks indicate that LLMs are effective summarizers for long scientific documents, both in terms of lexical alignment and human preferences. However, when summarization tasks deviate from the training distribution (e.g., lay summarization), LLMs cannot consistently match the features of human summaries. Thus, LLMs do not usher the "death" of summarization (Pu et al., 2023) but the transition from reference-based evaluation to more nuanced, domain-specific evaluation protocols, as envisioned by Jones (1998) 25 years ago.

4.2 Guiding Summarizer Intentions

As uncertainty in intentionality is intrinsic in language generation (Giulianelli et al., 2023), we argue that summarization performance relates to the capacity of a system to adapt its behavior given not only the source documents but also the target communicative intentions. We investigate language model intent adaptability by changing the summarization context via prompting, and measuring how it affects the perceived intention in the summaries (see Figure 4.1 for an example). Let P_θ be a language model parameterized by θ. Then, we define summarization as a conditional sequence generation:

$$S \sim P_\theta(\cdot \mid D, p_{I_1}, \dots, p_{I_N}), \tag{4.1}$$

where D is an input article and $p_{I_1}, \dots, p_{I_N}$ are prompts inducing the intentions $I_1, \dots, I_N$. In this work, we consider three types of intentions that are important to adapt scientific summaries to different goals, namely *conciseness*, *narrative perspective*, and *keyword coverage*:

- **Conciseness**: as one of the most important intentional factors in summarization, conciseness defines the compression rate between the source document and the summary. We are interested in the ability of LLMs to follow specific conciseness instructions ranging from short abstracts (6-8 sentences) to longer lay summaries (14 or more sentences).

- **Narrative Perspective**: depending on the perspective of the summarizer, the summary is written using first or third-person narrative. In our experiments, we consider that paper abstracts from arXiv and PubMed use first-person narrative, and other summaries such as paper reviews from MuP (Cohan et al., 2022) and lay summaries from eLife (Goldsack et al., 2022) use third person narrative.

- **Keyword Coverage**: by guiding the coverage of keywords, we can indirectly manipulate the level of abstractiveness of a summary and favor simplified language, which is particularly relevant for lay summarization. Furthermore, when reference summaries are available, keyword guidance provides a mechanism to guide LLM generation using relatively small *keyword models* (Section 4.2.1.2) that demand fewer resources to fine-tune.

Since we specify intention guidance via prompts, we have no control over how the language models trade off their unconditional summarization behavior and the intention instructions. To address this issue, we modify the decoding process in Eq. 4.1 to include classifier-free guidance (CFG) weighting (Ho and Salimans, 2022; Sanchez et al., 2023):

$$\hat{P}_\theta(S|D, p_I, p_\varepsilon) \propto \frac{P_\theta(S|D, p_I)^\gamma}{P_\theta(S|D, p_\varepsilon)^{\gamma-1}}, \tag{4.2}$$

where $\gamma \geq 1$ is the guidance strength and p_ε is a “non-intentional” summarization prompt:

```
Write a summary of the article above.
```

When $\gamma > 1$, the next-token probabilities are changed so that the generated summary is closer to the target intentions than its default summarization behavior elicited by the prompt p_ε.

To measure the adherence of summaries to intention prompts, we define reference-free *intention control* metrics $k_I(y_I, S)$ that gauge the intentional alignment of S with respect to a target value y_I. In the next sections, we detail the intention prompts p_I and control metrics k_I used in our experiments.

4.2.1 Intention Prompt Templates

The next component of our methods refers to the way we prompt language models to induce the intentions described above. The general prompt template is defined as follows:

```
Article: {text}
{p_I1 ... p_IN}
```

where the placeholder {text} denotes the input article and $\{p_{I_1} \dots p_{I_N}\}$ refers to the concatenation of intention prompts, namely conciseness, narrative perspective, and keyword coverage.

4.2.1.1 Style Intention Prompts

Conciseness prompt We define an intention prompt for a target number of sentences $y_{\text{conciseness}}$:

```
Write a summary of the article above in {y_conciseness}
sentences.
```

We choose to specify the target in sentences as it was found to be effective in previous work (Goyal et al., 2022) and in our own experiments.

Narrative perspective prompt For paper abstract generation tasks, we use the following prompt to instruct the model to write in the usual first-person plural perspective:

```
Write in first person ''we'' when applicable.
```

For other tasks, we do not have to prompt the model to use third person voice, as we observe this is the default behavior of the language models we evaluate (see Tables 4.7 and 4.10).

POS Tag	Description
FW	Foreign word
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
SYM	Symbol
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present

Table 4.1: Part-of-speech tags used to filter keywords for summary guidance. Tag descriptions are taken from the Penn Treebank Project (Marcus et al., 1993).

4.2.1.2 Keyword Coverage Prompt

The last prompt type instructs the model to focus on a collection of keywords to generate the summary:

```
Focus on the following keywords: {y_keywords}.
```

where the $\{y_{\text{keywords}}\}$ placeholder indicates a comma-separated list of terms, which are provided by a *keyword model*.

Keyword model In our experiments, we use pre-trained encoder-decoder summarizers as keyword generators. Specifically, we use FactorSum (Fonseca et al., 2022) and Bart (Lewis et al., 2020) for abstract generation and lay summarization respectively. Using these models, we generate summaries for each document in the evaluation dataset. Then, we extract part-of-speech tags for tokens in summaries using the NLTK library (Bird et al., 2009) and keep as keywords only nouns, verbs, adjectives, foreign words, and symbols. The full list of keyword POS tags is presented in Table 4.1.

4.2.2 Intention Control Metrics

In this section, we identify the controllability metrics k_I for stylistic features (conciseness and narrative perspective) and keyword coverage. Although we present summary-level definitions, in our experiments we report dataset-level metrics, that is, their average over all evaluation samples.

Conciseness In this work, we measure the conciseness of a summary S by counting the number of sentences $|S|$. Then, the conciseness controllability metric k_I for a summary S and a target number of sentences y_I is defined by:

$$k_{\text{conciseness}}(y_I, S) = \texttt{abs}(y_I - |S|), \tag{4.3}$$

where `abs` is the absolute difference function.

Narrative perspective We define the narrative perspective metric based on a text classifier $f_{\text{narrative}}(s)$ that maps each summary sentence $s \in S$ to the label $\hat{y}_I \in \{\texttt{first}, \texttt{third}\}$. Then, we defined the summary-level perspective controllability as the percentage of sentences written in the target perspective y_I:

$$k_{\text{narrative}}(y_I, S) = \frac{100}{|S|} \sum_{s_i \in S} \mathbb{1}[f_I(s_i) = y_I], \tag{4.4}$$

where $|S|$ is the number of sentences in S. For the classifier $f_{\text{narrative}}(s)$, we use a simple heuristic where the sentence is classified as first-person perspective if 1) it starts with a first-person pronoun `FPP` $\in \{$`"we"`, `"our"`$\}$ or 2) if it contains the pattern `", FPP"`, i.e., a comma followed by a `FPP`.

Keyword Coverage We also measure the coverage of a target set of keywords y_I in summaries using the ROUGE-1 recall metric (Lin, 2004):

$$k_{\text{keywords}}(y_I, S) = \texttt{ROUGE-1}_{\text{recall}}(y_I, S). \tag{4.5}$$

4.3 Experimental Setup

In this section, we present the experimental settings spanning three styles of scientific summaries: abstracts, reviews, and lay summaries.

4.3.1 Datasets

Multi-perspective scientific summarization The MuP dataset (Cohan et al., 2022) is a corpus of summaries extracted from publicly available scientific peer reviews[1] capturing multiple summarization perspectives for a single document. In our experiments, we use 1,060 papers and review summaries from the validation set, covering topics primarily from the AI, Machine Learning, and Natural Language Processing fields.

Abstract generation We use arXiv and PubMed (Cohan et al., 2018), which are two large-scale benchmarks for abstract generation. In our experiments, we report results for 1,000 random samples from each dataset test set.

Lay summarization We also evaluate LLMs on the task of lay summarization using the eLife dataset (Goldsack et al., 2022). In contrast to reviews and abstracts, these summaries are much longer (around 18 sentences on average), more abstractive, and exhibit a strong bias towards background content, which makes them easier to parse by non-specialists. We report results on the 241 samples from the eLife test set.

4.3.2 Models

In this section, we describe the the models used in the summarization experiments. Refer to Appendix A.2 for additional details and generation parameters.

Supervised baselines For the abstract generation task, we use BIGBIRD (Zaheer et al., 2020), a transformer-based model that uses a sparse attention mechanism to handle long input sequences. Also, we include FACTORSUM, the model we introduce in Chapter 3. For abstract generation, we use the FACTORSUM checkpoints provided by Fonseca et al. (2022)[2]. We perform inference with a fixed budget of 6 and 8 sentences for arXiv and PubMed respectively, and content guidance from BIGBIRD summaries.

For lay summarization, we fine-tune a BART-base model (Lewis et al., 2020) on the 4,346 training samples of the eLife dataset (Goldsack et al., 2022). We use the Hugging Face summarization training script (commit `5c67682`[3]) with the hyperparameters listed in Table 4.2. We choose the checkpoint with the best ROUGE-2 (F1) score on

[1] https://openreview.net/

[2] https://github.com/thefonseca/factorsum

[3] https://github.com/huggingface/transformers/commits/5c67682b169576c4859700d551090ff79d450a9a/examples/pytorch/summarization/run_summarization.py

Checkpoint	bart-base
Epochs	30
Batch size	4
Optimizer	Adam (β_1=0.9; β_2=0.999)
Learning rate	5×10^{-5}
Weight decay	linear
Max. target length	1024
Validation metric	ROUGE-2 (F1)

Table 4.2: Fine-tuning parameters for BART-base on the eLife summarization dataset.

the the validation set. The fine-tuning process takes about 8 GPU hours (2 Nvidia GTX 1080 12GB GPUs).

LLAMA-2 A collection of pre-trained and fine-tuned large language models (LLMs) ranging in scale from 7 billion to 70 billion parameters (Touvron et al., 2023). Unless otherwise stated, we report results for the *chat* variant with 7B parameters (16-bit floating point), with 4,096 maximum context tokens using nucleus sampling (Holtzman et al., 2020) with temperature 0.8 and $p = 0.95$. The intention prompts described in Section 4.2.1 are wrapped into the following model-specific instruction-tuning prompt: `[INST] {instruction} [/INST]`, where `{instruction}` is an intention prompt.

GPT-3.5 A proprietary model based on INSTRUCTGPT (Ouyang et al., 2022). We use the model version `gpt-3.5-turbo-0301` via the chat completion API endpoint[4]. Although we do not have access to the implementation details behind commercial APIs, they were extensively studied in recent work (Zhang et al., 2024; Goyal et al., 2022). This model version is trained on data up to September 2021, and we generate summaries in December 2023.

4.3.3 Evaluation Metrics

In addition to the intention control metrics defined in Section 4.2.2, we report ROUGE (Lin, 2004) and the following abstractiveness metrics: 1) novel n-grams (percentage of n-grams in summary that are absent in the source document) and 2) the MINT

[4] https://platform.openai.com/docs/api-reference/chat

abstractiveness score by Dreyer et al. (2023)[5]. For the lay summarization task, we also report the Flesch-Kincaid Grade Level (FKGL) readability score (Kincaid et al., 1975).

4.4 Results and Discussion

In this section, we discuss our experimental results comparing human and LLM-generated summaries (Section 4.4.1), and the controllability experiments on the abstract generation and lay summarization tasks (Section 4.4.2).

4.4.1 Human versus LLM Summaries: Reviewer Perspectives

We leverage the MuP dataset (Cohan et al., 2022) for a controlled comparison between human and machine summaries. First, we select documents from the validation set with more than one human-written summary and randomly choose one of the summaries to serve as a reference and a second human-written summary as a fictitious system summary. This setting puts humans in a similar condition as usual reference-based summarization benchmarks, that is, humans are evaluated on their ability to *guess* reference summaries.

As machine summarizers, we evaluate LLAMA-2 and GPT-3.5. To minimize the confounding factors related to summary length, we employ conciseness prompts (Section 4.2.1.1) to instruct the LLMs to generate summaries with a number of sentences such that the number of generated tokens approximates the human summaries[6]. In Table 4.3, we report metrics comparing LLM and human summaries for 1,060 samples from the MuP validation set.

In addition to automatic evaluation, we perform human evaluation based on 30 random samples of human and LLM-generated summaries. Similarly to Goyal et al. (2022), we ask evaluators to perform blind A/B judgments based on the paper abstract and a triplet consisting of one human and two LLM summaries. Evaluators are tasked to elect from the triplet which summary (or summaries) they judge is the best/worst alternative to the paper abstract. For each answer, the evaluators need to provide a short justification (refer to Table 4.4 for an example). Before performing the tasks, evaluators were presented with the following short description of the tasks:

[5] The MINT abstractiveness score (Dreyer et al., 2023) measures both contiguous and non-contiguous overlapping of text spans, combining ideas from previous metrics such as *density* (Grusky et al., 2018) and *perfect fusion* (Durmus et al., 2020).

[6] We divide the number of tokens in each human summary by the average number of tokens per sentence generated by the LLMs, resulting in different conciseness targets for each document.

Metric	Ref	Human	LLAMA	GPT-3.5
ROUGE-1	100	35.55	**37.35**	36.39
ROUGE-2	100	8.12	**9.20**	8.51
ROUGE-L	100	20.55	**24.47**	24.02
Average Tokens	115	113	112	113
% Third Person	99.8	99.6	99.3	1.0
% Novel bigram	66.4	**67.0**	39.9	34.1
MINT (Dreyer et al., 2023)	0.79	**0.92**	0.54	0.46

Table 4.3: Metrics comparing human-written, LLAMA-2-7B, and GPT-3.5 summaries to reference reviews (Ref) from the MuP validation set. We report average tokens per summary and the percentage of sentences using *third person* perspective. *Novel bigram* measures the percentage of novel bigrams in summaries. The MINT abstractiveness score (Dreyer et al., 2023) ranges from 0 to 1.

The goal of this study is to evaluate machine-generated summaries of scientific articles. Each reference article abstract will be presented with 3 alternative summaries. Your task is to identify which of the alternatives is the best according to your personal preferences and experience reading papers. The criteria may include good coverage of the abstract content, factuality issues, and linguistic fluency, among others. It is expected that in some cases the alternative summaries cover details that cannot be verified based on the reference abstract. You can disregard the factuality of those details if you consider it appropriate.

Each summary triplet is annotated by two graduate students specializing in NLP.

ROUGE penalizes abstractiveness in human summaries The results in Table 4.3 show that LLMs achieve comparable albeit higher ROUGE scores compared to human reviewers, which by standard reference-based protocols, would be an indicator that LLM summaries have “superior quality.” However, we observe quantitative differences between human and machine reviews across other metrics. Notably, human summaries are significantly more abstractive as measured by the MINT score and the percentage of novel bigrams in the summaries. Given the more creative nature of human reviews, it is expected that they offer more possibilities for lexical disagreement than a conservative summary that prioritizes content from the source document.

Reference Abstract (Zhou et al., 2020): Learning disentangled representations is regarded as a fundamental task for improving the generalization, robustness, and interpretability of generative models. However, measuring disentanglement has been challenging and inconsistent, often dependent on an ad-hoc external model or specific to a certain dataset. To address this, we present a method for quantifying disentanglement that only uses the generative model, by measuring the topological similarity of conditional submanifolds in the learned representation. This method showcases both unsupervised and supervised variants. To illustrate the effectiveness and applicability of our method, we empirically evaluate several state-of-the-art models across multiple datasets. We find that our method ranks models similarly to existing methods. We make our code publicly available at https://github.com/stanfordmlgroup/disentanglement.

Summary A: The article proposes a new evaluation metric for disentanglement in generative models, which is based on the concept of persistent homology. The metric measures the topological similarity between submanifolds conditioned on different latent dimensions in the generative model, and can be used to evaluate the disentanglement of different models and datasets. The proposed metric provides a more reliable and robust way of evaluating disentanglement compared to existing methods, and can be applied across different model architectures and datasets without requiring additional training or hyperparameter tuning.

Summary B: The paper proposes a novel metric for evaluating disentanglement by taking a manifold-topological perspective on the representations learnt. The key insight is that for a disentangled representation, when we fix a certain factor of variation at different values the topology of the conditional sub-manifolds should be similar. Using this insight the paper proposes a metric for disentangling which does not require annotations of the factors of variation and is more general than previous such tests.

Summary C: Researchers have proposed an unsupervised method for evaluating the level of disentanglement in deep generative models. This evaluation metric uses topology and density analysis of samples to measure the degree of topological similarity across submanifolds conditioned on disentangled latent dimensions versus entangled ones. The approach shows consistent results across 10 models and three datasets, without relying on training additional models or introducing a dataset-specific preprocessing step.

Questions

1. Given the reference abstract above, which of the alternative summaries do your prefer the most? You can select multiple if there are equally good summaries.
2. Justify your selection in the text box below. A possible reason could be “Summary A better represents the motivation of the paper and is more coherent.”
3. Which summary is the worst? (Like the previous question, you can choose multiple if no summary is clearly worse than the others)
4. Justify your choice in the text box below. A possible reason could be “Summary B presents non-factual research results.”

Table 4.4: Sample human evaluation task with a reference abstract, 3 candidate summaries, and questions for the evaluators.

Vote	Human	LLAMA	GPT-3.5	Cohen's κ
Best ↑	16.9	40.0	**43.1**	0.06
Worst ↓	36.1	**27.9**	36.1	0.17

Table 4.5: Human preferences (percentage) for best/worst MuP summaries. To account for multiple summary choices, we use the weighted Cohen's κ agreement statistic (Artstein and Poesio, 2008a).

Best	%	Worst	%
Coverage	**64.1**	**Coverage**	**42.9**
Coherence	12.8	Relevance	11.1
Fluency	6.4	Factuality	9.5
Conciseness	3.8	Informativeness	7.9

Table 4.6: Frequency of top criteria for human preferences on best and worst MuP summaries. Cohen's κ score is -0.16 and 0.1 for best and worst votes respectively.

Humans strongly prefer LLM summaries The results in Table 4.5 reveal the marked preference of evaluators towards LLM summaries, which account for 83% of votes for best summary. The preference for the worst summary is more balanced, with humans and GPT-3.5 taking the majority of votes. Also, we note that Cohen's κ scores (Artstein and Poesio, 2008a) are low, which indicates the high subjectivity of the task.

To understand the factors underlying human preferences, we categorize the evaluators' comments into quality criteria, as shown in Table 4.6. For both best and worst summary choices, the dominant reason is *coverage* of research aspects such as background or experimental results. Interestingly, *factuality* only appears as a third factor for choosing a bad summary, which is an indicator of the high quality of LLM summaries. However, we note that our evaluation protocol only includes the reference abstract, which makes more difficult to identify factuality issues.

Takeaways Despite the advantage of LLMs on reference-based metrics and human evaluation, *our aim with this experiment is not to claim that LLM summaries are superior*. Instead, we emphasize the limitations of evaluation based on reference summaries, as subjective preferences in communicative intentions impact ROUGE scores and even judgments of human annotators. Since LLM summaries exhibit human-level quality in terms of coherence, fluency, and informativeness, we argue that the major challenge for LLM summarizers lies in the adaptability to different tasks such as abstract generation and lay summarization, which we explore in Section 4.4.2.

Model	R-1	R-2	R-L	Control Metrics k_I		
				Concise↓	Narrative↑	Keywords↑
Reference	100	100	100	1.8	29.7	42.5
BIGBIRD	45.0	18.3	39.8	1.9	46.2	42.3
FACTORSUM	**48.9**	**20.1**	**43.8**	1.9	50.9	100
"Non-intentional" baseline prompt p_ε						
LLAMA	42.0	14.3	37.3	1.1	0.2	49.3
GPT-3.5	42.9	14.4	37.8	1.1	0.1	51.2
$p_{\text{conciseness}}$						
LLAMA	42.6	14.3	37.7	0.4	1.3	47.7
GPT-3.5	42.5	13.8	37.4	**0.3**	0.1	50.5
LLAMA-CFG	42.4	14.0	37.4	0.3	2.2	39.3
$p_{\text{conciseness}} + p_{\text{narrative}}$						
LLAMA	43.9	14.7	38.7	0.3	83.6	48.4
GPT-3.5	44.0	15.0	38.9	0.3	42.2	53.2
LLAMA-CFG	43.4	14.4	38.2	0.3	**84.5**	40.9
$p_{\text{conciseness}} + p_{\text{keywords}} + p_{\text{narrative}}$						
LLAMA	45.2	15.7	39.8	0.4	84.0	64.4
GPT-3.5	47.0	17.6	41.4	0.6	61.3	68.6
LLAMA-CFG	44.5	14.9	39.1	0.5	81.5	**74.2**

Table 4.7: Summarization results on the arXiv test set (1,000 samples) using different intention prompts p_I (defined in Section 4.2.1). We report ROUGE (F1) and intention control metrics $k_I(y_I, S)$ for conciseness, narrative perspective (first person), and keyword recall. The target conciseness in $p_{\text{conciseness}}$ is 6 sentences. The list of keywords in p_{keywords} is derived from FACTORSUM summaries as described in Section 4.2.1.2.

