Multi-Document Summarization: From Ideational to Opinionated Sources

Miao Li

ORCID: 0000-0002-1669-7063

School of Computing and Information Systems

The University of Melbourne

Submitted in total fulfillment of the requirements of the degree of

Doctor of Philosophy

August 2025

In memory of my grandfather.

Abstract

The thesis alleviates research gaps in multi-document summarization (MDS) by investigating how to integrate information from both ideational and opinionated sources.

For ideational sources such as news articles, we construct heterogeneous graphs to capture relationships among source documents. To incorporate the heterogeneous graph representations into pre-trained language models (PLMs), we propose a multi-task training framework with objectives for heterogeneous graph compression and text summarization. Our model achieves the state-of-the-art performances on widely-used datasets.

For opinionated sources such as product reviews, we introduce a challenging new benchmark derived from scientific reviews, characterized by explicit and complex inter-document relationships. To model these relationships, we design a baseline model that incorporates the conversational structure by manipulating Transformer attention mechanisms. Experimental results show that this simple baseline significantly outperforms other strong models. Manual analysis on the generated meta-reviews further reveals that most models cannot recognize and resolve conflicts in reviews, highlighting a promising direction for future research. In addition, we propose an aspect-aware opinion consolidation framework, inspired by strategies employed by human meta-reviewers, to enhance the quality and transparency of summarization models. Experiments demonstrate that prompting with this framework produces better meta-reviews than other competitive prompting approaches.

We further investigate decomposing the summarization process across domains by leveraging underlying aspects in opinion summarization (e.g., price and cleanliness for hotels), and examine how this decomposition framework can improve both summary quality and evaluation. Experimental results show that this approach produces higher-quality meta-reviews in most domains, with greater aspect coverage and more faithful generations than strong baselines, while remaining competitive with our prior approach in the scientific domain. Moreover, the framework generates intermediate steps that assist human meta-reviewers in producing better reviews more efficiently. Overall, this work demonstrates that incorporating reasoning traces into prompting enhances the effectiveness of large language models (LLMs) in MDS.

Declaration

I hereby declare that:

- Unless explicitly indicated otherwise, the content of this dissertation is original and has not been submitted, either in whole or in part, for consideration towards any other degree or qualification at this or any other university;
- This dissertation represents my own work and does not include material resulting from collaborative efforts, except where explicitly stated within the text and the Acknowledgments section;
- This dissertation contains fewer than 65,000 words (including appendices, bibliography, footnotes, tables and equations) and includes fewer than 150 figures.

Miao Li
ORCID: 0000-0002-1669-7063
August 2025

Preface

The thesis is based on the work conducted within my PhD candidature. I declare that I am the main contributor while in collaboration with my supervisors, Eduard Hovy and Jey Han Lau from The University of Melbourne, and Mirella Lapata from The University of Edinburgh. The work in the thesis has been published in the following papers.

- **Miao Li**, Jey Han Lau, Eduard Hovy, and Mirella Lapata. 2025. Decomposed Opinion Summarization with Verified Aspect-Aware Modules. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24805–24841, Vienna. Association for Computational Linguistics.

- **Miao Li**, Jey Han Lau, and Eduard Hovy. 2024. A Sentiment Consolidation Framework for Meta-Review Generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10158–10177, Bangkok, Thailand. Association for Computational Linguistics.

- **Miao Li**, Eduard Hovy, and Jey Han Lau. 2023. Summarizing Multiple Documents with Conversational Structure for Meta-Review Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7089–7112, Singapore. Association for Computational Linguistics.

- **Miao Li**, Jianzhong Qi, and Jey Han Lau. Compressed heterogeneous graph for abstractive multi-document summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 11, pp. 13085-13093. 2023.

Besides the above publications that comprise the main content of the thesis, there are other publications that I have during the PhD, which are in different research topics from the thesis and so are not included.

- **Miao Li**, Ming-Bin Chen, Bo Tang, Shengbin Hou, Pengyu Wang, Haiying Deng, Zhiyu Li, Feiyu Xiong, Keming Mao, Cheng Peng, and Yi Luo. 2024. NewsBench: A Systematic Evaluation Framework for Assessing Editorial Capabilities of Large Language Models in Chinese Journalism. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9993–10014, Bangkok, Thailand. Association for Computational Linguistics.