4.4.2 Controlling Style and Keyword Coverage

In this section, we gauge the performance of LLMs on benchmarks for abstract generation and lay summarization. Our goal is to compare the performance of zero-shot inference in LLMs to supervised baselines, and most importantly, to assess the flexibility of LLMs, via intention prompts, to replicate preferences such as narrative perspective and conciseness. In addition to LLAMA-2 and GPT-3.5, we evaluate strong fine-tuned encoder-decoder baselines: BIGBIRD and FACTORSUM (Zaheer et al., 2020; Fonseca et al., 2022) for abstract generation and BART (Lewis et al., 2020) for lay summarization. Also, we report results for LLAMA-CFG, denoting LLAMA-2 with classifier-free guidance (refer to Section 4.2 for details).

LLMs capture style intents We analyze the effects of style prompts compared to the baseline prompt p_ε. For abstract generation, LLMs are able to follow conciseness instructions ($p_{\text{conciseness}}$) consistently, as indicated by the significant reduction in the average deviation from the target number of sentences ($k_{\text{conciseness}}$ in Table 4.7). In this aspect, GPT-3.5 exhibits superior performance compared to LLAMA-2, achieving a deviation of 0.28 and 0.53 sentences for arXiv and PubMed, respectively.

By adding narrative perspective prompts ($p_{\text{conciseness}} + p_{\text{narrative}}$), we observe a similar level of controllability, with the percentage of sentences in first person voice increasing from zero to 84% for LLAMA-2 ($k_{\text{narrative}}$ in Table 4.7). This change of perspective also results in higher similarity to reference summaries (as measured by ROUGE), which validates the effectiveness of narrative guidance. Interestingly, GPT-3.5 generates fewer sentences in first person perspective, balancing first and third voices to achieve better fluency compared to LLAMA-2 summaries.

Long summaries remain challenging While conciseness prompts are effective for abstract generation, LLMs cannot achieve low conciseness deviation for long lay summaries (Table 4.10). To evaluate the limits of conciseness guidance, we compare LLAMA-2 and GPT-3.5 summaries (100 samples from the eLife validation split) with target conciseness varying from 1 to 16 sentences (Figure 4.2). We find that both LLAMA-2 and GPT-3.5 reproduce conciseness intentions almost perfectly up to 6 sentences when it starts to degrade its instruction adherence. In all cases, the model outputs do not saturate the maximum output length of 512 tokens. Thus, we hypothesize that longer generation targets conflict with the notion of "summary" that LLMs learn from pre-training data. However, our conclusions are limited to the investigated

Model	R-1	R-2	R-L	Control Metrics k_I		
				Conciseness↓	Narrative↑	Keywords↑
Reference	100	100	100	2.5	8.8	44.0
BIGBIRD	44.7	19.1	40.6	3.6	10.1	48.8
FACTORSUM	**47.3**	**20.0**	**43.3**	1.3	10.5	100
Baseline prompt p_ε: *Write a summary of the article above.*						
LLAMA	43.4	15.5	39.0	2.3	0	44.1
GPT-3.5	41.8	14.7	37.6	2.7	0	43.3
$p_{\textbf{conciseness}}$: *Write a summary of the article above in* $\{y_{\textit{conciseness}}\}$ *sentences.*						
LLAMA	43.4	15.2	39.2	1.0	0	44.3
GPT-3.5	43.9	15.4	39.6	**0.5**	0	49.7
LLAMA-CFG	43.4	15.1	39.2	0.7	0.2	41.5
$p_{\textbf{conciseness}} + p_{\textbf{narrative}}$: *Write a summary of the article above in* $\{y_{\textit{conciseness}}\}$ *sentences. Write in* ***first person "we"*** *when applicable.*						
LLAMA	44.1	15.8	39.6	1.0	62.7	45.1
GPT-3.5	44.1	15.6	39.8	0.7	22.3	50.4
LLAMA-CFG	43.7	15.7	39.3	0.8	**70.6**	43.8
$p_{\textbf{conciseness}} + p_{\textbf{keywords}} + p_{\textbf{narrative}}$: *Write a summary of the article above in* $\{y_{\textit{conciseness}}\}$ *sentences. Focus on the following keywords:* $\{y_{\textit{keywords}}\}$. *Write in* ***first person "we"*** *when applicable.*						
LLAMA	44.6	16.5	40.2	1.4	64.5	66.2
GPT-3.5	46.9	18.5	42.4	1.2	30.6	69.6
LLAMA-CFG	44.1	15.9	39.7	1.4	67.9	**76.4**

Table 4.8: Summarization results on the PubMed test sets (1,000 samples) using different intention prompts p_I. We report ROUGE (F1) and intention control metrics $k_I(y_I, S)$ for conciseness, narrative perspective (first person), and keyword recall. The target conciseness $y_{\text{conciseness}}$ is 8 sentences. The list of keywords y_{keywords} is derived from FACTORSUM summaries as described in Section 4.2.1.2.

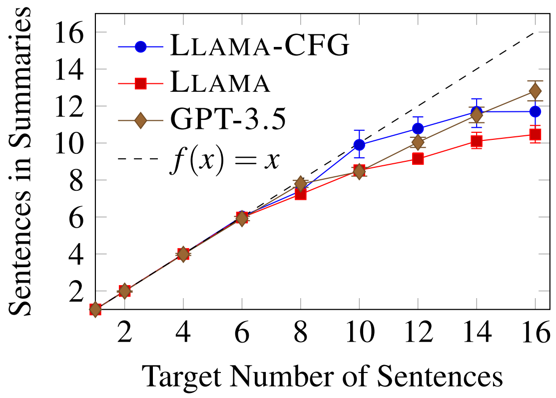

Figure 4.2: Number of sentences in generated summaries subject to varying conciseness targets (100 samples from eLife validation set).

LLMs, and results could change with upcoming models capable of handling significantly longer input contexts (Achiam et al., 2023; Jiang et al., 2023).

Keyword prompts improve lexical alignment We introduce the keyword coverage prompt p_{keywords} with a keyword model derived from FACTORSUM summaries for arXiv and PubMed. The results in Tables 4.7 and 4.8 demonstrate that the keyword recall for LLAMA-2 and GPT-3.5 increases significantly with keyword guidance, leading to ROUGE scores higher than BIGBIRD summaries. Similarly, keyword prompt improves lexical alignment in lay summarization (Table 4.10). In this case, in addition to keyword recall, we also observe that keyword guidance results in better readability scores and higher abstractiveness (Table 4.11), which are desirable properties for lay summaries.

Classifier-free guidance improves intention control We report results for LLAMA-2 with classifier-free guidance (CFG), using $\gamma = 1.5$ for all prompt variants[7] (Tables 4.7 and 4.10). In all cases, we observe an improvement in intention control metrics versus LLAMA-2 with regular decoding. In Figure 4.2, we can clearly observe how LLAMA-CFG can follow conciseness instructions for up to 10 sentences but fail to produce longer summaries.

LLMs perform well out-of-distribution One important concern is that LLAMA-based models might simply be memorizing their training data, which likely include scientific

[7] We set $\gamma = 1.5$ based on the results reported by Sanchez et al. (2023) and our experiments on the validation set.

Model	R-1	R-2	R-L	Control Metrics k_I	
				Concise↓	Narrative↑
BIGBIRD	21.2	2.5	19.3	5.0	57.8
"Non-intentional" baseline prompt p_ε					
LLAMA	45.3	14.2	42.6	1.0	0.3
GPT-3.5	46.6	15.5	43.9	1.1	0.1
$p_{\textbf{conciseness}} + p_{\textbf{narrative}}$					
LLAMA	46.4	14.6	43.3	**0.3**	**89.6**
GPT-3.5	**47.9**	**15.9**	**45.1**	0.7	18.5

Table 4.9: Summarization results for 500 arXiv samples published between December 2023 and January 2024. We report ROUGE (F1) and intention control metrics $k_I(y_I, S)$ for conciseness and narrative perspective (first person). The prompts p_ε, $p_{\textbf{conciseness}}$ and $p_{\textbf{narrative}}$ are defined in Section 4.2.1.

papers that overlap with the arXiv and PubMed summarization datasets. To check this possibility, we collected 500 articles (category `cs.CL`) submitted to arXiv between December 2023 and January 2024 and removed all the information before the introduction section (including title, abstract, and other information). Then, we evaluate the same models without further fine-tuning, using the narrative perspective and conciseness prompts. The results in Table 4.9 show that instruction-tuned models achieve even higher ROUGE scores (compared to Table 4.7), suggesting that memorization is not a cause for their good summarization performance. In contrast, we observe a pronounced loss of performance BIGBIRD, which tends to generate repeated sentences.

4.5 Related Work

Our work fits in the context of recent research of large language models applied to summarization, mostly for news articles (Goyal et al., 2022; Zhang et al., 2024). In this work, we explore summarization in the scientific domain and its specific challenges related to document length (Beltagy et al., 2020; Fonseca et al., 2022) and technical writing style. Furthermore, we consider our work as a contribution towards understanding production variability by humans and language models, which was explored

Model	**R-1**	**R-2**	**R-L**	**Control Metrics k_I**		
				Concise↓	**Narrative↑**	**Keywords↑**
Reference	100	100	100	10.8	99.5	44.1
BART	**48.1**	**13.9**	**30.5**	**2.0**	**99.9**	100
"Non-intentional" baseline prompt p_ε						
LLAMA	34.5	7.7	21.0	8.1	99.3	28.4
GPT-3.5	29.3	6.7	19.3	9.2	1.0	27.4
$p_{\textbf{conciseness}}$						
LLAMA	39.4	8.5	23.7	3.9	99.4	32.6
GPT-3.5	40.6	8.7	25.5	3.1	99.2	38.0
LLAMA-CFG	37.4	7.9	23.2	3.0	98.7	32.3
$p_{\textbf{conciseness}} + p_{\textbf{keywords}}$						
LLAMA	44.0	10.3	25.5	2.3	99.6	60.2
GPT-3.5	42.4	9.9	26.4	4.7	99.7	61.5
LLAMA-CFG	42.9	9.6	25.1	3.1	99.7	**81.5**

Table 4.10: Summarization results on the eLife test set (241 samples) using different intention prompts p_I (defined in Section 4.2.1). We report ROUGE (F1) and intention control metrics $k_I(y_I, S)$ for conciseness, narrative perspective (third person), and keyword recall. The target conciseness in $p_{\textbf{conciseness}}$ is 14 sentences. The list of keywords in $p_{\textbf{keywords}}$ is derived from BART summaries as described in Section 4.2.1.2.

on tasks such as translation and open-domain dialogue by Giulianelli et al. (2023).

Finally, our work is related to attempts to improve the prompt-adherence in generation tasks. Pu and Demberg (2023) apply prompt-based methods to control summarization according to target audience (expert vs. layman) and style (formal vs. informal). Kumar et al. (2022) propose a non-autoregressive generation method to introduce soft and hard constraints, including adherence to keywords. In a different direction, (Sanchez et al., 2023) showed that Context-Free Guidance (CFG) (Ho and Salimans, 2022), a technique originally used in text-to-image models, improves the performance of language models across several tasks. While Sanchez et al. (2023) applied CFG to tasks such as question answering, code generation and translation, our work is the first

Model	FKGL↓	Novel Bigram↑
Reference	10.8	**66.6**
BART	**10.2**	57.8
"Non-intentional" baseline prompt p_ε		
LLAMA	15.69	31.5
$p_{\textbf{conciseness}}$		
LLAMA	14.8	28.4
LLAMA-CFG	14.1	27.0
$p_{\textbf{conciseness}} + p_{\textbf{keywords}}$		
LLAMA	10.9	47.2
LLAMA-CFG	12.8	55.8

Table 4.11: Effects of intention prompts and classifier-free guidance (CFG) on readability (FKGL) and abstractiveness (novel bigrams) on the eLife test set (241 samples).

to explore this technique for summarization controllability.

4.6 Limitations

Model scale and proprietary APIs Our experiments operate at model scales up to 7 billion parameters (LLAMA-2), and we expect to observe more performance gains for larger models (Wei et al., 2022). By including results from the OpenAI API, we can get an estimation of the level of performance of larger models, although we do not have access to information about the number of parameters of the underlying model and whether it includes extra machinery on top of language modeling.

Excluded models and reproducibility issues Despite our best prompting efforts, we could not make encoder-decoder instruction-tuned models T0 (Sanh et al., 2022) and FLAN-T5 (Chung et al., 2022) generate long summaries (at least six sentences), which makes them unsuitable for comparison. We also tried other models that claim high ROUGE performance but do not provide public code (Pang et al., 2023), fine-tuned checkpoints for the scientific papers (Guo et al., 2022; Xiong et al., 2022), or reproducible results with the provided resources (Phang et al., 2022).

Human evaluation We limit our analysis of variability in communicative intents from the point of view of summarizers, but we believe human evaluators are also subject to these factors. An interesting future investigation could explore how different annotation conditions and guidelines potentially bias human judgments and to what extent human evaluation is still valid as the gold standard.

Other summarization domains Related work in news summarization explores the generation of summaries based on guidelines for topic and entity coverage (Ahuja et al., 2021; Maddela et al., 2022). We believe that the methods we presented in this chapter could be applied to such coverage requirements. However, given that news summarization has been extensively investigated in previous work, we choose to perform a comprehensive evaluation of summarization models in the scientific domain.

4.7 Conclusion

Our experimental results on four scientific summarization benchmarks confirm that large language models are effective summarizers, as measured by conventional lexical overlap metrics and human preferences. Moreover, we find that LLMs can follow intentional prompts for style and keyword coverage, especially for short summaries. This summarization controllability can be further improved using simple decoding changes such as classifier-free guidance. However, longer summaries that require a higher level of abstractiveness are not easily achievable with our prompting techniques. Our findings suggest that analyzing summaries as an expression of communicative intentions leads to informative and actionable insights for future model improvements, where traditional reference-based evaluation shows its limits.

Finally, our results suggest that large language models dramatically reduce the costs of deploying high-quality summarization services. As a consequence, we expect that much of the information available on social media and other sources will be abridged versions generated by commoditized summarization systems. In this context, special attention is needed regarding existing issues of generative language models such as hallucination (McKenna et al., 2023). Crucially, we have shown that language models can adapt their outputs according to specific communication intentions, which might bias the summary contents towards the (potentially harmful) beliefs and desires of the actors behind those systems.

Chapter 5

Controlling Coverage of Domain Concepts

In the previous chapters, we investigate the controllability of stylistic features and simple lexical guidance in summaries. However, some generation tasks require control of more abstract factors such as the coverage of domain-specific concepts, e.g., *background*, *objectives*, and *results* in a paper abstract. In this chapter, we set out to address the problem of controllability of concepts in summaries by designing a three-stage pipeline: per-concept *annotation*, *planning*, and *editing* (APE). While our approach demonstrates effectiveness in changing concept coverage in lay summarization, it provides a potential diagnostic: this kind of controllability might be hindered by poor understanding of domain concepts by large language models (LLMs).

Although LLMs exhibit a remarkable capacity to leverage in-context demonstrations, it is still unclear to what extent they can learn new facts or *concept definitions* via prompts. To address this question, we examine the capacity of instruction-tuned LLMs to follow in-context *concept annotation guidelines* for *zero-shot* sentence labeling tasks. We design guidelines that present different types of *factual* and *nonfactual* concept definitions, which are used as prompts for zero-shot sentence classification tasks. Our results show that although concept definitions consistently help in task performance, only the larger models (with 70B parameters or more) have a limited ability to work under nonfactual contexts. Importantly, only proprietary models such as GPT-3.5 can recognize nonsensical guidelines, which we hypothesize is due to more sophisticated alignment methods. Altogether, our simple evaluation method reveals significant gaps in concept understanding between the most capable open-source language models and the leading proprietary APIs.

5.1 Controlling Concept Coverage in Summaries

While controlling simple style intentions such as narrative perspective is straightforward with prompts (see Chapter 4), guiding large language models toward a particular coverage of concepts requires more fine-grained control. An adequate balance of concept coverage is crucial to fulfill the purpose of a summary. For instance, in scientific lay summarization (Goldsack et al., 2022), there is a preference for research background content over methodological details, which helps a lay reader to parse the material. To achieve this level of control, we propose to structure the summarization task to explicitly reason about concepts, in a three-stage inference process: *annotation*, *planning*, and *editing* (APE).

Annotation step First, we prompt the model to describe the main points about a scientific concept c, that is, to generate a *concept note* n_c, for $c \in \{$*"background and related work", "research questions and motivation", "research methods and experiments", "experimental results", "research conclusions"*$\}$. We define the annotation prompt p_c for a concept c as follows:

```
Article: {D}
For the article above, describe the main {c} in {b_c}
sentences.
```

where $\{D\}$ is a document and $\{b_c\}$ is the target number of sentences for n_c. Having a separate annotation step allows better interpretability of the model inference, as one can check the correctness of each concept note and diagnose potential knowledge gaps for specific concepts.

Planning step In this step, it is possible to post-process the concept notes generated in the previous step. For example, this post-processing could favor notes with specific entity coverage or eliminate content with quality issues. For simplicity, in our experiments, we simply tokenize each note n_c into sentences and filter out sentences with less than 10 words as well as incomplete sentences that do not end with a period (".") character. We denote this quality filter as the function IS_VALID in Algorithm 2.

Editing step In the final step, we prompt the language models to edit the concatenated notes $\hat{n}_c$ resulting from the planning step, into a more fluent summary-like text. The editing prompt p_e is defined as follows:

```
{notes}
Rephrase the content above into a fluent text, using
transition words.
```

where {notes} is the concatenation of post-processed notes from the planning step.

This multi-step approach (refer to Algorithm 2 for details) facilitates the generation of several types of summaries from the same set of initial concept notes n_c, by choosing different planning and editing strategies. Secondly, one can use different language models for the annotation and editing step, which allows the use of a more computationally expensive editor model at a lower cost in terms of the number of tokens compared to using the source document. While the planning step could be expressed as a prompt in the edit step, we prefer to implement it as a separate deterministic algorithm that does not depend on the LLM probabilistic behavior.

Algorithm 2: The Annotate-Plan-Edit (APE) algorithm for guiding concept distribution in a summary of document *D*. The variables *concepts* and *budgets* refer to the concept labels and the target number of sentences per concept, respectively. The variables *annotator* and *editor* denote language models responsible for generating and editing concept notes.

```
Input: D, concepts, budgets, annotator, editor
Output: S
notes ← ∅;
i ← 0;
for c ∈ concepts do
    b_c ← budgets[i];
    p_c ← ANNOTATION_PROMPT(D, c, b_c);
    n_c ← annotator.GENERATE(p_c);
    n_c ← SENTENCE_TOKENIZE(n_c);
    for s ∈ n_c do
        if IS_VALID(s) then
            notes ← notes ∪ {s};
        end
    end
    i ← i + 1;
end
p_e ← EDITOR_PROMPT(notes);
S ← editor.GENERATE(p_e)
```

5.1.1 Experimental Setup

In this section, we describe the dataset, models, and metrics used in our experimental setup, which is very similar to the lay summarization experiment in Section 4.3. We evaluate the models on the eLife lay summarization benchmark (Goldsack et al., 2022), which includes summaries with an unusual imbalance of concepts. These summaries have around 18 sentences on average, with more than 50% of the content dedicated to background information, which makes this dataset suitable for our controllability experiments.

As annotator and editor models, we use LLAMA-2 (7B-chat version; Touvron et al. 2023) and GPT-3.5 (version `gpt-3.5-turbo-0613`) via the chat completion API endpoint[1]. We set a target budget of 18 sentences, which in the APE annotation step is divided per concept as follows: *background and related work* (12), *research questions and motivation* (1), *research methods and experiments* (1), *experimental results* (2), and *research conclusions* (2). This distribution of concepts is derived from the average number of sentences in the eLife validation set. Finally, we also report results for the BART supervised baseline (Lewis et al., 2020). All training and generation details are the same as described in Section 4.3.2.

5.1.1.1 Metrics

In addition to ROUGE (Lin, 2004) and the conciseness control metric from Section 4.2.2, we also report the distribution of sentences per concept in summaries. Let S_k be a summary and $g(s_i)$ be a concept classifier that maps a sentence $s_i \in S_k$ into a concept $c \in [C]$. We define the concept weight w_k^c for a summary S_k as the total number of sentences assigned to concept c:

$$w_k^c = \sum_{s \in S_k} \mathbb{1}[c = g(s_i)]. \tag{5.1}$$

Then, we obtain a distribution of content per concept w^c by averaging over all summaries $S_k \in \mathcal{S}$:

$$w^c = \frac{1}{|\mathcal{S}|} \sum_{k=1}^{|\mathcal{S}|} w_k^c. \tag{5.2}$$

A classifier for scientific concepts The concept coverage metrics described above depend on the sentence-level concept classifier $g(s)$ (Eq. 5.1). To train the classifier, we combine the three datasets containing annotated sentences from scientific papers:

[1] `https://platform.openai.com/docs/api-reference/chat`

- **PubMed-RCT** (Dernoncourt and Lee, 2017): contains sentences from 200,000 abstracts of randomized controlled trials (RCT), each one classified according to their rhetorical role as BACKGROUND, MOTIVATION, METHODS, RESULTS or CONCLUSION. For our model, we use a subset of 20,000 abstracts provided by the dataset authors[2].

- **ArtCorpus** (Liakata et al., 2010): a dataset with 35,040 sentences from 225 scientific papers (full text), each annotated by 20 chemistry specialists according to a taxonomy of core scientific concepts (Liakata and Soldatova, 2008). Since the original CoreSC concepts contain 11 fine-grained classes, we post-processed the dataset to merge more detailed categories, so the resulting concepts match the five categories from the PubMed-RCT dataset.

- **CSAbstruct** (Cohan et al., 2019): contains 2,189 computer science abstracts with sentence-level rhetorical role labels. We use samples from BACKGROUND, MOTIVATION, METHODS, and RESULTS, and ignore the ones labeled as OTHER.

From the datasets above, we obtain 110,010, 17,741, and 16,891 sentences for training, validation, and test respectively. We fine-tune a SciBERT model (Beltagy et al., 2019) articles[3] for 5,000 steps and choose the checkpoint with best accuracy on the validation set (84.8 F1 score). The resulting model classifies a sentence into one of the following concepts: BACKGROUND, MOTIVATION, METHODS, RESULTS or CONCLUSION.

5.1.2 Results and Discussion

As discussed in Section 4.4.2, generating long summaries is challenging for both LLAMA-2 and GPT-3.5. The results in Table 5.1 and Figure 5.1 demonstrate that performing per-concept annotation with APE helps to generate longer summaries, as seen by the lower conciseness deviation and the higher number of sentences covering research background. Interestingly, GPT-3.5 exhibits a strong bias towards background content by default, which raises the question of whether this model has more exposure to this kind of summary during pre-training. In contrast, LLAMA-2 offers a more neutral coverage of concepts, and the APE pipeline is very effective in changing the conceptual distribution of its summaries.

---

[2]Dataset available at https://github.com/Franck-Dernoncourt/pubmed-rct.
[3]https://huggingface.co/allenai/scibert_scivocab_uncased

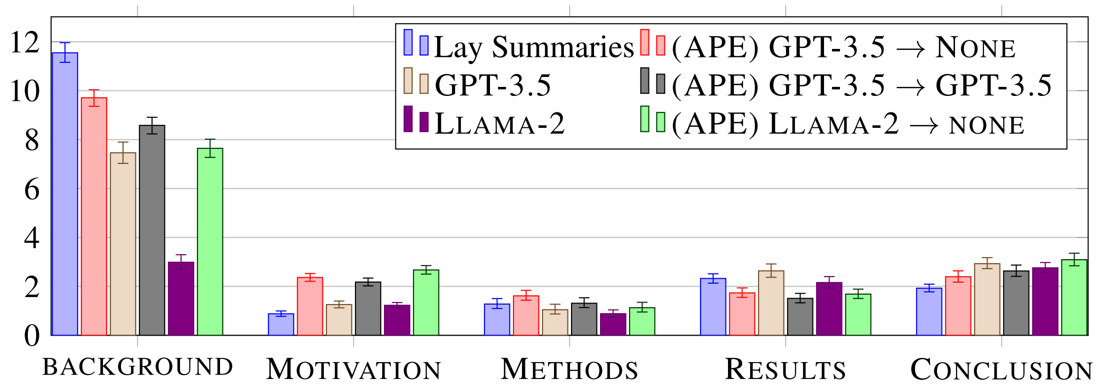

Figure 5.1: Average number of sentences per scientific concept in summaries (Eq. 5.2) on the eLife test set. APE summarizers are represented as ANNOTATOR → EDITOR. Error bars indicate the 95% confidence interval.

Overall, we observe that our simple APE summarizer can achieve some level of controllability in terms of concept coverage, without impacting ROUGE scores significantly. However, we still observe a large gap between human and machine-generated lay summaries mainly regarding the coverage of background information. We attribute this difference to the high abstractiveness of lay summaries (see Table 5.1) and, possibly, to the limited understanding of scientific concepts by the LLMs. In the rest of this chapter, we investigate the latter issue in more depth.

5.2 Gauging LLM Concept Understanding

In the previous section, we illustrated how the understanding of domain concepts can be crucial to tailor summaries to specific audiences. We argue that measuring concept understanding is an useful proxy to check if a summarization model exhibits more sophisticated understanding of the underlying subject (e.g., the methods of a scientific investigation), or if it is leveraging lexical-level statistics. In this section, we propose methods to gauge concept understanding when language models have access only to their prior knowledge, and also to in-context concept definitions (*guidelines*) in two different domains: scientific and financial documents.

Large language models are known to distill knowledge from vast datasets during the pre-training phase (Brown et al., 2020). Such knowledge can be queried via prompting, which allows the application of LLMs to several knowledge-intensive tasks in zero-shot and few-shot settings. In particular, recent work demonstrates promising applications of LLMs to reduce the cost of data annotation in several domains (Wang

Model	R-1	R-2	R-L	Concise $k_I \downarrow$	Nov. Bigram
BART	**48.1**	**13.9**	**30.5**	**2.0**	57.8
Prompt: *Write a summary of the article above in* ***18 sentences.***					
LLAMA-2	39.17	8.44	23.66	8.0	28.3
GPT-3.5	43.42	9.82	27.23	4.4	32.5
APE (our method)					
Annotator → Editor	Sentences per concept: $[12,1,1,2,2]$				
LLAMA-2 → NONE	40.86	9.53	25.25	3.2	33.4
GPT-3.5 → NONE	43.10	10.17	26.77	2.8	32.9
GPT-3.5 → GPT-3.5	43.44	9.43	26.97	3.5	41.9

Table 5.1: Summarization performance on the eLife test split (241 samples). We report ROUGE (F1) scores, conciseness control k_I (see Section 4.2.2), and novel bigrams in summaries (%). The NONE editor (APE) outputs are the annotator's concept notes.

et al., 2021; Agrawal et al., 2022; Zhu et al., 2023).

Most data labeling efforts based on LLMs leverage in-context demonstrations to elicit the desired concepts. However, previous work suggests that language models cannot learn from in-context ground-truth labels but just leverage demonstrations to infer the task format and label space (Min et al., 2022). In contrast, human annotators typically follow *guidelines*, which in addition to examples, include concept definitions (Liakata and Soldatova, 2008) that complement and modify the annotator's prior concept understanding to align with the labeling goals.

In this section, we assess the capacity of LLMs to follow analogous in-context *concept annotation guidelines* for sentence classification tasks. Our goal is to verify if language models can learn from in-context definitions and change their behavior consistently in downstream tasks (Onoe et al., 2023). To this end, we design several types of guidelines that represent both *factual* and *nonfactual* concept definitions[4]. Our assumption is that learning from concept definitions would imply the capacity to reason in contexts that contradict the model's prior knowledge.

In our experiments, we evaluate the LLAMA-2 model by Touvron et al. (2023)

[4]In this work, we denote as *nonfactual* those concept definitions that disagree with commonsense understanding, which we assume to be prevalent in a *default world model* (Wu et al., 2023) derived from the pre-training data. We formalize *factual* and *nonfactual* guidelines in Section 5.2.1.1.

Consider the following concepts:
- Background: A sentence that provides context, foundational knowledge, or relevant information about the research topic, existing theories, prior studies, or the broader scientific field in which the research is situated.

(more definitions c_K: $\delta(c_K)\ldots$)

- Conclusion: A sentence that summarizes the key takeaways, implications, interpretations, or insights derived from the study's results.

`Classify the text below into one of the categories listed above. Be concise and write only the category name.`

`Text: Therefore, the phase transition can be classified as essentially driven by Coulomb interactions.`
`Concept:` **`Conclusion`**

Figure 5.2: An abridged example of zero-shot sentence classification using a concept guideline prompt. We perform controlled interventions in **concept definitions** (pairs of concept labels c_K and their descriptions $\delta(c_K)$) while keeping the `task prompt` fixed. We aim to gauge the capacity of the model to learn new concepts during inference, *without in-context demonstrations.*

(7B, 13B, and 70B-parameter chat variants), FALCON-180B-CHAT (Almazrouei et al., 2023), GPT-3.5, and GPT-4 (Achiam et al., 2023) on zero-shot sentence classification tasks (as illustrated in Figure 5.2). The tasks require the recognition of scientific concepts, for which labels are likely present in the models' pre-training data. (Liakata and Soldatova, 2008). To control for pattern memorization, we also annotate a novel dataset of company disclosures with financial concepts based on the Integrated Reporting framework (Cheng et al., 2014).

In both domains, we observe a consistent classification performance improvement when models have access to concept labels paired with their factual concept definitions (compared to just a list of labels). However, when presented with nonfactual guidelines, only larger models (70B parameters or more) tend to output predictions consistent with guidelines. Still, we observe that scaling alone is not sufficient, as FALCON-180B-CHAT is outperformed by LLAMA-2-70B-CHAT in most settings. Importantly, only proprietary models are able to recognize unsolvable tasks, that is, ones for which the guidelines provide nonsensical concept labels. Finally, we find that the performance of more capable models such as GPT-3.5 is more strongly correlated to the degree of guideline factuality compared to LLAMA-2-7B, suggesting that the former model has a more nuanced concept understanding.

Overall, some of our findings reinforce previous studies focusing on few-shot learning using perturbed labels (Wei et al., 2023) and chain-of-thought reasoning (Saparov and He, 2022). However, our classification tasks require the model to generalize *only from concept definitions*, without demonstrations. Additionally, unlike previous work, we provide extensive experiments using state-of-the-art open-source models. Although these models may approach the aggregate performance of proprietary APIs, our results reveal important gaps in terms of concept understanding, especially in nonfactual scenarios and regarding the ability to recognize nonsensical tasks.

5.2.1 Concept Classification with Guidelines

Let S and C be random variables representing sentences (from the set of token sequences $\mathcal{S}$) and corresponding latent concepts to be inferred (from the concept set $\mathcal{C}$; e.g., whether the sentence conveys scientific *background* or *methods*). To specify the task, we introduce annotation guidelines G, which specify concept labels and their definitions. Then, the concept annotation process for a sentence s given the guideline g is formalized as follows:

$$c_s = \arg\max_{c' \in \mathcal{C}} P(C = c' \mid K, G = g, S = s), \tag{5.3}$$

where c_s is the inferred concept and K represents prior domain knowledge about the concepts of interest. Then, we define a language model P_θ that approximates Eq. 5.3 through conditional generation:

$$y_s = P_\theta(\cdot \mid \texttt{prompt}_G(g); \texttt{prompt}_t(s)), \tag{5.4}$$

where y_s is a concept label, that is, a sequence of tokens corresponding to a concept c_s. The functions $\texttt{prompt}_G$ and $\texttt{prompt}_t$ are textual templates that describe concept guidelines and a concept classification task respectively (see Figure 5.2). The guideline prompt $\texttt{prompt}_G$ includes a list of concept definitions:

```
Consider the following concept categories:
- {c_1}:  {δ(c_1)}
...
- {c_K}:  {δ(c_K)}
```

where $\delta(c_K)$ is a function that maps the concept label c_K to its definition (refer to Section 5.2.1.1 for details). Then, we define the task $\texttt{prompt}_t$ as follows:

```
Classify the text below into one of the categories listed
above.  Be concise and write only the category name.

Text:  {input sentence s}
{domain} Concept:
```

where the placeholder {domain} is a domain indicator (in this work, we use either Scientific for scientific concepts or the empty string for financial concepts).

The language model parameters θ capture the prior conceptual knowledge K acquired during pre-training and instruction-tuning. In addition, we hypothesize that θ encodes the conditional dependency between the guidelines G and the concepts C as the guidelines express relationships between concept labels and their task-specific definitions. By performing controlled interventions in the guidelines, we aim to measure how concept understanding is affected by the following factors: 1) the lexical information of concept labels and concept definitions; 2) the degree of factuality of concept definitions. In the next section, we present guidelines designed to capture such factors.

5.2.1.1 Concept Guidelines

We associate each concept $c_i \in \mathcal{C}$ to a natural language definition $d_j \in \mathcal{D}$ through a injective *concept definition* function δ: $\mathcal{C} \to \mathcal{D}$, where each $c_i \in \mathcal{C}$ and $d_j \in \mathcal{D}$ are sequences of tokens. Each association $\delta(c_i) = d_j$ represents a *factual definition* if $i = j$ and a *nonfactual definition* otherwise. Then, a concept guideline is formalized as a tuple $G = \langle \mathcal{C}, \mathcal{D}, \delta \rangle$. Depending on the choice of the function δ, we derive different types of factual and nonfactual concept guidelines, which we describe below.

Factual guidelines G_f The factual guidelines combine the concept labels with their corresponding factual definitions, that is, $\delta(c_i) = d_i$ for all i. To illustrate, a factual guideline prompt would include the following definition for the scientific concept BACKGROUND:

> *Background: A sentence that provides context, foundational knowledge, or relevant information about the research topic, existing theories, prior studies, or the broader scientific field in which the research is situated.*

The factual guideline serves as a control baseline to compare against other types of guidelines. For each of the scientific and financial concepts explored in this paper, we use definitions generated by GPT-3.5. These definitions are further reviewed for quality and redacted to remove explicit mentions of label names. We provide the complete list of factual definitions and prompts in Section 5.2.2.4.

Out-of-dictionary guidelines G_{OOD} We replace real concept labels c_i from factual guidelines with out-of-dictionary (OOD) words (e.g., `Snizzlewump`). With those OOD words, we remove the dependency with respect to prior knowledge tied to the lexical information of concept labels. The OOD labels are generated by GPT-3.5 using the prompt *Give me a list of random out-of-dictionary words*. The resulting words are: `Flibberknock`, `Quibblesnatch`, `Blibberflop`, `Ziggledorf`, `Snizzlewump`, `Wobblequark`, `Jibberplunk`, `Crumblefluff`, `Splonglewort`, `Dinglewhack`.

Empty-definition guidelines G_ε As a variant of the factual and OOD guidelines above, we replace each definition with an empty string, that is, $\delta(c_i) = \varepsilon$ for all i. We denote these guidelines $G_{f,\varepsilon}$ and $G_{\text{OOD},\varepsilon}$ respectively, and use them to gauge the contribution of concept definitions compared to the factual guideline G_f baseline.

Nonfactual guidelines G_n A guideline is considered nonfactual when at least one concept c_i is paired with a definition from another concept c_j, that is, $\delta(c_i) = d_j$ for $i \neq j$. Since δ is injective, we have the number of nonfactual definitions (the *degree of nonfactuality* of a guideline) ranging from two to $|\mathcal{C}|$.

5.2.2 Experimental Setup

To experiment with different types of guidelines defined in Section 5.2.1.1, we choose concepts $\mathcal{C}$ for which the language models have exposure via pre-training data. The first domain we explore relates to rhetorical roles in scientific articles (Section 5.2.2.1), which is extensively covered in the literature (Liakata et al., 2012). Since the pre-training data for LLMs likely include various scientific concept classification datasets, we annotate a novel dataset of sentence-level financial concepts (Section 5.2.2.2). In addition to controlling for label memorization, our financial annotation based on the Integrated Reporting framework (Cheng et al., 2014) covers concepts that are technical but arguably more accessible than scientific rhetoric. In Sections 5.2.2.3 and 5.2.2.5, we detail the LLM baselines used in the experiments and the classification task hyper-parameters.

5.2.2.1 Scientific Concepts Dataset

To test a model's knowledge of scientific concepts we use the ARTCorpus dataset (Liakata and Soldatova, 2008), which consists of 35,040 sentences from 225 chemistry papers annotated by experts. Each sentence is labeled with one of the 11 *Core Scientific Concepts* (CoreSC) derived from the EXPO ontology (Soldatova and King, 2006).

In the CoreSC scheme, the scientific concepts are structured hierarchically, with concepts such as *hypothesis*, *motivation*, and *goal* being different sub-types of scientific *objectives*. Since the dataset is relatively small, we observed that some classes were too fine-grained, resulting in a strong label imbalance. To address this issue, we merged some of the categories that shared the same parent concept, yielding the following set of categories: *Background*, *Objective*, *Methods*, *Results*, and *Conclusion*. This classification scheme is also used in other PubMed-derived datasets such as the PubMed RCT (Dernoncourt and Lee, 2017). In our experiments, we use 500 sentences (100 samples per scientific concept) sampled from the ARTCorpus training split.

5.2.2.2 Financial Concepts Dataset

In this section, we introduce the methodology used to collect and annotate a dataset of company disclosures with financial concepts.

Data Collection We collected narrative sections from 10-K annual reports extracted from the Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system (SEC, 2014), which is used by companies to submit documents to the United States Securities and Exchange Commission (SEC). For each report, we use the following sections: Item 1 - Business, Item 7 - Management's Discussion and Analysis, and Item 7A - Quantitative and Qualitative Disclosure about Market Risk. The reports are published in December 2021 by companies in the S&P 500 index (S&P Global, 2024) with the largest market capitalization across 11 industry sectors.

Annotation Scheme Several reporting standards (IFRS[5], GAAP[6]) and ontologies such as FIBO (Bennett, 2013) have been developed, but they are often too technical and complicated to derive a simple taxonomy of financial concepts. Fortunately, the Integrated Reporting (IR) framework[7] offers a suitable set of domain concepts for the

[5] https://www.ifrs.org
[6] https://www.investopedia.com/terms/g/gaap.asp
[7] https://integratedreporting.org

task. It defines a set of reporting elements that deliver a holistic view of how the company uses capital to generate value (in this case, value in a broad sense, not just financial). In this work, we use one dimension of the IR framework related to *capitals*, which is the pool of funds available to an organization for use in the production of goods or the provision of services. The capital concept types are: *Financial*, *Manufactured*, *Intellectual*, *Human*, *Social and relationship*, and *Natural*.

Annotation Process Our annotation process is inspired by the General Scientific Concepts guidelines (Liakata and Soldatova, 2008), but our concepts typology is not hierarchical and does not account for instances of concepts (i.e., assigning identifiers for each concept instance). Before engaging in the annotation task, the annotators were presented with the textual guidelines listed in Figures 5.4 and 5.5. The web-based annotation interface (Figure 5.3) is implemented using Label Studio[8]. The interface shows the sample sentence and requests the annotator to classify it in one of the six capital concepts (`Financial`, `Manufactured`, `Intellectual`, `Human`, `Social and relationship`, and `Natural`) or `None` if the content is not related to any capital. The annotator also has the option to indicate a secondary capital, if applicable.