- Zhuohan Xie, **Miao Li**, Trevor Cohn, and Jey Han Lau. 2023. DeltaScore: Fine-Grained Story Evaluation with Perturbations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5317–5331, Singapore. Association for Computational Linguistics.

- Rui Zhang, Bayu Distiawan Trisedya, **Miao Li**, Yong Jiang, and Jianzhong Qi. A benchmark and comprehensive survey on knowledge graph entity alignment via representation learning. *The VLDB Journal* 31, no. 5 (2022): 1143-1168.

- Mengxue Zhao, Yang Yang, **Miao Li**, Jingang Wang, Wei Wu, Pengjie Ren, Maarten de Rijke, and Zhaochun Ren. Personalized Abstractive Opinion Tagging. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1066-1076. 2022.

Acknowledgements

The PhD journey has been truly wonderful, despite its ups and downs. I did my main study in School of Computer Science and Information Systems, The University of Melbourne and spent around one year visiting in School of Informatics, The University of Edinburgh. I am deeply grateful to everyone who supported me and helped make these years unforgettable.

First and foremost, I would like to express my deepest gratitude to my supervisors in Melbourne, Prof. Eduard Hovy and Dr. Jey Han Lau. It has been a privilege to work with them. From them, I have learned a lot, not only to conduct right and meaningful research but also how to be a kind and thoughtful person. I still remember the meetings when we were discussing promising ideas and interesting results. In particular, getting to know Ed is the most precious part of my experience in Melbourne. It is difficult to put into words how much his support and guidance have meant to me. He instilled in me the importance of aiming high of tackling big problems and doing impactful work, and encouraged me to be 'arrogant' of my own research and humble of learning from others. Not living up to his expectation remains my greatest regret during my PhD, but his encouragement will stay with me forever.

Then I would like to extend my heartfelt thanks to Prof. Mirella Lapata, who supervised me during my time as a visiting PhD student in Edinburgh. It was a true privilege to be invited to visit her exceptional group after we met at ACL 2023 in Toronto. I have been deeply inspired by the vibrant intellectual atmosphere in the group and by her unwavering passion and leading insight for research. Despite her demanding schedule, she was always accessible and generous with her time, offering thoughtful feedback and guidance whenever I needed it. I am especially grateful for the care and effort she put into my work—even editing table colours in my paper draft and sending me interesting papers late at night. None of it would have been possible without her support.

I would also like to thank Prof. Tim Baldwin, Prof. Trevor Cohn, and Dr. Lea Frermann from the UniMelb NLP group and Prof. Ivan Titov from the EdinburghNLP group, who gave me great advice in doing research.

During my years in Melbourne, I have been blessed with the friendship of many amazing people. I would like to particularly thank all friends in the UniMelb NLP group, including Aili Shen, Aso Mahmudi, Biaoyan Fang, Chunhua Liu, Dalin Wang, Fajri Koto, Fan Jiang, Fei Liu, Gisela Vallejo, Haonan Li, Hung Thinh Truong, Jinghui Liu, Jinrui Yang, Jiyu Chen, Jun Wang, Kemal Kurniawan, Ming-Bin Chen, Qiongkai Xu, Rena Gao, Rongxin Zhu, Rui Xing, Shima Khanezar, Takashi Wada, Viktoria Schram, Xudong Han, Yanbei Jiang, Yilin Geng, Yuan Li, Yulia Otmakhova, Yuxia Wang, Zenan Zhai, Zheng Wei Lim, and Zhuohan Xie (in an alphabet order), and other friends from CIS or Melbourne Connect, including Emma Zhang, Gracie Pu, Haopeng Li, Haoyu Gao, Kabir Manandhar Shrestha, Markus Hiller, Mo Zhang, Pok Him Siu, Shiquan Yang, Shu Liu, Suleman Rasheed, Tina Zhang, Yanchuan Chang, Ying Ma (in an alphabet order). If I missed anyone here, it was totally my fault. Knowing you in person is also a privilege, and I can not wait to see you all in future.