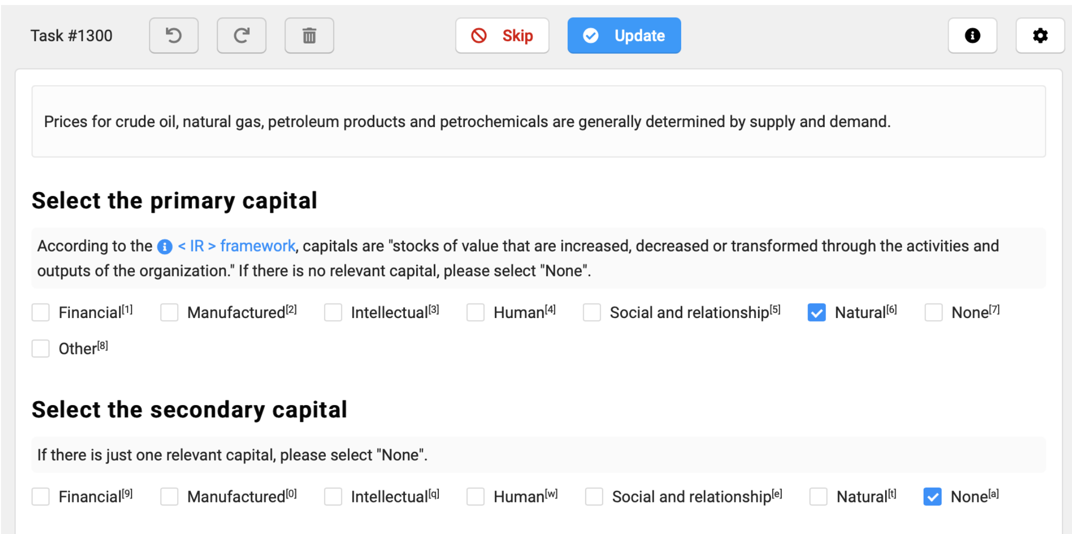

Figure 5.3: The annotation interface for the annotation of financial concepts. Given a sample sentence, annotators are requested to assign one of six capital concepts or `None`, if not applicable.

[8]https://github.com/HumanSignal/label-studio

The concepts described in this section follow closely the definitions of the International <IR> framework, and should be sufficient to perform the annotation. According to the IR framework, capitals are "stocks of value that are increased, decreased or transformed through the activities and outputs of the organization." They can be classified in financial, manufactured, intellectual, social and relationship, and human <IR> framework, section 2C).

Financial capital
The pool of funds that is available to an organization for use in the production of goods or the provision of services. It can be obtained through financing or generated through operations and investments. Example: *"The discussion also provides information about the financial results of our business segments to provide a better understanding of how those segments and their results affect the financial condition and results of operations of Ameren as a whole."*

Manufactured capital
Manufactured physical objects (excluding natural physical objects) that are available to an organization for use in the production of goods or the provision of services, including, buildings, equipment, and infrastructure (such as roads, ports, bridges, etc). Example: *"Due to the long lead time for the manufacture, repair, and installation of the components, the energy center is expected to return to service in late June or early July 2021."*

Intellectual capital
Organizational, knowledge-based intangibles, including: Intellectual property, such as patents, copyrights, software, rights and licences "Organizational capital" such as tacit knowledge, systems, procedures and protocols. Example: *"The absence of revenues from a software licensing agreement with Ameren Missouri decreased margins $5 million."*

Figure 5.4: Guidelines for financial concept annotation provided to human labelers. Concepts: Financial, Manufactured, and Intellectual capitals.

Human capital
People's competencies, capabilities and experience, and their motivations to innovate, including their: 1) alignment with and support for an organization's governance framework, risk management approach, and ethical values; 2) sbility to understand, develop and implement an organization's strategy; 3) loyalties and motivations for improving processes, goods and services; 4) Other matters related to people management. Example: *"As the situation rapidly evolved, we remained focused on safely serving our customers and protecting the health and safety of our employees."*

Social and relationship capital
The institutions and the relationships within and between communities, groups of stakeholders and other networks, including: 1) shared norms, and common values and behaviours; 2) key stakeholder relationships, and the trust and willingness to engage that an organization has developed and strives to build and protect with external stakeholders; 3) intangibles associated with the brand and reputation that an organization has developed; 4) an organization's social licence to operate. Example: *"In March 2020, the MoPSC issued an order in Ameren Missouri's July 2019 electric service regulatory rate review, approving nonunanimous stipulation and agreements."*

Natural capital
All renewable and non-renewable environmental resources and processes that provide goods or services that support the past, current or future prosperity of an organization, including air, water, land, minerals, and biodiversity. Example: *"These amounts include the 700 MWs of wind generation projects discussed below, which will support Ameren Missouri's compliance with the state of Missouri's requirement of achieving 15% of native load sales from renewable energy sources beginning in 2021."*

Figure 5.5: Guidelines for financial concept annotation provided to human labelers. Concepts: Human, Social and Relationship, and Natural capitals.

Annotation Round	Annotator Agreement		
	A_{12}	A_{13}	A_{23}
Round 1	0.27	0.35	0.35
Round 2	0.45	0.60	0.35

Table 5.2: Annotator agreement on capital labels. A_{ij} is the weighted Cohen's κ between annotators i and j.

Hiring and training Two final-year undergraduate students and one graduate student with a background in finance/economics were hired as annotators. The compensation was 10 British Pounds per hour of work, with an estimated effort of 50 sentences per hour[9]. Each annotator received one-to-one training about the motivation, annotation scheme, and guidelines, which included a description of each financial concept, examples, and general instructions covering edge cases.

Agreement assessment In the first round of annotation, each labeler worked on the same report (with 1,291 sentences) and then the agreement was estimated using the weighted Cohen's κ statistic (Artstein and Poesio, 2008b). The weighted version was adopted because each annotator is allowed up to two choices for capitals, so partial agreements are also taken into account. Formally, we define the disagreement weight d as the symmetric difference between the sets of labels $L_{a,i}$ and $L_{b,i}$ assigned to sample i by annotators a and b respectively:

$$d_i(a,b) = |(L_{a,i} \cup L_{b,i}) \setminus (L_{a,i} \cap L_{b,i})|.$$

The disagreements $d_i(a,b)$ are used in the weighted Cohen's κ formulation by Artstein and Poesio (2008b), Section 2.6.2.

The first round aimed at gauging the annotator's understanding of the guidelines and also, collecting feedback to improve the instructions. After analysis of the results, a one-to-one review session was delivered to give feedback about some common misconceptions and disagreements. A second report (562 sentences) was released to assess the effect of the improved guidelines. As shown in Table 5.2, the scores improved significantly in the second round, suggesting the changes in the guidelines were ef-

[9]The annotation tasks were performed in 2021, when the UK minimum wage was 8.91 British Pounds (GOV.UK, 2024).

Company	Sector	Sentences	Labelers
Monster Beverage	Consumer Staples	1291	3
Chevron	Energy	562	3
Netflix	Communication Services	648	1
Amazon	Consumer Discretionary	609	1
Sherwin- Williams	Materials	687	1

Table 5.3: Statistics for annual reports annotated with financial concept labels.

fective[10]. The final concept labels for the first two reports were chosen by majority voting.[11]

Final annotation Following Liakata and Soldatova (2008), the first two phases of annotation are used for quality assessment, and each subsequent report is annotated by just one annotator. Table 5.3 details the annotation statistics. In our experiments, we use a balanced sample of 540 sentences, with 90 sentences for each of the 6 financial concepts. The data was not released or stored in public servers to avoid potential contamination.

5.2.2.3 Models

In our experiments, we use the leading open-source and proprietary instruction-tuned language models currently available, covering a wide range of sizes (from 7B to 180B for open-source models). We focus on instruction-tuned models as non-instruct models require the task specification via in-context samples (Xie et al., 2021; Min et al., 2022), which in our early experiments resulted in poor performance when mixed with concept guidelines.

- **LLAMA-2** (Touvron et al., 2023), a family of open-source large language models that achieve state-of-the-art results at the moment of this writing. We use the LLAMA-2-CHAT variants (7B, 13B, and 70B parameters), which are pre-trained on 2 trillion tokens of data and fine-tuned via supervised fine-tuning and Rein-

[10]Due to the inherent ambiguity of the task, the agreement scores are moderate. We discuss this issue in the limitations section.
[11]Voting ties were adjudicated by the guideline's author.

forcement Learning with Human Feedback (RLHF). Unless otherwise stated, all mentions of LLAMA-2 in this work refer to the chat variants.

- **GPT-3.5** and **GPT-4** (Achiam et al., 2023), two proprietary models that offer the best instruction-following capabilities at the time of this writing. In our experiments, GPT-3.5 and GPT-4 refer to the `gpt-3.5-turbo-0613` and `gpt-4-0613` models respectively[12], which are invoked via the chat completions API[13].

- **FALCON-180B** (Almazrouei et al., 2023), a 180-billion language model trained on 3.5 trillion tokens from the RefinedWeb dataset (Penedo et al., 2023). We use the FALCON-180B-CHAT version that is fine-tuned on further instruction, question answering, and chat datasets. In contrast to the other language models above, it does not use Reinforcement Learning with Human Feedback (RLHF) in its fine-tuning phase.

5.2.2.4 Concept Definitions for LLMs

Ideally, we would use the same guidelines provided to humans (Figure 5.4) for annotation with LLMs. However, the human guidelines are not uniform across concepts and contain several examples and references to external content. Thus, to minimize confounding factors related to differences in definitions across concepts, we use model-generated concept definitions. Specifically, we prompt GPT-3.5 (refer to Section 5.2.2.3 for API usage details) to provide a short description of a concept in the context of a sentence annotation task. For scientific concepts, we use the following prompt:

```
We need to classify sentences in scientific articles
according to the information they convey: background,
motivation, method, results, or conclusion. Please
provide a short definition for each of those labels to
be used in annotation guidelines.
```

Then, we review and edit the definitions to remove explicit mentions of label names such as *A sentence is classified as "Motivation" when it explains (...)*. The final scientific concepts are provided in Table 5.4.

Similarly, we generate definitions for financial concepts using the prompt:

[12] https://platform.openai.com/docs/models
[13] https://platform.openai.com/docs/guides/gpt

```
We need to classify sentences in company disclosure
reports according to the capital information they convey:
financial, manufactured, intellectual, human, social
and relationship, or natural.  Based on the Integrated
Reporting framework, please provide a short definition
for each of those labels to be used in annotation
guidelines.
```

The financial concept definitions (after review) are provided in Table 5.5.

5.2.2.5 Concept Classification Details

For concept classification, we perform conditional generation (Eq. 5.4) using the Hugging Face transformers library (Wolf et al., 2020). To build the model inputs, we apply the prompt templates defined in Section 5.2.1, replacing the placeholders with the concept labels, definitions, the input sentence, and a concept domain indicator. Since the input sentences are short (around 30 words on average), no truncation is applied. We provide details on the prompts and inference parameters in Appendix A.3.

Finally, the classification prompts are wrapped into model-specific prompts. For the LLAMA-2 models, we use the following prompt:

```
[INST] {prompt_G}

{prompt_t} [/INST]
```

And for FALCON-180B, we use the following prompt:

```
User: {instruction} {prompt_G}

{prompt_t}
Falcon:
```

where $\texttt{prompt}_G$ and $\texttt{prompt}_t$ are defined in Section 5.2.1. Note that we do not use system prompts for both models, as we found that system prompts result in more verbose outputs.

Post-processing Since we use unconstrained generation for classification, in some instances the output includes extra dialog verbiage and even explanations for the predictions. By examining these outputs, we can gain more detailed insights into the model behavior, for instance, when it refuses to classify sentences with out-of-dictionary labels (refer to details in Section 5.2.3). To extract labels from those outputs, we apply a post-processing heuristic that checks if any of the labels is a substring of the output.

Concept	Definition
Background	A sentence that provides context, foundational knowledge, or relevant information about the research topic, existing theories, prior studies, or the broader scientific field in which the research is situated. It helps readers understand the background against which the research is conducted.
Motivation	A sentence that explains the reasons, objectives, or goals behind the research. It often includes statements about the research gap, the problem being addressed, the significance of the study, and why the research is important.
Method	A sentence that describes the research methods, techniques, procedures, and data collection processes used in the study. This category also encompasses details about the experimental design, data analysis, and any materials or instruments utilized.
Result	A sentence that presents the empirical findings, outcomes, observations, or data generated by the research. It includes quantitative and qualitative results, statistical analyses, tables, figures, and any other information related to the research findings.
Conclusion	A sentence that summarizes the key takeaways, implications, interpretations, or insights derived from the study's results. It often discusses the broader significance of the findings, suggests future research directions, and may reiterate the study's contributions to the field.

Table 5.4: Scientific concept definitions used in sentence classification guidelines.

Concept	Definition
Financial	A sentence that pertains to monetary resources, assets, liabilities, revenues, expenses, or any other financial information related to the company's operations, investments, and financial performance.
Manufactured	A sentence that refers to physical assets, infrastructure, and tangible resources such as buildings, machinery, equipment, or any other manufactured or constructed items that contribute to the company's value.
Intellectual	A sentence that relates to intangible assets, knowledge, intellectual property, patents, trademarks, copyrights, research and development activities, or any other intellectual assets that enhance the company's competitiveness and innovation.
Human	A sentence that involves information about the company's workforce, including employees, skills, expertise, training, recruitment, talent development, and any other human resources aspects that contribute to the company's success.
Social and relationship	A sentence that deals with the company's relationships and interactions with external stakeholders, communities, customers, suppliers, partners, and any other social or relationship-based assets that affect the company's operations and reputation.
Natural	A sentence that addresses environmental resources, sustainability efforts, ecological impacts, conservation initiatives, or any other aspects related to the company's use of natural resources and its environmental responsibility.

Table 5.5: Financial concept definitions used in sentence classification guidelines.

If there is a single substring that meets this requirement, it is considered as the prediction. Finally, for OOD guidelines, we replace the OOD label predictions with the corresponding factual labels, so the performance metrics can be computed with respect to the ground-truth labels.

5.2.3 Results and Discussion

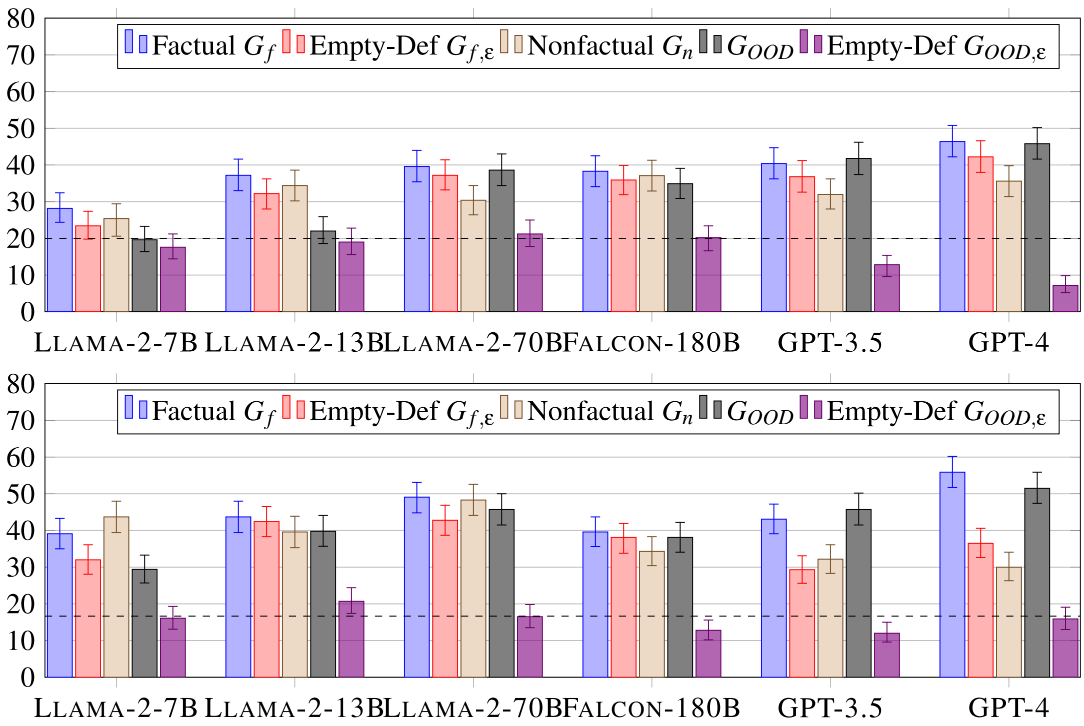

Figure 5.6: Concept classification accuracy for different **scientific (top)** and **financial (bottom)** concept guidelines. In this experiment, the nonfactual guideline G_n is a random permutation where *all concept definitions* are nonfactual. *Empty-Def* refers to the empty-definition factual ($G_{f,\varepsilon}$) and out-of-vocabulary guidelines ($G_{OOD,\varepsilon}$). Error bars represent the 95% confidence interval and the dashed line indicates the random classifier baseline.

LLMs Leverage Concept Labels and Definitions Using a factual guideline G_f as a reference, results from Figure 5.6 show that removing concept definitions (guideline $G_{f,\varepsilon}$) reduces consistently the accuracy of concept classification. However, the classification performance without concept definitions is still significantly higher than the random baseline, which suggests that the models have relevant prior knowledge related

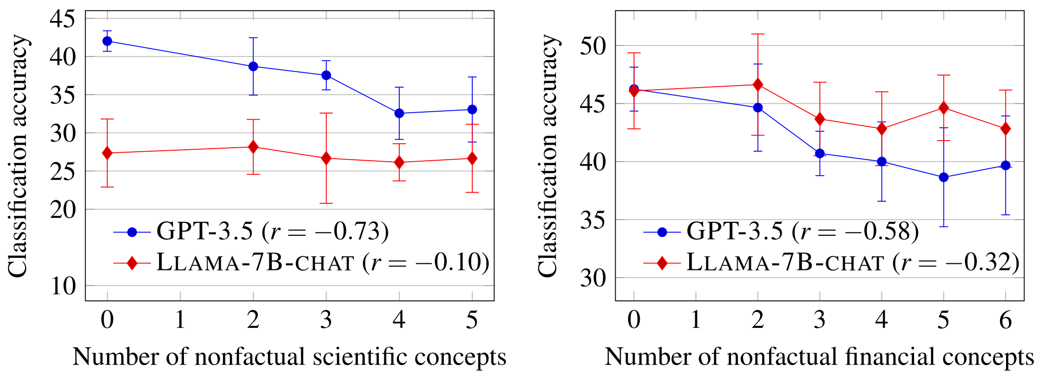

Figure 5.7: Concept classification accuracy results for different levels of nonfactuality of **scientific (left)** and **financial (right)** concept guidelines. We sample 10 guidelines for each nonfactuality level and average the classification accuracies. Error bars represent the standard deviations.

to the concept label lexical information. The average accuracy loss when removing concept definitions is 3.7% and 8.2% for scientific and financial concepts respectively, which indicates that financial definitions have a stronger influence on model predictions.

Nonfactual Understanding Emerges with Scaling As pointed out above, the lexical information from both concept labels and definitions contributes to task performance. However, we want to verify if the *associations* between labels and definitions are relevant. In Figure 5.6, we observe that the smaller LLAMA-2-7B and LLAMA-2-13B models have a similar performance under the factual G_f and nonfactual guideline G_n settings. In contrast, there is a consistent drop in accuracy for LLAMA-2-70B, GPT-3.5, and GPT-4, which indicates that these models are effectively changing the labels according to the nonfactual semantics. Despite having more than two times the number of parameters of LLAMA-2-70B, FALCON-180B behaves similarly to the smaller LLAMA-2 models when conditioned with the nonfactual guideline. In this case, scaling is not a sufficient condition to improve understanding in nonfactual contexts.

As further evidence of the capacity of GPT models to follow guidelines, we sample nonfactual concept guidelines such that they are balanced with respect to the number of nonfactual concepts. Then, we evaluate the classification performance for each guideline on the same data samples and average the results for guidelines with the same number of nonfactual concepts. The curves in Figure 5.7 show decreasing classification performance as the scientific guidelines become more nonfactual, with GPT-3.5

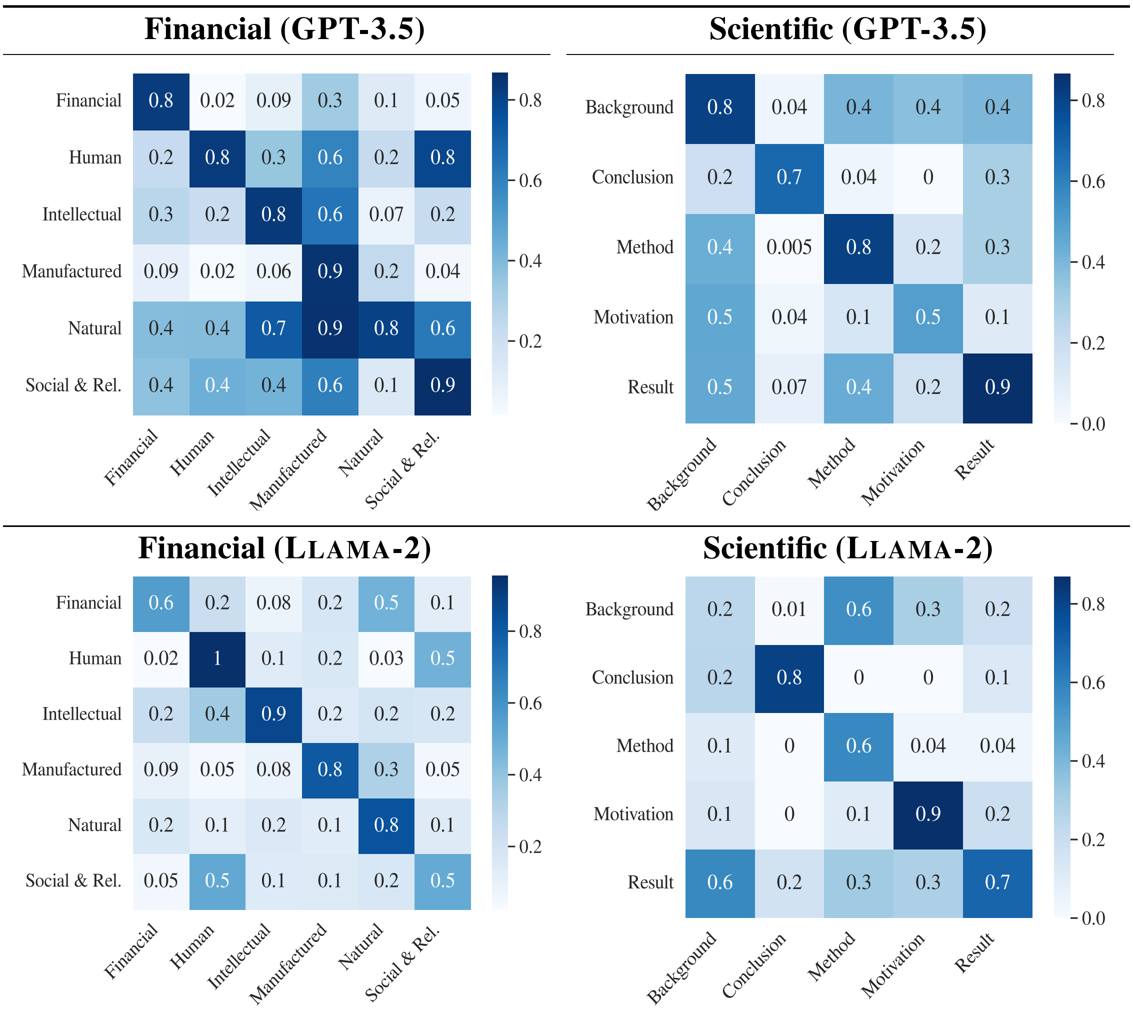

Figure 5.8: Guideline adherence scores per **financial** and **scientific** concept for GPT-3.5 and LLAMA-2. Each cell A_{ij} shows the fraction of concept predictions that adhere to definitions $\delta(c_j) = d_i$, where the rows indicate original factual labels c_i that are randomly replaced by labels c_j (columns). Off-diagonal results indicate nonfactual definitions.

results having a stronger Pearson correlation of -0.73 compared to -0.10 for LLAMA-2-7B. A similar trend is observed for financial guidelines, as additional evidence that GPT-3.5 is more sensitive to nonfactual guidelines. However, the ability to adhere to nonfactual guidelines is not uniform across concepts. In Figure 5.8, we observe that some concept changes (e.g., from scientific RESULTS to CONCLUSION) are followed much less frequently. We hypothesize that the semantic similarity between concepts may impact the accuracy in nonfactual settings. We leave the study of such factors for future work.

Larger Models Can Rename Existing Concepts We consider the effects of removing the lexical information from concept labels by using out-of-dictionary labels (G_{OOD} guideline). Again, the largest models (70B or more parameters) tend to perform on par with the original factual guideline G_f across both domains. Even though the models are not learning entirely new concepts, it is remarkable that they can associate novel labels with abstract concepts and leverage them to solve tasks. We believe this ability might be relevant for natural language reasoning problems that require symbolic formulation (Pan et al., 2023).

Proprietary Models Recognize Unknown Concepts While LLAMA-2-70B has performance similar to GPT-3.5 on most settings, when presented with out-of-dictionary (OOD) labels without definitions ($G_{\text{OOD},\varepsilon}$ guideline), it predicts labels randomly. This result confirms that OOD labels provide no information related to scientific or financial concepts. However, GPT-3.5 and GPT-4 behave differently, often refusing to assign concepts to the sentences and instead generating outputs such as *None of the categories listed above are appropriate for classifying the given text*. For instance, GPT-3.5 refuses to classify 58% and 51% of sentences from scientific and financial documents respectively, while the open-source models always predict one of the nonsensical labels. We hypothesize that this ability to recognize unknown concepts is derived from careful alignment efforts (Ouyang et al., 2022), which presents an avenue for improving open-source language models.

Agreement with Human Annotators Using sample sentences from the second report of the financial annotation, we measure the agreement of the models' financial concept predictions (using factual concept guidelines) to each human annotator. We find that LLAMA-2-7B and GPT-4 achieve average Cohen's κ scores on par with expert annotators (Table 5.6). This result is in line with previous work showing that LLMs can be a useful tool in annotation pipelines (Wang et al., 2021).

5.3 Related Work

The summarization of scientific articles informed by sentence concepts (or rhetorical status) is explored by Teufel and Moens (2002). Their approach consists of one for sentence relevance and a second classifier for concepts such as paper *aim*, *background*, and *contrast* to previous research. With the output of the classifiers, they can tailor

Model	Annotation Agreement			
	A_1	A_2	A_3	Avg
Human Average	0.46	0.43	**0.37**	0.42
LLAMA-2-7B	0.46	0.47	0.35	0.43
GPT-4	**0.47**	**0.53**	0.35	**0.45**

Table 5.6: Financial concept annotation agreement to annotators A_1, A_2, and A_3. Results are non-weighted Cohen's κ on a subset of sentences for which human annotators assigned at least one capital concept.

summaries for nonexpert readers focusing on background information. Using a different concept classification scheme, the CoreSC, Liakata et al. (2012) proposes a similar solution for extractive summarization of chemistry papers.

Previous work examined if LLMs exhibit human-like conceptual *grounding* (Piantadosi and Hill, 2022). Patel and Pavlick (2021) demonstrate that LLMs such as GPT-3 can generalize spatial and color concepts in some settings. Using several nonfactual reasoning tasks such as arithmetic, chess, and drawing Wu et al. (2023) show that some proprietary models have limited capacity for reasoning under nonfactual conditions. Our work explores concept classification tasks that are more abstract than spatial concepts but still simpler than the more complex tasks proposed by Wu et al. (2023). As a consequence, we can more precisely control the level of nonfactuality of the tasks while keeping the same level of difficulty.

A variety of approaches related to editing factual knowledge of LLMs has been explored recently (Onoe et al., 2023; Meng et al., 2022; Zhu et al., 2020). This line of work proposes different ways to edit memory related to entities and assess if the model outputs in different contexts are consistent with the newly introduced facts. Those approaches focus on updating model parameters while we examine model behavior under in-context concept edits. Min et al. (2022) study the role of in-context demonstrations for various classification tasks. They conclude that associations of samples and labels do not strongly influence the model's performance, suggesting that non-instruct LLMs cannot learn new information from the demonstrations. In contrast, our results suggest that instruction-tuned models are consistently influenced by in-context concept definitions.

Our evaluation protocol is similar to Wei et al. (2023) work as they use flipped and "semantically-unrelated" labels in task demonstrations. While they focus on in-context learning, our experiments are zero-shot tasks including *only concept definitions*. Thus, our setting is arguably harder, requiring the models to generalize from guidelines (not examples) that are relatively agnostic with respect to the classification task. Furthermore, we put a significant effort into evaluating open-source models that are state-of-the-art at the time of this submission. To our knowledge, this kind of evaluation is not addressed by previous work and is relevant to inform the improvement of open-source initiatives.

The potential of LLMs as zero-shot and few-shot data annotators has been demonstrated in medical (Agrawal et al., 2022), social science (Zhu et al., 2023), and other language understanding tasks (Wang et al., 2021). Our work provide further evidence that instruction-tuned models can perform concept classification with agreement scores comparable to expert annotators. Additionally, we show that similarly to humans, LLMs can leverage concept guidelines to improve the annotation quality.

5.4 Limitations

Opacity of Proprietary Models The experimental results from Section 5.2.3 confirm that the proprietary models excel in almost all classification settings. However, we cannot determine if the main cause for the best performance is the scale, training data, or fine-tuning methods since we do not have access to their implementation details.

Inference costs for Large Models Our experiments are severely limited by the computing requirements of the larger open-source LLM models. For instance, FALCON-180B-CHAT requires around 400GB of memory for inference, equivalent to 5 Nvidia A100-80GB GPUs. Thus, we limit our nonfactual guideline experiments (Figure 5.7) to include only a subset of possible permutations for LLAMA-2-7B and GPT-3.5.

Consequences of Nonfactual Performance to Other Tasks In this work, we measure the capacity of several language models to work under nonfactual contexts. Future investigation efforts could explore how this ability correlates to a potential reduction of hallucinations in generative tasks or even improved performance in natural language reasoning problems (Pan et al., 2023).

Financial Annotation Agreement Due to the ambiguity of financial annotation the task, the agreement scores we report in Section 5.2.2.2 are relatively moderate (average $\kappa = 0.47$ for the second round in Table 5.2). One of the main factors for disagreement is that some sentences are complex and may contain multiple capitals, as illustrated in the following example (passages conveying capitals are underlined):

> *Such factors include the duration and scope of the pandemic, including any resurgences of the pandemic, and the impact on our workforce and operations; the negative impact of the pandemic on the economy and economic activity, including travel restrictions and prolonged low demand for our products; the ability of our affiliates, suppliers and partners to successfully navigate the impacts of the pandemic; the actions taken by governments, businesses and individuals in response to the pandemic; the actions of OPEC and other countries that otherwise impact supply and demand and correspondingly, commodity prices; the extent and duration of recovery of economies and demand for our products after the pandemic subsides; and Chevron's ability to keep its cost model in line with changing demand for our products.*

In many cases, annotators chose a non-intersecting subset of the capitals, which counts as a disagreement (even though both are partially correct). Those voting ties were reviewed and adjudicated by the author of the guidelines. Previous scientific annotation projects like (Liakata et al., 2012) also report a moderate agreement ($\kappa = 0.55$, median of the best annotators), which demonstrates the difficulty in annotating technical documents.

Finally, even though report agreements between LLM annotations and humans in Table 5.6, our experiments are not designed to fairly compare annotation quality. Before annotation, humans received training and guidelines that are more comprehensive than LLM guidelines. Secondly, humans were able to "calibrate" their labels according to previously annotated sentences, whereas LLMs do not have access to this memory.

5.5 Conclusion

Our experiments on guiding the coverage of domain concepts indicate that structuring a complex summarization task into smaller and more tractable steps is a promising direction. The *annotate-plan-edit* (APE) pipeline provides a way to control the emphasis of specific scientific concepts in summaries and also, a means to interpret the model understanding of the documents by examining the intermediate concept notes. Our investigation reveals gaps in concept coverage between human and machine-written summaries, which motivated a comprehensive assessment of scientific concepts by large language models.

By using factual and nonfactual concept guidelines for sentence classification, we demonstrate measurable gaps in concept understanding between leading open-source and proprietary instruction-tuned models. While some level of nonfactual concept understanding emerges with scaling, open-source models cannot recognize nonsensical (out-of-dictionary) guidelines, which the closed APIs can address more consistently. One question to be addressed in future work would be to investigate potential correlations between the capacity of reasoning in nonfactual contexts and other common generation issues such as hallucination.

Chapter 6

Conclusions

This thesis investigates techniques for controllable summarization of long technical documents, focusing on the scientific domain. By examining the recent progress of single-document summarization in Chapter 2, we learn that summarization (as other fields in Artificial Intelligence) seemingly does not escape the *bitter lesson*: simple general learning principles (at scale) tend to be more effective in the long term than summarization-specific inductive biases. However, scaling alone cannot account for subjective and extrinsic factors that constrain the summarization tasks. This reality calls for innovations in how we formulate and evaluate summarization tasks.

In this setting, the *context factors* proposed by Jones (1998) provide a useful framework for constraining summarization tasks not only in terms of lexical similarity to reference summaries but more nuanced *purpose* factors of summarization. To realize this view in the age of large neural language models, we propose a simple energy-based model of summarization in which we shift the modeling of context factors from learning to inference. The investigation includes explicit optimization procedures in a factorized energy-based model or the direct specification of context factors via textual prompts for language models.

Chapter 3 addresses the first research question, *how to adapt a summarization model to diverse context factors at inference time*. It introduces FACTORSUM, a model that separates training and optimization with respect to context factors such as summary budget and content coverage. This factorized model achieves superior lexical alignment with reference scientific abstracts compared to models with up to 4 times more parameters. Additionally, FACTORSUM demonstrates promising domain adaptation when using a model trained on scientific articles to summarize government reports without further in-domain training. This method was also applied to legal documents

(Malik et al., 2024) and lay summarization of biomedical research articles (Phan et al., 2023; Goldsack et al., 2023), demonstrating the flexibility of the factorized model.

The next set of experiments investigates *whether large language models can adapt to diverse styles of scientific summarization* including abstracts, peer reviews, and lay summaries. The results from Chapter 4 indicate a competitive performance of LLMs as measured by ROUGE, compared to strong supervised baselines trained specifically for abstract generation. Also, the human evaluation experiments show a significant preference for machine-generated summaries, mostly due to better coverage of scientific concepts. Moreover, we show how to improve lexical alignment controllability in LLM summaries by using keywords extracted from FACTORSUM summaries. This approach balances the token-level alignment of FACTORSUM with the high fluency of LLM summaries.

Lastly, we explore the question of *how to control the coverage of scientific concepts in lay summaries*. In Chapter 5, we design a *annotate-plan-edit* (APE) model that breaks summarization into per-concept sub-tasks. Our findings indicate that large language models cannot reproduce a similar distribution of concepts (with a strong emphasis on scientific background information) as found in human-written lay summaries. These results motivate a further investigation of the capacity of LLMs to identify scientific and financial concepts in technical documents. To this end, we design sentence classification experiments using state-of-the-art LLMs subject to factual and perturbed concept definitions. As expected, our results point out that model scale correlates with concept understanding. However, increasing scale alone is not sufficient to achieve high performance, as we find that the Falcon model with 180 billion parameters is outperformed by some of the smaller open-source counterparts. Finally, there are some counterfactual settings (nonsensical tasks) that can only be handled by proprietary APIs, which indicates there is still a gap in task understanding compared to open-source models.

Our research provides evidence of the effectiveness of large language models in summarizing long technical documents. However, we also find relevant limitations related to fine-grained control of concept coverage in lay summaries that deviate from a typical summary in the training data. Furthermore, we found cases where small supervised models like FACTORSUM are still useful, especially in contexts where large-scale models are not feasible or more precise control of surface-level properties of summaries is required.

Leveraging context factors provides a promising avenue for future summarization models and evaluation benchmarks. The summarization directions depicted by Jones (1998) remain relevant as an open research question, which is, fortunately, more tractable due to the recent developments in language modeling.

6.1 Limitations and Future Work

The current research space related to generative language modeling is extremely dynamic, and many of the model capabilities and limitations we present in this thesis might demand constant reevaluation as novel (and even larger) large language models are released. In this section, we highlight some of the research directions that we believe will remain relevant to improve summarization systems.

Focus on summary purpose As evaluation protocols based on linguistic fluency and general-purpose importance become obsolete, it is expected that actual summary goals and user needs will be explored in more detail. As mentioned in Chapter 2, recent work conducted surveys among summary readers in universities to find out what is missing in summaries produced by current systems, and how language simplification affects the perception of quality (Ter Hoeve et al., 2022; August et al., 2024). It is likely that more such field research based on context factors will inform future summarization developments.

Crowdsourced comparative benchmarks One of the main challenges we found in evaluating summarization models refers to scaling human evaluation. Especially for technical documents such as scientific papers, the evaluators must be at least moderately knowledgeable about the research topic, and a slight shift in the domain may affect the quality of judgments substantially. Other issues include the length of documents and the inherent subjectivity of the task. A possible solution to this problem is using crowdsourced benchmarks based on blind comparisons between model outputs (Chiang et al., 2024) and ranking estimation using the Bradley-Terry model (Bradley and Terry, 1952). Although most current leaderboards of this kind are general-purpose evaluations, we expect that derived versions will be developed targeting specific quality dimensions (e.g., *which paper abstract is the best suited to a lay audience?*) or communities, such as computer scientists or financial specialists.

Cognitive and socially-grounded models Textual information in scientific articles reflects a real-world process that involves resources and people with different interests within complex organizations. Modeling this context and also the cognitive processes involved in crafting a summary (Cardenas et al., 2024) might improve the quality and interpretability of the model outputs.

Multimodality This thesis and the majority of research in summarization concentrates on textual data. However long documents such as research papers and financial reports contain relevant information represented in other modalities such as tables and diagrams, which could be leveraged by multimodal large language models (Koh et al., 2023). More generally, a summarization model would not only consume multiple modalities as input but also generate multimodal summaries (Zhu et al., 2018; He et al., 2023), potentially increasing their informativeness.

Appendix A

Additional Results and Implementation Details

In this appendix, we present additional validation results, sample summaries, and model hyperparameters for models covered in chapters 3 and 4.

A.1 Validation Results for FactorSum

Table A.1 shows the validation scores for the ensemble experiments corresponding to the test results in Table 3.7.

Ranker	PubMed			arXiv		
	R-1	**R-2**	**R-L**	**R-1**	**R-2**	**R-L**
PEGASUS + BigBird Ensemble Summaries						
TextRank	44.02	18.40	38.73	44.20	16.97	37.66
FactorSum	44.83	**19.12**	40.90	44.77	17.36	39.84
FactorSum-Oracle	48.77	21.72	44.40	49.18	20.07	43.72
Summary Views						
TextRank	42.17	16.82	37.62	42.49	16.39	37.58
FactorSum	**45.33**	18.69	**41.62**	**47.16**	**18.57**	**42.57**
FactorSum-Oracle	51.64	23.27	47.48	53.27	22.75	48.08

Table A.1: ROUGE F1 scores on the **validation sets** for the ensemble experiments.

Additionally, in Table A.2, we provide validation scores corresponding to the in-domain summarization test results in Table 3.6.

Model		PubMed				arXiv				GovReport			
		R-1	R-2	R-L	Len	R-1	R-2	R-L	Len	R-1	R-2	R-L	Len
Previous work													
PEGASUS		43.73	18.77	40.15	181	43.07	16.39	38.66	170	-	-	-	-
BigBird		45.28	19.77	41.60	186	46.02	18.54	41.35	164	-	-	-	-
BART-large		-	-	-	-	-	-	-	-	53.06	19.11	50.12	597
Guidance													
Budget	**Content**	**FactorSum - no content guidance**											
Oracle	-	47.37	19.16	43.33	210	48.85	18.85	43.92	165	59.54	23.98	55.82	642
Fixed	-	45.33	18.69	41.62	205	47.16	18.57	42.57	165	58.41	23.90	54.83	650
Model	-	44.71	18.07	40.84	185	46.35	18.22	41.79	165	57.55	23.68	53.92	639
FactorSum - content guidance													
Oracle	Lead	48.19	19.99	44.25	205	49.61	19.27	44.75	164	60.47	24.93	56.89	649
Fixed	Lead	46.10	19.21	42.43	202	47.98	18.99	43.41	165	59.19	24.72	55.78	649
Fixed	Model	**47.20**	**20.17**	**43.48**	197	**49.16**	**20.17**	**44.59**	164	**60.00**	**25.33**	**56.49**	648
Model	Model	47.00	20.15	43.19	179	48.17	20.06	44.07	164	59.35	25.25	55.84	641

Table A.2: ROUGE F1 scores and average words per summary on the **validation sets** for different types of guidance during inference. *Lead* guidance is the first k sentences from the source document (Section 3.3.2.2). Model guidance is provided by BART-large for GovReport and BigBird for PubMed and arXiv. The choice of budget guidance values is described in Appendix 3.2.2.1. Results for models marked with † are taken from the original publications. Underlined results are statistically equivalent to the best methods ($p < 0.05$).