During my time in Edinburgh, I have been grateful that we have great discussion on research and amazing lunches together with students or members from Mirella's group. They are Akash Gupta, Ashutosh Adhikari, Alex Gurung, Agostina Calabrese, Iñigo Alonso, Irina Saparina, Julen Etxaniz, Katya Taktasheva, Laura Perez-Beltrachini, Louis Mahon, Litu Ou, Parag Jain, Sophie Fischer, Tom Hosking, Tom Sherborne, Ulrich Germann, and Yao Fu (in an alphabet order). I would also thank my friends from the 4.08 hot desk group for visiting students in Informatics Forum, especially Alberto Carlo Maria Mancino, Alessio Devoto, Chiara Manganini, Chuang Liu, Dongqi Pu, Hosein Mohebbi, Hongru Wang, Kaci Ibazizene, Marcell Fekete, Nathan Godey, Victoria Brami, Wen Lai, Wenhao Zhu, Xuehui Yu, Xuran Ge, Zhaowei Wang (in an alphabet order). We are from all corners of the globe (i.e., Australia, China, France, Germany, Hungary, Italy, Netherlands, Spain), each bringing our unique cultures, perspectives, and experiences. Together, we created a vibrant, diverse, and supportive community that I will always cherish.

Last but not least, I would like to thank my parents, my sister, and all my family for their unconditional love and unwavering support throughout this journey. None of this would have been possible without them. I am especially grateful to my wife, Dr. Ruijie Meng, who has walked beside me through every step of this journey. Her love, support, and encouragement have lifted me in countless ways, and together we have grown, strengthened, and become better versions of ourselves.

Table of contents

Chapter 1

Introduction

Text summarization is the task of enabling computers to automatically produce a shorter version of a text while preserving the most salient information from the original (El-Kassas et al., 2021a; Lin and Ng, 2019). It has been a popular research area of natural language generation (NLG) and natural language processing (NLP) (Jurafsky and Martin, 2025). With vast amounts of texts generated daily across domains such as news, scientific research, and social media, text summarization can help address the growing challenge of information overload for humans in the digital age by automatically distilling key information into concise summaries. Because summaries enable us to quickly grasp the main content from lengthy or complex texts, text summarization could enhance the time efficiency for us to process information. For example, if computers could generate summaries for news articles (Fabbri et al., 2019), scientific papers (Lu et al., 2020), or even medical record documents (Jain et al., 2022), it would save time for readers to stay informed about world events, for researchers to understand others' research, and for doctors to quickly grasp patients' medical histories.

The input of text summarization could be a single document or multiple documents. Based on the number of input documents, text summarization is typically categorized into single-document summarization (SDS) and multi-document summarization (MDS), as illustrated in Fig. 1.1. MDS generates the summary for a cluster of documents, while SDS produces the summary only for a single document. For example, MDS can generate a summary for multiple news articles to provide an overview of an event from various sources, while

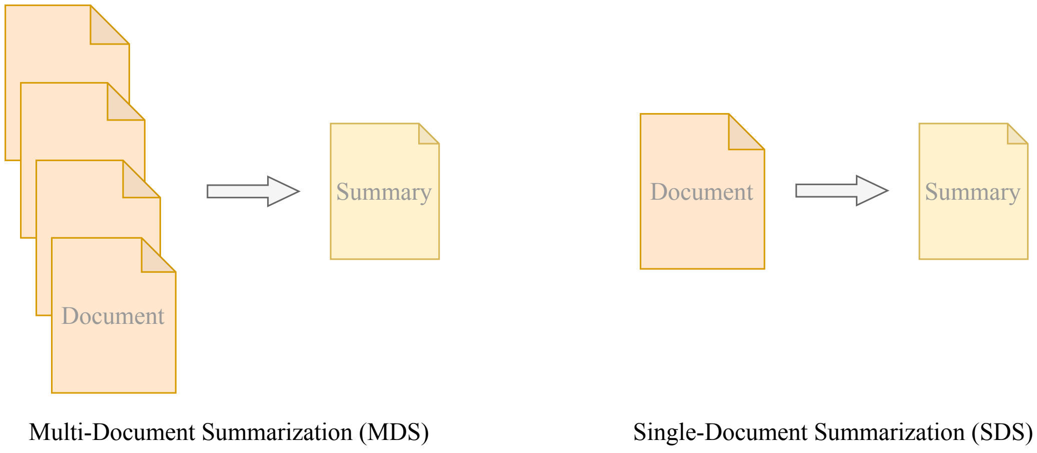

Document
Summary
Multi-Document Summarization (MDS)
Document
Summary
Single-Document Summarization (SDS)

Fig. 1.1 The difference between multi-document summarization and single-document summarization.