A.2 Summarization Model Parameters

Table A.3 lists the model hyperparameters for the experiments described in Chapter 4.

BIGBIRD	
Checkpoint (arXiv)	bigbird-pegasus-large-arxiv
Checkpoint (PubMed)	bigbird-pegasus-large-pubmed
Number of parameters	576.9M
Max source length	3072
Generation beams	5
Length penalty	0.8
FACTORSUM	
Checkpoints	bart-base
Number of parameters	139.4M
Max source length	1024
Max source length	128
Generation beams	4
Length penalty	1.0
LLAMA-2	
Checkpoint	Llama-2-7b-chat-hf
# parameters	7B
Max context length	2048
Parameter type	float16
Nucleus temperature	0.8
Nucleus top-*p*	0.95
GPT-3.5	
Model	gpt-3.5-turbo-0301
temperature	1
top_p	1
presence_penalty	0
frequency_penalty	0
All models (except FACTORSUM)	
Max target length (arXiv, MuP)	256
Max target length (PubMed, eLife)	512

Table A.3: Model details and generation parameters.

A.3 Concept Classification Model Parameters

Table A.4 lists the model hyperparameters for the experiments described in Chapter 5.

LLAMA-2	
Number of parameters	7B / 13B / 70B
Max context length	4096
FALCON-180B-CHAT	
Number of parameters	180B
Max context length	2048
LLAMA-2 and **FALCON-180B-CHAT**	
Parameter type	float16
Nucleus temperature	0.8
Nucleus top-p	0.95
GPT-3.5 and **GPT-4**	
Model GPT-3.5	gpt-3.5-turbo-0611
Model GPT-4	gpt-4-0611
temperature	1
top_p	1
presence_penalty	0
frequency_penalty	0
All models	
Max generation tokens	128

Table A.4: Summary of generation details and parameters.

A.4 Sample Summaries for FACTORSUM

In Tables A.5 to A.16, we list samples of summaries from PubMed, arXiv, and GovReport test sets. We compare BigBird and BART summaries to FACTORSUM with lead and model-provided content guidance. Budget guidance is fixed to the same values as described in Section 3.3.2.

Reference abstract (Tokens: 156)
we report a 2.5-year - old girl who presented with hoarseness of voice since 3 months of age and failure to thrive . chest x - ray showed cardiomegaly with a deviation of the trachea and mediastinum to the right side . two - dimensional echocardiography showed decreased flow across the right pulmonary artery , a small atrial septal defect (asd) with a right - to - left shunt , and a dilated right atrium and right ventricle with severe tricuspid regurgitation suggestive of severe pulmonary hypertension . a silent large patent ductus arteriosus was also seen . multiple detector computerized tomography aortogram confirmed the findings of absent right pulmonary artery and hypoplastic right lung with small cystic lesions suggestive of congenital cystic adenomatoid malformation in the right lower lobe .
BigBird summary Tokens: 48; ROUGE-1: 45.16; ROUGE-2: 31.52; ROUGE-L: 39.78
we report a 2.5-year - old girl with congenital absence of the right pulmonary artery with associated congenital cystic adenomatoid malformation of the right lower lobe, patent ductus arteriosus, and atrial septal defect, who presented with ortner's syndrome due to severe pulmonary hypertension. .

Table A.5: Sample abstract and BigBird summary from the PubMed test set (ID = 5836).

FactorSum - Fixed budget and Lead content guidance Tokens: 199; ROUGE-1: 47.17; ROUGE-2: 26.58; ROUGE-L: 44.65
unilateral absence of pulmonary artery (uapa) is a rare congenital abnormality , with an estimated prevalence of 1 in 200,000. while some patients with uapa are totally asymptomatic , others may have severe pulmonary hypertension. we report a 2.5-year - old girl with congenital absence of the right pulmonary artery with associated congenital cystic adenomatoid malformation (ccam) , patent ductus arteriosus (pda) , and atrial septal defect , who presented with ortner 's syndrome due to severe pulmonary hypertension. a 2.5-year - old girl presented with hoarseness of voice noticed since 3 months of age , breathlessness for the past 15 days and failure to thrive. pulmonary agenesis is usually unilateral , right sided absence of pulmonary artery being more common. the syndrome has since been described in adults with various cardiovascular disorders , but reports in children are less common. congenital cystic adenomatoid malformation (ccam) is a rare cause of congenital cyanotic heart disease. the child was intubated and ventilated and started on pressors , but sustained a cardiac arrest on the 4th hospital day from which she could not be resuscitated.
FactorSum - Fixed budget and BigBird content guidance Tokens: 142; ROUGE-1: 42.42; ROUGE-2: 23.66; ROUGE-L: 38.64
congenital cystic adenomatoid malformation (ccam) is a rare cause of congenital cyanotic heart disease. pulmonary agenesis is usually unilateral , right sided absence of pulmonary artery being more common. the syndrome has since been described in adults with various cardiovascular disorders , but reports in children are less common. unilateral absence of pulmonary artery (uapa) is a rare congenital abnormality , with an estimated prevalence of 1 in 200,000. we report a 2.5-year - old girl with congenital absence of the right pulmonary artery with associated congenital cystic adenomatoid malformation (ccam) , patent ductus arteriosus (pda) , and atrial septal defect, who presented with ortner 's syndrome due to severe pulmonary hypertension. while some patients with uapa are totally asymptomatic , others may have severe pulmonary hypertension.

Table A.6: Summaries generated by FactorSum for a sample from PubMed test set (ID = 5836). Reference abstract is presented in Table A.5.

Reference abstract (Tokens: 240)
the navier - stokes - fourier theory of viscous , heat - conducting fluids provides parabolic equations and thus predicts infinite pulse speeds . naturally this feature has disqualified the theory for relativistic thermodynamics which must insist on finite speeds and , moreover , on speeds smaller than c. the attempts at a remedy have proved heuristically important for a new systematic type of thermodynamics : extended thermodynamics . that new theory has symmetric hyperbolic field equations and thus it provides finite pulse speeds.extended thermodynamics is a whole hierarchy of theories with an increasing number of fields when gradients and rates of thermodynamic processes become steeper and faster . the first stage in this hierarchy is the 14-field theory which may already be a useful tool for the relativist in many applications . the 14 fields and further fields are conveniently chosen from the moments of the kinetic theory of gases.the hierarchy is complete only when the number of fields tends to infinity . in that case. the pulse speed of non - relativistic extended thermodynamics tends to infinity while the pulse speed of relativistic extended thermodynamics tends to c , the speed of light. [...]
BigBird summary Tokens: 105; ROUGE-1: 37.27; ROUGE-2: 13.75; ROUGE-L: 32.29
the paradox of pulse speeds in extended thermodynamics has been known for 50 years. it seems to have been caused by eckart s theory of irreversible processes which assumed a constant pulse speed and a fixed temperature. recently it has been shown by boillat & ruggeri that, as the number of moments increases, the pulse speed tends to infinity in the non - relativistic kinetic theory of gases and the relativistic case by which the pulse speed tends to c. these results put an end to the long - standing paradox of pulse speeds. they are reviewed in detail.

Table A.7: Sample summaries from the PubMed test set (ID = 1475).

FACTORSUM - fixed budget and Lead content guidance Tokens: 216; ROUGE-1: 50.12; ROUGE-2: 14.18; ROUGE-L: 44.77
the pulse speed problem is one of the most important questions in thermodynamics , but it is a question that can be answered , and has to be answered. we derive a set of thermodynamic processes for a , a , b , c , and d. in this paper , we prove that a co - vector exists in which the entropy is a function of the thermal equation of the state. the thermodynamics of viscous , heat - conducting gases is studied by means of the determination of the 14 fields of the field equations. in this paper , we review the recent developments in the field of non - equilibrium thermodynamics. it is possible , and indeed common , to make a specific choice for the fields u and the concavity postulate is contingent upon that choice. the first moments in the kinetic theory of gases are obtained from a homogeneous system where the acceleration waves and their speeds of propagation are to be calculated from the homogeneous systems. the heat fluxes f(x , p , t) of the atoms , viz. the paper deals with the thermodynamics of a non - degenerate gas. in the non - relativistic limit ,
FACTORSUM - fixed budget and BigBird content guidance Tokens: 190; ROUGE-1: 44.67; ROUGE-2: 12.75; ROUGE-L: 42.64
it is then a simple problem of linear algebra to prove that the entropy density h = ha is concave as a function of f. the pulse speed problem is one of the most important questions in thermodynamics , but it is a question that can be answered , and has to be answered. it is possible , and indeed common , to make a specific choice for the fields u and the concavity postulate is contingent upon that choice. the first moments in the kinetic theory of gases are obtained from a homogeneous system where the acceleration waves and their speeds of propagation are to be calculated from the homogeneous systems. we derive a set of thermodynamic processes for a , a , b , c , and d. in this paper , we review the recent developments in non - relativistic kinetic theory. this paper presents the results of an investigation of the non - degenerate gases in the system. we conclude that no paradox of infinite speeds can arise in extended thermodynamics. in the non - relativistic limit ,

Table A.8: Summaries generated by FACTORSUM for a sample from PubMed test set (ID = 1475). Reference abstract is presented in Table A.7.

Reference abstract (Tokens: 157)
the metrization of the space of neural responses is an ongoing research program seeking to find natural ways to describe , in geometrical terms , the sets of possible activities in the brain . one component of this program are the _ spike metric , notions of distance between two spike trains recorded from a neuron . alignment spike metrics work by identifying " equivalent " spikes in one train and the other . we present an alignment spike metric having @xmath0 underlying geometrical structure ; the @xmath1 version is euclidean and is suitable for further embedding in euclidean spaces by multidimensional scaling methods or related procedures . we show how to implement a fast algorithm for the computation of this metric based on bipartite graph matching theory . @xmath2center for studies in physics and biology , rockefeller university , new york ny @xmath3harvard faculty of arts and sciences , cambridge ma
BigBird summary Tokens: 89; ROUGE-1: 28.71; ROUGE-2: 3.00; ROUGE-L: 22.77
we propose a spike metric that is consistent with the time - coding hypothesis of spike generation and has all of the desirable properties of an @xmath0 norm . when @xmath1 , this metric is equal to the victor - purpura metric @xcite . when @xmath2 , this metric is equal to the hungarian algorithm @xcite , @xcite , @xcite . when @xmath3 , this metric is equal to the monge algorithm @xcite , @xcite , @xcite .

Table A.9: Sample summaries from the arXiv test set (ID = 5946).

FactorSum - Fixed budget and Lead content guidance Tokens: 165; ROUGE-1: 42.51; ROUGE-2: 6.32; ROUGE-L: 38.33s
we consider the problem of finding a minimum weight matching on a bipartite graph such that each edge of the graph has a weight or cost. the value of the metric between two spike trains is the @xmath0 norm of the difference between their estimated rate functions . this metric preserves the integrity of individual spikes instead of viewing them as contributions to a rate function . we present a spike metric satisfying two important desiderata : that it be grounded in the time - coding hypothesis of spike generation , and that it is closely related to the euclidean @xmath1 norm . our proof uses the tools of graph theory and combinatorial optimization , and demonstrates that they can be usefully brought to bear on problems in computational biology. the cost of aligning two spikes on different trains is @xmath0. we present a new metric for comparing two spike trains , which we call the hungarian metric.
FactorSum - Fixed budget and BigBird content guidance Tokens: 150; ROUGE-1: 41.61; ROUGE-2: 5.88; ROUGE-L: 38.69
we present a spike metric satisfying two important desiderata : that it be grounded in the time - coding hypothesis of spike generation , and that it is closely related to the euclidean @xmath1 norm . the value of the metric between two spike trains is the @xmath0 norm of the difference between their estimated rate functions . the hungarian algorithm is a special case of algorithms to solve general matching problems and assignment problems. the metric is equal to the sum of the costs incurred by the most efficient alignment. we consider the problem of finding a minimum weight matching on a bipartite graph such that each edge of the graph has a weight or cost. we present a new metric for comparing two spike trains , which we call the hungarian metric. our metric is a minimization over all possible matchings .

Table A.10: Summaries generated by FactorSum for a sample from arXiv test set (ID = 5946). Reference abstract is presented in Table A.9.

Reference abstract (Tokens: 145)
in this paper we discuss the chemical evolution of elliptical galaxies and its consequences on the evolution of the intracluster medium (icm) . we use chemical evolution models taking into account dark matter halos and compare the results with previous models where dark matter was not considered . in particular , we examine the evolution of the abundances of some relevant heavy elements such as oxygen , magnesium and iron and conclude that models including dark matter halos and an initial mass function (imf) containing more massive stars than the salpeter (1955) imf , better reproduce the observed abundances of mg and fe both in the stellar populations and in the icm (asca results) . we also discuss the origin of gas in galaxy clusters and conclude that most of it should have a primordial origin .
BigBird summary Tokens: 109; ROUGE-1: 36.21; ROUGE-2: 9.57; ROUGE-L: 28.45
we discuss a model for the chemical evolution of elliptical galaxies in which supernovae (sne)-driven galactic winds play an important role in the formation of these objects . in this model the star formation is assumed to stop after the occurrence of the galactic wind and the galaxy evolves passively thereafter . the star formation is assumed to stop after the occurrence of the galactic wind and the galaxy evolves passively thereafter . the model includes the most recent ideas on sn progenitors and nucleosynthesis , indicating that sne ia originate from long living stars whereas sne of type ii originate from short living stars .

Table A.11: Sample summaries from the arXiv test set (ID = 6213).

FactorSum - fixed budget and Lead content guidance Tokens: 159; ROUGE-1: 43.73; ROUGE-2: 12.27; ROUGE-L: 39.43
this is a very strong conclusion since it implies a very fast process for the formation of big ellipticals at variance with the hierarchical clustering scenario for galaxy formation. we show how abundance ratios in stellar populations and gas in ellipticals can be used to constrain the amount and concentration of dark matter in these objects. in order to reproduce realistic galaxies , namely with the right colors and luminosities. we discuss the chemical evolution of elliptical galaxies in the framework of a simple model based on the idea that the efficiency of star formation should be inversely proportional to the dynamical timescale. we find that the efficiency of star formation in elliptical galaxies is inversely proportional to the dynamical timescale. in particular , we show that it is not possible to explain the increase of the [mg / fe] ratio in the nuclei of ellipticals as a function of galactic luminosity.
FactorSum - fixed budget and BigBird content guidance Tokens: 147; ROUGE-1: 43.70; ROUGE-2: 14.93; ROUGE-L: 39.26
we discuss the chemical evolution of elliptical galaxies in the framework of a simple model based on the idea that the efficiency of star formation should be inversely proportional to the dynamical timescale. in particular , we show that the efficiency of star formation increases with the total mass of the galaxy and that the more massive galaxies develop a galactic wind before the less massive ones. we show that the presence of dark matter in elliptical galaxies plays a crucial role in determining the onset and the entity of galactic winds. in this paper we discuss the possibility of an inverse wind scenario for the formation of elliptical galaxies. the model includes the most recent ideas on sn progenitors and nucleosynthesis , indicating that sne ia originate from long living stars whereas sne of type ii originate from short living stars.

Table A.12: Summaries generated by FactorSum for a sample from arXiv test set (ID = 6213). Reference abstract is presented in Table A.11.

Reference abstract (Tokens: 474)
Congress frequently faces questions about whether and how to commemorate people and events that have influenced the nation's history. Congress often has chosen to do so by establishing national memorials or by conferring a national designation on existing state, local, or private memorials. The National Park Service (NPS) defines national memorials within the National Park System as "primarily commemorative" works that need not be at sites historically associated with their subjects. The Commemorative Works Act (CWA; 40 U.S.C. §§8901-8910) was enacted to govern the establishment process for memorials located in the District of Columbia (Washington, DC) or its environs that are under the jurisdiction of the NPS or the General Services Administration. The CWA includes provisions related to memorial location, design, construction, and perpetual maintenance. Memorials in Washington, DC, include those with the word national in the name and those that are essentially national memorials but do not bear that title. For memorials outside the District of Columbia, no specific law or set of regulations governs their establishment. Congress has established a number of federally administered national memorials throughout the nation, most often as units of the National Park System but also under management of other federal agencies. Various nonfederal entities undertaking commemorative efforts also have petitioned Congress for assistance or statutory recognition, and some individual memorial organizers have titled their works as national memorials without congressional recognition. To clarify options for Congress when considering commemoration of individuals, groups, and events through memorials, this report discusses several types of congressional involvement in memorials outside the District of Columbia. For purposes of the report, these are characterized as high federal involvement (e.g., congressional establishment of a national memorial under federal agency administration); medium federal involvement (e.g., congressional authorization for a memorial to be located on federal property or to receive federal funds); low federal involvement (e.g., statutory recognition without additional federal support); and no federal involvement (e.g., a self-declared national memorial). The report provides examples of memorials of each type and discusses some options for Congress, with regard to both individual memorial designations and consideration of whether to systematize criteria for memorials outside Washington, DC, similar to the CWA's provisions for District of Columbia memorials. Because this report focuses specifically on memorials outside the District of Columbia, please see CRS Report R41658, Commemorative Works in the District of Columbia: Background and Practice, by Jacob R. Straus, for discussion of memorials governed by the CWA in Washington, DC, and its environs.

Table A.13: Sample abstract from the GovReport test set (ID = 681). Model predictions are presented in the next pages.

BART-large summary Tokens: 390; ROUGE-1: 56.11; ROUGE-2: 19.38; ROUGE-L: 54.23
Virtually all societies attempt to remember and memorialize individuals, groups, and events as part of the preservation of shared rhetoric and history. In the United States, there are hundreds, and possibly thousands, of commemorative works to commemorate such events. Decisions about which places to commemorate are made by many different entities, including Congress, federal agencies, state and local governments, and private citizens, among others. Some memorials receive federal involvement, while others do not. For example, for memorials on federal land in the District of Columbia, the Commemorative Works Act (CWA) requires that Congress authorize the creation of a new memorial. No systematic law or set of regulations governs the establishment of memorials outside Washington, DC. However, Congress also has established or recognized numerous memorials nationwide, and some have been designated by the executive branch. Federal agencies may be classified as ”high,” ”medium,” ”low,” or ”none.” Memorials with high federal involvement typically are located on U.S. federal land; receive federal funds for design, construction, and maintenance; and are managed by federal agencies. These include memorials established by Congress as units of the National Park System or under the administration of another agency. Other memorials with no federal involvement are those that do not receive any direct federal involvement (i.e., memorials designated by Congress but not administered by a federal agency). Congress, executive branch officials, and other interested parties may place plaques, memorials, and similar works at federal sites in remembrance of a person, group, or event. or on nonfederal land. The National Park Service (NPS) and the General Services Administration (GSA) maintain some of the nation’s largest memorials. NPS provides assistance to other federal agencies with assistance in managing memorials located on its lands, including NPS-designated ”NPS-affiliated areas.” Other agencies, primarily the Army and Air Force, have similar relationships with NPS affiliated areas. Congress has appropriated funds to both NPS and GSA to help fund memorials created on the lands they manage.

Table A.14: Summary generated by BART for a document from GovReport test set (ID = 681). Reference summary is presented in Table A.13.

FACTORSUM - fixed budget and Lead content guidance
Tokens: 548; ROUGE-1: 63.37; ROUGE-2: 24.77; ROUGE-L: 57.30

Beyond these federally endorsed memorials, a wide variety of other entities have established and maintained memorials throughout the country with no federal connection, including some titled as ”national memorials,”In the United States, there are hundreds, and possibly thousands, of memorials to various individuals, groups, and events. Decisions about which people, groups, or events to memorialize are made by many different entities, including Congress, federal agencies, state and local governments, and private citizens, among others. For example, the CWA governs the establishment of memorials on federal lands in the District of Columbia, with provisions for the creation, design, construction, and maintenance of such works. In other areas, various laws, regulations, and policies may provide for different groups and governments to decide what should be commemorated and how. For certain types of commemorations, Congress has taken a more systematized approach. No systematic law or set of regulations governs the establishment of memorials outside Washington, DC. ,Some of these memorials include multiple facilities such as a visitor center or kiosk in addition to the primary commemor For example, the George Washington Masonic National Memorial in Alexandria, VA, and the National Memorial for Peace and Justice in Montgomery, AL, are privately established and maintained. In some cases, memorials located outside of the District of Columbia have been called ”national” memorials without being so designated by Congress, such as through the establishment of a program to identify nonfederal memorials deserving of a national designation. A distinction is drawn between memorials located within and outside of Washington, DC, because of the exclusive role the CWA gives Congress to authorize new memorials on federal land in the District of Columbia, and the role of federal agencies—primarily the National Park Service (NPS) and the General Services Administration (GSA)—in maintaining District-based memorials once dedicated. This report considers the extent of federal involvement in memorials located outside the District of Columbia. Congress also could potentially consider a program to provide grants to nonfederal entities for constructing and/or maintaining national memorials outside of Washington, DC. While many such works are established without federal involvement, Congress also has established or recognized numerous memorials nationwide, and some have been designated by the executive branch. For purposes of this report, federal involvement in memorials outside the District of Columbia may be classified as ”high,” ”medium,” ”low,” or ”none.” For example, P. L. Other variations of federal-nonf For a discussion of the process for creating a new NPS unit and associated issues, see CRS Report RS20158, National Park System: Establishing New Units. Legislation designating these national memorials often includes explicit language stating that the memorial is not an NPS unit and that federal funds shall not be provided for the memorial. In some instances, Congress authorizes a memorial to be created on federal land and administered by a federal agency.

Table A.15: FACTORSUM summary with BART content guidance for a document from GovReport test set (ID = 681). Reference summary is presented in Table A.13.

FACTORSUM - fixed budget and BART-large content guidance Tokens: 548; ROUGE-1: 63.37; ROUGE-2: 24.77; ROUGE-L: 57.30
In other areas, various laws, regulations, and policies may provide for different groups and governments to decide what should be commemorated and how. Beyond these federally endorsed memorials, a wide variety of other entities have established and maintained memorials throughout the country with no federal connection, including some titled as ”national memorials,”In the United States, there are hundreds, and possibly thousands, of memorials to various individuals, groups, and events. Decisions about which people, groups, or events to memorialize are made by many different entities, including Congress, federal agencies, state and local governments, and private citizens, among others. This report considers the extent of federal involvement in memorials located outside the District of Columbia. For example, the CWA governs the establishment of memorials on federal lands in the District of Columbia, with provisions for the creation, design, construction, and maintenance of such works. No systematic law or set of regulations governs the establishment of memorials outside Washington, DC. In some cases, memorials located outside of the District of Columbia have been called ”national” memorials without being so designated by Congress, such as through the establishment of a program to identify nonfederal memorials deserving of a national designation. For purposes of this report, federal involvement in memorials outside the District of Columbia may be classified as ”high,” ”medium,” ”low,” or ”none.” Legislation designating these national memorials often includes explicit language stating that the memorial is not an NPS unit and that federal funds shall not be provided for the memorial. For a discussion of the process for creating a new NPS unit and associated issues, see CRS Report RS20158, National Park System: Establishing New Units. While many such works are established without federal involvement, Congress also has established or recognized numerous memorials nationwide, and some have been designated by the executive branch. For certain types of commemorations, Congress has taken a more systematized approach. Congress also could potentially consider a program to provide grants to nonfederal entities for constructing and/or maintaining national memorials outside of Washington, DC. A distinction is drawn between memorials located within and outside of Washington, DC, because of the exclusive role the CWA gives Congress to authorize new memorials on federal land in the District of Columbia, and the role of federal agencies—primarily the National Park Service (NPS) and the General Services Administration (GSA)—in maintaining District-based memorials once dedicated. ,Some of these memorials include multiple facilities such as a visitor center or kiosk in addition to the primary commemor For example, the George Washington Masonic National Memorial in Alexandria, VA, and the National Memorial for Peace and Justice in Montgomery, AL, are privately established and maintained. In some instances, Congress authorizes a memorial to be created on federal land and administered by a federal agency. For example, P. L. Other variations of federal-nonf

Table A.16: Summary generated by FACTORSUM with BART content guidance for a document from GovReport test set (ID = 681). Reference summary is presented in Table A.13. Note that this summary uses the same set of summary views as FACTORSUM with Lead content guidance in Table A.15, just changing their presentation order.

A.5 Sample Summaries for LLM Summarizers

In Tables A.23 to A.22, we provide sample summaries for the MuP, arXiv, PubMed, and eLife datasets, generated by models presented in chapter 4.

Reference Abstract: We extend the scalar sector of the neutrinophilic two-Higgs-doublet model, where small masses of Dirac neutrinos are obtained via a small vacuum expectation value of the neutrinophilic -doublet scalar field which has a Yukawa interaction with only right-handed neutrinos. A global symmetry is used for the neutrinophilic nature of the second -doublet scalar field and also for eliminating Majorana mass terms of neutrinos. By virtue of an appropriate assignment of the -charges to new particles, our model has an unbroken symmetry, under which the lightest -odd scalar boson can be a dark matter candidate. In our model, is generated by the one-loop diagram to which -odd particles contribute. We briefly discuss a possible signature of our model at the LHC.
Prompt: *Write a summary of the article above in 6 sentences. Write in first person "we" when applicable.*
LLAMA-2: We present a new model for the origin of neutrino masses, which is based on the neutrino Yukawa coupling and the Higgs mechanism. We extend the Standard Model with a new scalar field, which is a singlet under the SM gauge group and has a non-zero yukawa coupling to neutrinos. This new scalar field generates neutrino masses through the Higgs mechanism, and its vev generates a small neutrino mass matrix. We show that the scenario of spontaneous breaking of the @xmath2 symmetry is not allowed by the measurement of the invisible decay of the @xmath30 boson. We also discuss the possibility of dark matter candidates being involved in the loop diagram, which can generate a small vev for the new scalar field.
GPT-3.5: The article discusses the possibility that neutrinos may not be Majorana fermions but instead Dirac fermions. The neutrinophilic two Higgs doublet model is introduced as a new physics model where neutrinos are regarded as Dirac fermions. This model explains the smallness of neutrino masses by a tiny vacuum expectation value of the neutrinophilic scalar without extremely small yukawa coupling constants for neutrinos. The article extends the neutrinophilic two Higgs doublet model to introduce a dark matter candidate, which may suppress the soft-breaking parameter by extending the model. The new particle discovered at the LHC is likely to be the SM Higgs boson, and may provide information on the origin of particle masses. Collider experiments can test the predictions of this model, which could not be spoiled by certain assumptions.

Table A.17: Sample summaries for Kanemura et al., 2013 (arXiv test set).

Reference Abstract
This study examines the composition and activity of the planktonic community during the polar night in the high Arctic Kongsfjord, Svalbard. Our results are the first published evidence of bioluminescence among zooplankton during the Arctic polar night. The observations were collected by a bathyphotometer detecting bioluminescence, integrated into an autonomous underwater vehicle, to determine the concentration and intensity of bioluminescent flashes as a function of time of day and depth. To further understand community dynamics and composition, plankton nets were used to collect organisms passing through the bathyphotometer along with traditional vertical net tows. Additionally, using a moored bathyphotometer closed to the sampling site, the bioluminescence potential itself was shown not to have a diurnal or circadian rhythm. Rather, our results provide evidence for a diel vertical migration of bioluminescent zooplankton that does not correspond to any externally detectable changes in illumination.
Prompt: *Write a abstract for the article above in 8 sentences. Write in first person "we" when applicable.*
LLAMA-2 Summary
We conducted a study on plankton distribution and bioluminescence in the polar night at Kongsfjorden, Svalbard, during January 2010. We used a remotely operated vehicle (ROV) equipped with bioluminescence, depth, and temperature sensors to collect data on plankton abundance and distribution. Our results showed that bioluminescence potential was detected throughout the water column, with higher intensity at depth during the day and increased surface bioluminescence at night. We found significant differences in acoustic backscatter coefficients between day and night, indicating the presence of zooplankton migration. Our plankton enumerations showed an increase in the majority of the most abundant zooplankton taxa at night, including Pseudocalanus spp., which are known to be present throughout the year in this region. We also found that Metridia lucens, Metridia longa, Oncaea borealis, Thysanoessa inermis, and Thysanoessa longicaudata likely contributed to the increase in high-intensity bioluminescent flashes at 15 and 45 m during the night. Our results suggest that there is a coordinated movement of biomass indicative of vertical migration in this region.

Table A.18: LLAMA-2 summary for Berge et al., 2012 (PubMed test set).

GPT-3.5 Summary
The article discusses the importance of bioluminescence and other adaptations in marine environments. Bioluminescence has evolved independently over 40 times and has various functions including defensive functions such as counter-illumination and burglar alarm and offensive mechanisms such as prey attraction and intra-specific communication. Some adaptations involve entering a dormant state and overwintering at depth, such as seen in the copepods calanus glacialis and c. hyperboreus. The study aimed to characterize plankton abundance and distribution patterns during a time of year rarely studied, using vertical net tows and autonomous underwater vehicle (AUV) surveys. The AUV was fitted with upward and downward facing RD instruments, a CTD, and a bioluminescence bathyphotometer. The data collected off the coast of Svalbard showed bioluminescence was detected throughout the water column both day and night, and higher bioluminescence was present at depth during the day and increased surface bioluminescence at night. Vertical net hauls showed that there was an increase above 60 meters in the majority of the most abundant zooplankton taxa at night, indicating dvm. Lastly, plankton enumerated from the ¿ 20m net collection of the bp exhaust suggests that during the day, the greatest biomass occurred at 45m and was dominated by copepod nauplii, copepod eggs, and the tintinnid acantostomella norvegica.

Table A.19: GPT-3.5 summary for Berge et al., 2012 (PubMed test set). Reference summary is shown in Table A.18.

Reference Lay Summary
Imagine a gymnastics competition in which participants take turns to cartwheel and somersault across the floor. The routines on display comprise sequences of precisely timed movements learned through practice. This is also true for many of the actions we perform every day, such as reaching for a cup of coffee. A region of the brain called the cerebellum helps us learn sequences of movements. But how does it do this? To find out, Khilkevich et al. came up with a new version of an old experiment. Rabbits were first trained to blink their eye in response to a specific external cue. This type of learning, called associative learning, has been shown before in the cerebellum. But Khilkevich et al. wondered whether the cerebellum could also use internal feedback signals from the eyeblink as a cue to learn the next movement? If so, this might explain how the cerebellum can chain movements together in a sequence. As predicted, Khilkevich et al. found that rabbits could learn to blink their eye in response to an initial signal, and then blink again in response to the first blink. Control experiments confirmed that the second eyeblink was coupled to the first, and not to the original cue. Moreover, on many trials the rabbits showed a third and even fourth eyeblink. This is because feedback signals from the first, second or third blink were the same. Thus, the feedback signals from the first blink triggered the second blink, feedback from the second triggered the third, and so forth. Rabbits could also learn to use a blink of the left eye as a cue for a blink of the right eye. Similar patterns of neuronal activity accompanied each blink, suggesting that the same mechanism generated them all. The cerebellum can thus use feedback from one movement as a cue to learn the proper timing of the next movement in a sequence. A key question is whether this mechanism of sequence learning extends beyond movement. The cerebellum has extensive connections to the brain's outer layer, the cortex, including many areas involved in cognition. Future experiments should test whether the cerebellum might help guide sequences of cortical activity during cognitive tasks.

Table A.20: GPT-3.5 summary for Khilkevich et al., 2018 (eLife test set).

GPT-3.5 Summary Prompt: *Write a summary of the article above in 14 sentences.*
The article focuses on how the cerebellum, which is responsible for accurate movement execution and learning, can learn and execute movement sequences. Previous studies have shown that cerebellar pathology leads to severe deficits in the execution and learning of sequences. To test the sufficiency of feedback signals from a movement as a cue for the next component in a sequence, the researchers trained rabbits using electrical stimulation of mossy fibers as a cue. Once subjects acquired a sequence of conditioned eyelid responses, the researchers tested whether the feedback information about the first component is sufficient for the cerebellum to use it as a new "CS" to learn the subsequent CR. The results show that the cerebellum can learn to chain together a sequence of inter-connected movement components by using feedback signals from one component to serve as a cue for the next component in the sequence. In vivo recordings from cerebellar cortex showed that Purkinje cell activity relates as strongly to sequences as it relates to single-component eyelid responses. The experiments suggest that the feedback signals are strong enough to serve as cues for the cerebellum to learn subsequent movements. The authors suggest that cerebellar mechanisms of learning single movement components can be extended to learning of movement sequences, providing a general framework for how the cerebellum can use feedback signals to learn to chain together correctly timed responses to produce a movement sequence. The researchers emphasize that most human movements are comprised of sequences, and movement sequences as well as multi-joint movements are sensitive to cerebellar dysfunction. The article provides insights into how researchers can improve the treatment of movement disorders involving the cerebellum.

Table A.21: GPT-3.5 summary for Khilkevich et al., 2018 (eLife test set).

Prompt: *Write a summary of the article above in 14 sentences.*
LLAMA-2 Summary
Most movements are composed of sequences, and the cerebellum is involved in learning and executing accurate movements. The cerebellum has been linked to learning and execution of movement sequences, and studies have shown that cerebellar dysfunction can lead to deficits in sequence learning or the inability to learn sequences at all. However, most of what we know about cerebellar mechanisms of learning comes from studies using single-component movements. Three possible ways of implementing cerebellar learning of movement sequences have been proposed, and one of these ways involves the use of feedback signals from one movement component to learn the next component. To test the hypothesis that the cerebellum can use feedback signals from a movement to learn the next movement in a sequence, researchers trained rabbits using electrical stimulation of mossy fibers as a cue. The cerebellum can learn to chain together a sequence of interconnected movement components by using feedback signals from one component to serve as a cue for the next component in the sequence. The design of most experiments does not permit distinguishing between different possibilities of sequence learning, but the use of eyelid conditioning to test the sufficiency of feedback signals (FS) for sequence learning provided direct evidence that the cerebellum can learn to use FS to chain together movement components. In the experiment, the first movement was a conditioned eyelid response (CR) elicited by a sensory stimulus, and the second movement was a sequence of movements produced by the same muscle or different muscle groups used for the different components of the sequence. The results showed robust learning of later components, including sequences of movements produced by the same muscle and sequences with different muscle groups used for the different components. The in vivo recordings from cerebellar cortex showed that Purkinje cell activity related as strongly to the latter components of sequences as it has been shown to relate to single-component eyelid responses. These findings provide a general framework for how the cerebellum can use feedback signals to learn to chain together appropriately timed responses to produce a movement sequence. The results demonstrate the importance of cerebellar learning mechanisms in mediating the acquisition of movement sequences and suggest that the cerebellum can use feedback signals from a movement to learn the next movement in a sequence.

Table A.22: LLAMA-2 summary for Khilkevich et al., 2018 (eLife test set).

Human 1 (Reference): This paper presents a fast method for solving certain classes of optimal control problems, by implementing a value iteration (VI) algorithm that takes advantage of convex duality. The basic idea is to take advantage of the fact that for two functions f_1, f_2, we have $(f_1 \Box f_2)^* = f_1^* + f_2^*$, where f^* denotes the Legendre-Fenchel transform of f, and $(f_1 \Box f_2)(x) = \inf\{f_1(x_1) + f_2(x_2) : x_1 + x_2 = x\}$ denotes the infimal convolution of f_1, f_2. Thus, the slow inf operation can be replaced with a fast addition operation. The paper provides analyses of the convergence, time complexity, and error of the algorithm. When compared to a "naive" VI algorithm in which each iteration takes $O(XU)$ (where X, U are the grid sizes of the state and input spaces, respectively), the algorithm presented in this paper can achieve $O(X+U)$. Finally, the paper provides a numerical example implemented in MATLAB.
Human 2: The paper develops an approximate value iteration method for an infinite-horizon, discounted-cost Markov Decision Processes (MDPs) that satisfy a given set of regulatory assumptions. The idea is to work on a dual space that replaces the DP value function by a reformulation written in terms of biconjugate operations. The authors present an algorithm that applies the conjugate operators iteratively, and present convergence and other structural results associated with the resulting errors. Numerical results compare the proposed approach against a traditional value iteration and a variant that generated dynamic discretization grids.
Prompt: *Write a summary of the article above in 4 sentences.*
LLAMA-2: The article discusses the problem of optimal control of continuous-state systems with discrete-time dynamics and stochastic cost. The authors propose a new algorithm, called Conjugate Value Iteration (ConjVI), which combines the power of the DP operator with the efficiency of discrete-time methods to solve this problem. ConjVI is based on a modified version of the Discrete Dual Domain (d-CDP) operator, which allows for a more efficient computation of the value function. The authors provide theoretical guarantees on the convergence of ConjVI and analyze its error, showing that it can achieve a linear time complexity of O(X+U) in each iteration, where X and U are the cardinalities of the discrete state and input spaces, respectively.
GPT-3.5: Value iteration (VI) is an algorithm used for problems in reinforcement learning (RL) and optimal control formulated as Markov decision processes (MDPs). However, VI suffers from a high computational cost for large-scale finite state spaces. A solution to this problem is to incorporate function approximation techniques and compute the output of the DP operator for a finite sample of the underlying continuous state space. The article proposes the conjugate VI (ConjVI) algorithm based on a modified version of the d-CDP operator that extends existing results in three directions for solving optimal control problems of discrete-time systems with a continuous state-input space.

Table A.23: Sample summaries for Kolarijani et al., 2021 (MuP validation set).

Bibliography

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Agrawal, M., Hegselmann, S., Lang, H., Kim, Y., and Sontag, D. (2022). Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022.

Ahuja, O., Xu, J., Gupta, A., Horecka, K., and Durrett, G. (2021). Aspect-news: Aspect-oriented summarization of news documents. *ArXiv preprint*, abs/2110.08296.

Almazrouei, E., Alobeidli, H., Alshamsi, A., Cappelli, A., Cojocaru, R., Debbah, M., Goffinet, É., Hesslow, D., Launay, J., Malartic, Q., et al. (2023). The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

Andreas, J. (2022). Language models as agent models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5769–5779, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Artstein, R. and Poesio, M. (2008a). Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4):555–596.

Artstein, R. and Poesio, M. (2008b). Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

August, T., Lo, K., Smith, N. A., and Reinecke, K. (2024). Know your audience: The benefits and pitfalls of generating plain language summaries beyond the" general" audience. *arXiv preprint arXiv:2403.04979*.

Aumiller, D., Chouhan, A., and Gertz, M. (2022). EUR-lex-sum: A multi- and cross-lingual dataset for long-form summarization in the legal domain. In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7626–7639, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., and Schneider, N. (2013). Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 178–186.

Beltagy, I., Lo, K., and Cohan, A. (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.

Beltagy, I., Peters, M. E., and Cohan, A. (2020). Longformer: The long-document transformer. *ArXiv preprint*, abs/2004.05150.

Bennett, M. (2013). The financial industry business ontology: Best practice for big data. *Journal of Banking Regulation*, 14(3-4):255–268.

Bhattacharya, P., Poddar, S., Rudra, K., Ghosh, K., and Ghosh, S. (2021). Incorporating domain knowledge for extractive summarization of legal case documents. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, ICAIL '21, page 22–31, New York, NY, USA. Association for Computing Machinery.

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Bradley, R. A. and Terry, M. E. (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1-7):107–117.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Cao, S. and Wang, L. (2022). HIBRIDS: Attention with Hierarchical Biases for Structure-aware Long Document Summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 786–807, Dublin, Ireland. Association for Computational Linguistics.

Carbonell, J. and Goldstein, J. (1998). The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336.

Cardenas, R., Gallé, M., and Cohen, S. B. (2024). On the trade-off between redundancy and cohesion in extractive summarization. *Journal of Artificial Intelligence Research*.

Chan, H. P., Wang, L., and King, I. (2021). Controllable summarization with constrained Markov decision process. *Transactions of the Association for Computational Linguistics*, 9:1213–1232.

Chang, Y., Lo, K., Goyal, T., and Iyyer, M. (2024). Booookscore: A systematic exploration of book-length summarization in the era of LLMs. In *The Twelfth International Conference on Learning Representations*.

Cheng, M., Green, W., Conradie, P., Konishi, N., and Romi, A. (2014). The international integrated reporting framework: key issues and future research opportunities. *Journal of International Financial Management & Accounting*, 25(1):90–119.

Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhang, H., Zhu, B., Jordan, M., Gonzalez, J. E., et al. (2024). Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., et al. (2022). Scaling instruction-finetuned language models. *ArXiv preprint*, abs/2210.11416.

Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., and Smith, N. A. (2021). All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296, Online. Association for Computational Linguistics.

Cohan, A., Beltagy, I., King, D., Dalvi, B., and Weld, D. (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.