SDS can summarize a single book to serve as a brief introduction. MDS is arguably more challenging than SDS because MDS requires the capability to correlate and integrate information drawn from multiple sources (Radev, 2000; Radev and McKeown, 1998). The multi-source information from different documents may have complex relationships (Radev, 2000). For example, the news articles may contradict, complement, or repeat each other. MDS must consider these relationships to properly integrate the multi-source documents. The input length of multiple documents is also generally longer and MDS systems will need to handle that. While text summarization has mostly focused on SDS (El-Kassas et al., 2021b), this thesis focuses to advance the development of MDS.

Various computational models have been developed to enable computers to perform automatic summarization. Depending on how the summaries are generated, text summarization can be further classified as either abstractive or extractive. Abstractive models generate summaries using words or expressions that may not appear in the original source documents, whereas extractive models selects and reuses sentences directly from the input. Although extractive models are simpler to build, their outputs are often lengthy, redundant, and less coherent, leading to a poor reading experience (Ma et al., 2020). In contrast, human-written summaries are inherently abstractive, as they typically paraphrase content and include words or expressions not found in the original text. As a result, abstractive text summarization more closely mirrors human summarization behaviours, but it demands more advanced natural language understanding and generation capabilities. This thesis focuses on abstractive summarization. In this thesis MDS refers to *abstractive* summarization, unless specified.

We will now introduce the critical research challenges for MDS in Section 1.1 and research questions that this thesis focuses on in Section 1.2.

1.1 Research Challenges

The development of computational models for automatic text summarization relies on benchmark datasets, effective modelling techniques, and reliable evaluation metrics, in line with the data-driven learning paradigm for artificial intelligence (AI) (Jurafsky and Martin, 2025). However, multi-document summarization (MDS) remains a challenging task due to the scarcity of high-quality benchmark datasets, the limitations of current modelling techniques, and the inadequacy of existing evaluation metrics.

1.1.1 Flawed Benchmark Datasets

The development of modern text summarization models relies on labelled benchmark data with human-written summaries in both training and evaluation (Fabbri et al., 2021; Lewis et al., 2020; Zhang et al., 2020a). However, there are few high-quality datasets with human-written summaries (Ma et al., 2020). This is because constructing MDS datasets with human-written summaries is time-consuming and labour-intensive. Input documents are usually long on their own and we have limited processing speed and working memory to understand the complex inter-document relationships (Baddeley, 2003; Deary et al., 2010). Writing summaries for multiple documents may also requires domain expertise, which makes it costly. Most existing MDS datasets tend to favour extractive models because most words or phrases in the ground truth summary are directly extracted from the input documents (Ma et al., 2020). Although the emergence of large language models (LLMs) has changed the training paradigm of summarisation models (in that it is now possible to build summarisation models that doesn't require extensive training) (Zhang et al., 2024a), we still need human-written summaries to evaluate the model performances. Lastly, even in MDS datasets that have ground truth summaries, they do not provide explicit cross-document relationships such as conflicts among input documents which is essential to investigate how well summarization models process these different relationships and integrate information across the documents (Fabbri et al., 2019; Ghalandari et al., 2020). Therefore, this motivates the construction of high-quality MDS datasets with human-written summaries and rich cross-document relationships.

1.1.2 Limited Modelling Techniques

To generate a summary of multiple documents, summarization models have to learn to correlate and integrate information from multiple sources (Ma et al., 2020). It is challenging to model multi-document inputs mainly for two reasons. One reason is that the input documents are usually long, so it poses a challenge to summarization models (Fabbri et al., 2019; Ghalandari et al., 2020). Most summarization models model the input documents by concatenating them into a long text sequence (Beltagy et al., 2020; Guo et al., 2022; Zhang et al., 2024a). However, such format is not ideal as structural information within and between documents are lost. Because the input is long, MDS will also need to model long-range dependencies in the input documents. However, modelling long-range dependencies in texts is a long-standing challenge (Jurafsky and Martin, 2025). The Transformer architecture, as the backbone of all current summarization models, can theoretically capture any dependencies in a long text as it can directly connect any two words in the input text sequence with the self-attention mechanism. However, in practice, the ability of Transformer-based models to retain and use relevant information from far-off words can degrade because of the attention dilution problem (Kitaev et al., 2020).