Cohan, A., Dernoncourt, F., Kim, D. S., Bui, T., Kim, S., Chang, W., and Goharian, N. (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.

Cohan, A., Feigenblat, G., Ghosal, T., and Shmueli-Scheuer, M. (2022). Overview of the first shared task on multi perspective scientific document summarization (MuP). In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 263–267, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Dernoncourt, F. and Lee, J. Y. (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.

Dong, Y., Mircea, A., and Cheung, J. C. K. (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.

Dou, Z.-Y., Liu, P., Hayashi, H., Jiang, Z., and Neubig, G. (2021). GSum: A general framework for guided neural abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4830–4842, Online. Association for Computational Linguistics.

Dreyer, M., Liu, M., Nan, F., Atluri, S., and Ravi, S. (2023). Evaluating the tradeoff between abstractiveness and factuality in abstractive summarization. In Vlachos, A. and Augenstein, I., editors, *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2089–2105, Dubrovnik, Croatia. Association for Computational Linguistics.

Duan, X., Zhang, Y., Yuan, L., Zhou, X., Liu, X., Wang, T., Wang, R., Zhang, Q., Sun, C., and Wu, F. (2019). Legal summarization for multi-role debate dialogue via controversy focus mining and multi-task learning. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 1361–1370, New York, NY, USA. Association for Computing Machinery.

Durmus, E., He, H., and Diab, M. (2020). FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Elaraby, M. and Litman, D. (2022). ArgLegalSumm: Improving abstractive summarization of legal documents with argument mining. In Calzolari, N., Huang, C.-R., Kim, H., Pustejovsky, J., Wanner, L., Choi, K.-S., Ryu, P.-M., Chen, H.-H., Donatelli, L., Ji, H., Kurohashi, S., Paggio, P., Xue, N., Kim, S., Hahm, Y., He, Z., Lee, T. K., Santus, E., Bond, F., and Na, S.-H., editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6187–6194, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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

Fabbri, A. R., Kryściński, W., McCann, B., Xiong, C., Socher, R., and Radev, D. (2021). Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Fan, A., Grangier, D., and Auli, M. (2018). Controllable abstractive summarization. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 45–54, Melbourne, Australia. Association for Computational Linguistics.

Fonseca, M. and Cohen, S. B. (2024). Can large language model summarizers adapt to diverse scientific communication goals? *arXiv preprint arXiv:2401.10415*.

Fonseca, M., Ziser, Y., and Cohen, S. B. (2022). Factorizing content and budget decisions in abstractive summarization of long documents. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6341–6364, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ghosh, S., Dutta, M., and Das, T. (2022). Indian legal text summarization: A text normalization-based approach. In *2022 IEEE 19th India Council International Conference (INDICON)*, pages 1–4, Kochi, India. Institute of Electrical and Electronics Engineers.

Gidiotis, A. and Tsoumakas, G. (2020). A divide-and-conquer approach to the summarization of long documents. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:3029–3040.

Giulianelli, M., Baan, J., Aziz, W., Fernández, R., and Plank, B. (2023). What comes next? evaluating uncertainty in neural text generators against human production variability. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14349–14371, Singapore. Association for Computational Linguistics.

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 Demner-fushman, D., Ananiadou, S., and Cohen, K., editors, *The 22nd Workshop on Biomedical Natural Language Processing and*

BioNLP Shared Tasks, pages 468–477, Toronto, Canada. Association for Computational Linguistics.

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*, pages 10589–10604, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

GOV.UK (2024). National Minimum Wage and National Living Wage rates. https://www.gov.uk/national-minimum-wage-rate. [Accessed 11-02-2024].

Goyal, T., Li, J. J., and Durrett, G. (2022). News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

Grenander, M., Dong, Y., Cheung, J. C. K., and Louis, A. (2019). Countering the effects of lead bias in news summarization via multi-stage training and auxiliary losses. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *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 6019–6024, Hong Kong, China. Association for Computational Linguistics.

Grusky, M., Naaman, M., and Artzi, Y. (2018). Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In Walker, M., Ji, H., and Stent, A., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Gu, N., Ash, E., and Hahnloser, R. H. (2021). Memsum: Extractive summarization of long documents using multi-step episodic markov decision processes. *arXiv preprint arXiv:2107.08929*.

Guo, M., Ainslie, J., Uthus, D., Ontanon, S., Ni, J., Sung, Y.-H., and Yang, Y. (2022). LongT5: Efficient text-to-text transformer for long sequences. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 724–736, Seattle, United States. Association for Computational Linguistics.

He, B., Wang, J., Qiu, J., Bui, T., Shrivastava, A., and Wang, Z. (2023). Align and attend: Multimodal summarization with dual contrastive losses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14867–14878.

He, J., Kryscinski, W., McCann, B., Rajani, N., and Xiong, C. (2022). CTRLsum: Towards generic controllable text summarization. In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5879–5915, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. (2015). Teaching machines to read and comprehend. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. *ArXiv preprint*, abs/1503.02531.

Ho, J. and Salimans, T. (2022). Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. (2020). The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2022). Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Huang, L., Cao, S., Parulian, N., Ji, H., and Wang, L. (2021). Efficient attentions for long document summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1419–1436, Online. Association for Computational Linguistics.

Iso, H., Wang, X., Suhara, Y., Angelidis, S., and Tan, W.-C. (2021). Convex Aggregation for Opinion Summarization. In *Findings of the Association for Computational*

Linguistics: EMNLP 2021, pages 3885–3903, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. (2023). Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Jones, K. S. (1998). Automatic summarising: factors and directions. *arXiv preprint cmp-lg/9805011*.

Kalchbrenner, N., Grefenstette, E., and Blunsom, P. (2014). A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188*.

Kikuchi, Y., Neubig, G., Sasano, R., Takamura, H., and Okumura, M. (2016). Controlling output length in neural encoder-decoders. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1328–1338, Austin, Texas. Association for Computational Linguistics.

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., and Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. *Institute for Simulation and Training, University of Central Florida*.

Kintsch, W. and Van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological review*, 85(5):363.

Klaus, S., Van Hecke, R., Djafari Naini, K., Altingovde, I. S., Bernabé-Moreno, J., and Herrera-Viedma, E. (2022). Summarizing legal regulatory documents using transformers. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2426–2430, New York, NY, USA. Association for Computing Machinery.

Koh, J. Y., Fried, D., and Salakhutdinov, R. R. (2023). Generating images with multimodal language models. In Oh, A., Neumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 21487–21506. Curran Associates, Inc.

Kornilova, A. and Eidelman, V. (2019). Billsum: A corpus for automatic summarization of us legislation. *arXiv preprint arXiv:1910.00523*.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Kryściński, W., Keskar, N. S., McCann, B., Xiong, C., and Socher, R. (2019). Neural text summarization: A critical evaluation. *arXiv preprint arXiv:1908.08960*.

Kumar, S., Paria, B., and Tsvetkov, Y. (2022). Gradient-based constrained sampling from language models. In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2251–2277, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., and Huang, F. (2006). A tutorial on energy-based learning. *Predicting structured data*, 1(0).

Lester, B., Al-Rfou, R., and Constant, N. (2021). The power of scale for parameter-efficient prompt tuning. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Levenshtein, V. I. et al. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (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.

Liakata, M., Saha, S., Dobnik, S., Batchelor, C., and Rebholz-Schuhmann, D. (2012). Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics*, 28(7):991–1000.

Liakata, M. and Soldatova, L. (2008). Guidelines for the annotation of general scientific concepts. *Aberystwyth University, JISC Project Report http://ie-repository. jisc. ac. uk/88*.

Liakata, M., Teufel, S., Siddharthan, A., and Batchelor, C. (2010). Corpora for the conceptualisation and zoning of scientific papers. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

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

Lin, C.-Y. and Hovy, E. (2000). The automated acquisition of topic signatures for text summarization. In *COLING 2000 volume 1: The 18th international conference on computational linguistics*.

Liu, F., Flanigan, J., Thomson, S., Sadeh, N., and Smith, N. A. (2015). Toward abstractive summarization using semantic representations. In Mihalcea, R., Chai, J., and Sarkar, A., editors, *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1077–1086, Denver, Colorado. Association for Computational Linguistics.

Liu, Y., Jia, Q., and Zhu, K. (2022a). Length Control in Abstractive Summarization by Pretraining Information Selection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6885–6895, Dublin, Ireland. Association for Computational Linguistics.

Liu, Y. and Lapata, M. (2019). Text summarization with pretrained encoders. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Liu, Y. and Liu, P. (2021). SimCLS: A simple framework for contrastive learning of abstractive summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1065–1072, Online. Association for Computational Linguistics.

Liu, Y., Liu, P., Radev, D., and Neubig, G. (2022b). Brio: Bringing order to abstractive summarization. *arXiv preprint arXiv:2203.16804*.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165.

Maddela, M., Kulkarni, M., and Preotiuc-Pietro, D. (2022). EntSUM: A data set for entity-centric extractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3355–3366, Dublin, Ireland. Association for Computational Linguistics.

Makino, T., Iwakura, T., Takamura, H., and Okumura, M. (2019). Global optimization under length constraint for neural text summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1039–1048, Florence, Italy. Association for Computational Linguistics.

Malik, M., Zhao, Z., Fonseca, M., Rao, S., and Cohen, S. B. (2024). Civilsum: A dataset for abstractive summarization of indian court decisions. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2241–2250.

Mann, W. C. and Thompson, S. A. (1987). *Rhetorical structure theory: A theory of text organization*. University of Southern California, Information Sciences Institute Los Angeles.

Mao, Z., Wu, C. H., Ni, A., Zhang, Y., Zhang, R., Yu, T., Deb, B., Zhu, C., Awadallah, A., and Radev, D. (2022). DYLE: Dynamic latent extraction for abstractive long-input summarization. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1687–1698, Dublin, Ireland. Association for Computational Linguistics.

Marcu, D. (1997). From discourse structures to text summaries. In *Intelligent Scalable Text Summarization*.

Marcu, D. (1998). To build text summaries of high quality, nuclearity is not sufficient. In *Working Notes of the AAAI-98 Spring Symposium on Intelligent Text Summarization*, pages 1–8.

Marcus, M., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330.

McKenna, N., Li, T., Cheng, L., Hosseini, M., Johnson, M., and Steedman, M. (2023). Sources of hallucination by large language models on inference tasks. In Bouamor, H., Pino, J., and Bali, K., editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2758–2774, Singapore. Association for Computational Linguistics.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. (2022). Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.

Mihalcea, R. and Tarau, P. (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.

Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., and Zettlemoyer, L. (2022). Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.

Nenkova, A. and McKeown, K. (2012). A survey of text summarization techniques. *Mining text data*, pages 43–76.

Onoe, Y., Zhang, M. J., Padmanabhan, S., Durrett, G., and Choi, E. (2023). Can lms learn new entities from descriptions? challenges in propagating injected knowledge. *arXiv preprint arXiv:2305.01651*.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Oved, N. and Levy, R. (2021). PASS: Perturb-and-select summarizer for product reviews. 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 351–365, Online. Association for Computational Linguistics.

Pan, L., Albalak, A., Wang, X., and Wang, W. Y. (2023). Logic-lm: Empowering large language models with symbolic solvers for faithful logical reasoning. *arXiv preprint arXiv:2305.12295*.

Pang, B., Nijkamp, E., Kryscinski, W., Savarese, S., Zhou, Y., and Xiong, C. (2023). Long document summarization with top-down and bottom-up inference. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1267–1284, Dubrovnik, Croatia. Association for Computational Linguistics.

Patel, R. and Pavlick, E. (2021). Mapping language models to grounded conceptual spaces. In *International Conference on Learning Representations*.

Penedo, G., Malartic, Q., Hesslow, D., Cojocaru, R., Cappelli, A., Alobeidli, H., Pannier, B., Almazrouei, E., and Launay, J. (2023). The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Peyrard, M. (2018). A simple theoretical model of importance for summarization. *arXiv preprint arXiv:1801.08991*.

Peyrard, M. (2019). *Principled Approaches to Automatic Text Summarization*. PhD thesis, Technische Universität Darmstadt, Darmstadt.

Phan, P., Tran, T., and Trieu, H.-L. (2023). VBD-NLP at BioLaySumm task 1: Explicit and implicit key information selection for lay summarization on biomedical long documents. In Demner-fushman, D., Ananiadou, S., and Cohen, K., editors, *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 574–578, Toronto, Canada. Association for Computational Linguistics.

Phang, J., Zhao, Y., and Liu, P. J. (2022). Investigating efficiently extending transformers for long input summarization. *ArXiv preprint*, abs/2208.04347.

Piantadosi, S. T. and Hill, F. (2022). Meaning without reference in large language models. *arXiv preprint arXiv:2208.02957*.

Pu, D. and Demberg, V. (2023). ChatGPT vs human-authored text: Insights into controllable text summarization and sentence style transfer. In Padmakumar, V., Vallejo, G., and Fu, Y., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 1–18, Toronto, Canada. Association for Computational Linguistics.

Pu, X., Gao, M., and Wan, X. (2023). Summarization is (almost) dead. *arXiv preprint arXiv:2309.09558*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf. [Accessed 11-02-2024].

Rajbhandari, S., Rasley, J., Ruwase, O., and He, Y. (2020). Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE.

Rush, A. M., Chopra, S., and Weston, J. (2015). A neural attention model for abstractive sentence summarization. In Màrquez, L., Callison-Burch, C., and Su, J., editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

Sanchez, G., Fan, H., Spangher, A., Levi, E., Ammanamanchi, P. S., and Biderman, S. (2023). Stay on topic with classifier-free guidance. *ArXiv preprint*, abs/2306.17806.

Sanh, V., Webson, A., Raffel, C., Bach, S. H., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Raja, A., Dey, M., Bari, M. S., Xu, C., Thakker, U., Sharma, S. S., Szczechla, E., Kim, T., Chhablani, G., Nayak, N. V., Datta, D., Chang, J., Jiang, M. T., Wang, H., Manica, M., Shen, S., Yong, Z. X., Pandey, H., Bawden, R., Wang, T., Neeraj, T., Rozen, J., Sharma, A., Santilli, A., Févry, T., Fries, J. A., Teehan, R., Scao, T. L., Biderman, S., Gao, L., Wolf, T., and Rush, A. M. (2022). Multitask prompted training enables zero-shot task generalization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Saparov, A. and He, H. (2022). Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. *arXiv preprint arXiv:2210.01240*.

SEC (2014). SEC.gov — Filings & Forms. https://www.sec.gov/edgar. [Accessed 11-02-2024].

See, A., Liu, P. J., and Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. In Barzilay, R. and Kan, M.-Y., editors, *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.

Shukla, A., Bhattacharya, P., Poddar, S., Mukherjee, R., Ghosh, K., Goyal, P., and Ghosh, S. (2022). Legal case document summarization: Extractive and abstractive methods and their evaluation. *arXiv preprint arXiv:2210.07544*.

Silva, N. F. F. d., Silva, M. C. R., Pereira, F. S. F., Tarrega, J. a. P. M., Beinotti, J. a. V. P., Fonseca, M., Andrade, F. E. d., and de Carvalho, A. C. P. d. L. F. (2021). Evaluating topic models in portuguese political comments about bills from brazil's chamber of deputies. In *Intelligent Systems: 10th Brazilian Conference, BRACIS 2021, Virtual Event, November 29 – December 3, 2021, Proceedings, Part II*, page 104–120, Berlin, Heidelberg. Springer-Verlag.

Soldatova, L. N. and King, R. D. (2006). An ontology of scientific experiments. *Journal of the royal society interface*, 3(11):795–803.

S&P Global (2024). S&P 500®. https://www.spglobal.com/spdji/en/indices/equity/sp-500. [Accessed 11-02-2024].

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. F. (2020). Learning to summarize with human feedback. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021. Curran Associates, Inc.

Sun, S., Shapira, O., Dagan, I., and Nenkova, A. (2019). How to compare summarizers without target length? pitfalls, solutions and re-examination of the neural summarization literature. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 21–29, Minneapolis, Minnesota. Association for Computational Linguistics.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*

27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada, pages 3104–3112.

Sutton, R. (2019). The bitter lesson. http://www.incompleteideas.net/IncIdeas/BitterLesson.html. [Accessed 11-02-2024].

Takase, S. and Okazaki, N. (2019). Positional encoding to control output sequence length. 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 3999–4004, Minneapolis, Minnesota. Association for Computational Linguistics.

Ter Hoeve, M., Kiseleva, J., and Rijke, M. (2022). What makes a good and useful summary? Incorporating users in automatic summarization research. In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V., editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 46–75, Seattle, United States. Association for Computational Linguistics.

Teufel, S. and Moens, M. (2002). Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational linguistics*, 28(4):409–445.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. *ArXiv preprint*, abs/2302.13971.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Vijayakumar, A. K., Cogswell, M., Selvaraju, R. R., Sun, Q., Lee, S., Crandall, D., and Batra, D. (2016). Diverse beam search: Decoding diverse solutions from neural sequence models. *ArXiv preprint*, abs/1610.02424.

Wang, D., Zhu, S., Li, T., and Gong, Y. (2009). Multi-document summarization using sentence-based topic models. In Su, K.-Y., Su, J., Wiebe, J., and Li, H., editors, *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 297–300, Suntec, Singapore. Association for Computational Linguistics.

Wang, S., Liu, Y., Xu, Y., Zhu, C., and Zeng, M. (2021). Want to reduce labeling cost? gpt-3 can help. *arXiv preprint arXiv:2108.13487*.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al. (2022). Emergent abilities of large language models. *ArXiv preprint*, abs/2206.07682.

Wei, J., Wei, J., Tay, Y., Tran, D., Webson, A., Lu, Y., Chen, X., Liu, H., Huang, D., Zhou, D., et al. (2023). Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-art natural language processing. In Liu, Q. and Schlangen, D., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Wu, Z., Qiu, L., Ross, A., Akyürek, E., Chen, B., Wang, B., Kim, N., Andreas, J., and Kim, Y. (2023). Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. *arXiv preprint arXiv:2307.02477*.

Xie, S. M., Raghunathan, A., Liang, P., and Ma, T. (2021). An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*.

Xiong, W., Gupta, A., Toshniwal, S., Mehdad, Y., and Yih, W.-t. (2022). Adapting pre-trained text-to-text models for long text sequences. *ArXiv preprint*, abs/2209.10052.

Yeh, J.-Y., Ke, H.-R., Yang, W.-P., and Meng, I.-H. (2005). Text summarization using a trainable summarizer and latent semantic analysis. *Information processing & management*, 41(1):75–95.

Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontañón, S., Pham, P., Ravula, A., Wang, Q., Yang, L., and Ahmed, A. (2020). Big bird: Transformers for longer sequences. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Zhang, J., Zhao, Y., Saleh, M., and Liu, P. J. (2020). PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

Zhang, R., Li, W., Liu, N., and Gao, D. (2016). Coherent narrative summarization with a cognitive model. *Computer Speech & Language*, 35:134–160.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zhang, T., Ladhak, F., Durmus, E., Liang, P., McKeown, K., and Hashimoto, T. B. (2024). Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Zhang, Y., Ni, A., Mao, Z., Wu, C. H., Zhu, C., Deb, B., Awadallah, A. H., Radev, D., and Zhang, R. (2021). Summˆ n: A multi-stage summarization framework for long input dialogues and documents. *arXiv preprint arXiv:2110.10150*.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Zheng, H. and Lapata, M. (2019). Sentence centrality revisited for unsupervised summarization. *arXiv preprint arXiv:1906.03508*.

Zhu, C., Rawat, A. S., Zaheer, M., Bhojanapalli, S., Li, D., Yu, F., and Kumar, S. (2020). Modifying memories in transformer models. *arXiv preprint arXiv:2012.00363*.

Zhu, C., Yang, Z., Gmyr, R., Zeng, M., and Huang, X. (2021). Leveraging lead bias for zero-shot abstractive news summarization. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR ’21, page 1462–1471, New York, NY, USA. Association for Computing Machinery.

Zhu, J., Li, H., Liu, T., Zhou, Y., Zhang, J., and Zong, C. (2018). MSMO: Multimodal summarization with multimodal output. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J., editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4154–4164, Brussels, Belgium. Association for Computational Linguistics.

Zhu, Y., Zhang, P., Haq, E.-U., Hui, P., and Tyson, G. (2023). Can chatgpt reproduce human-generated labels? a study of social computing tasks. *arXiv preprint arXiv:2304.10145*.