The other reason is that integration of multiple documents requires understanding of complex cross-document relationships (Radev, 2000). For example, any two input documents may contradict or complement each other. The generated summary can be grounded in the input documents only if the summarisation models can accurately capture these cross-document relationships. Otherwise, the generated summary may not be faithful (i.e., summaries are not consistent with the content in input documents). Although a number of approaches have been developed to better model multi-document inputs, it remains unclear how effectively they capture different cross-document relationships because they lack transparency in terms of their decision making processes. Ultimately, this means it is difficult to understand how cross-document information is integrated during summary generation (Ma et al., 2020; Xiao et al., 2022).

1.1.3 Suboptimal Evaluation Metrics

The quality of model-generated summaries is typically assessed with automatic evaluation metrics. Although various evaluation metrics have been developed, it is still challenging to accurately assess the summary quality for several reasons (Gehrmann et al., 2023a). Firstly, standard evaluation paradigm typically assumes a human-written summary as the sole ground truth, even though in practice there may be different ways we can summarise the documents. For example, the salient information of the input documents may vary for different users. Evaluation based on a single ground truth as such is flawed. Secondly, putting the issue of multiple ground truths aside, evaluation metrics must be able to effectively measure the semantic similarity between the human-written summary and the model-generated summary. This is because the ground truth summary may be paraphrased in different ways, and the evaluation should assess the semantic difference rather than relying on surface-form comparison, e.g., ROUGE (Lin and Hovy, 2003). That is, the generated summary may have the same meaning but use different words or phrases compared with the human-written summary, and an ideal evaluation metric should recognize that it is a good summary. Lastly, an ideal summarization system should not only generate plausible summaries but also justification for them. As such, evaluation metrics for MDS should assess not only the quality of the generated summary, but also how the model generates the summary from multiple documents. However, most existing metrics assess only the final summaries without accounting for the generation process.

1.2 Research Questions and Contributions

To understand and properly treat the summarization process, we need a theoretical foundation that encompasses syntactic, semantic, and reader-oriented considerations. Systemic Functional Linguistics (SFL) provides such as foundation by positing that human language operates through three primary metafunctions (Halliday, 1970). First, it expresses *ideational* content about the world—events, entities, actions, and processes—through the ideational metafunction. Second, it conveys *opinionated* content—attitudes, evaluations, emotions, judgments, and stance—via the interpersonal metafunction. Third, it organizes information coherently and enables the flow of discourse through the textual metafunction, which manages how meaning is packaged and presented in context. Summarisation typically focuses on texts categorized as ideational meanings (e.g., news articles) or interpersonal meanings (e.g., product reviews). Summarizing ideational texts involves identifying and presenting their key factual content, whereas summarizing opinionated texts requires synthesizing the overall stance they convey. This thesis investigates three research questions focused on the summarization of ideational and opinionated documents by tackling the challenges in Section 1.1.

1.2.1 How to integrate ideational information from multiple documents to generate better summaries?

For the first research question, we explore the summarization of multiple documents which are mainly composed of ideational information. Ideational documents such as news articles and scientific papers are primarily composed of ideational or objective information in the world. Summarizing ideational texts involves identifying and presenting their key factual content. A cluster of ideational documents such as news articles are connected with underlying facts that may have complex cross-document relationships, including contradiction, redundancy, and complementary information (Radev, 2000). To generate summaries that can integrate dispersed information from different ideational source documents, a summarization model needs to understand the relationships among the sources.

To handle the multi-document input, existing summarization models concatenate the input documents into a text sequence and are expected to somehow learn to integrate information from them through supervised training (i.e., by training them to predict the ground truth summary) (Xiao et al., 2022). However, they sometimes produce plausible but unfaithful summaries (Xiao et al., 2022). We hypothesize that incorporating explicit graph representation of inter-document relationships into the summarization process would help improve the quality of generated summaries. Although graphs have been used to represent source documents for MDS, they are in fact homogeneous (i.e., nodes or edges in the graphs are of a single type) which has limited capability for capturing cross-document relationships (Cui and Hu, 2021; Jin et al., 2020; Li et al., 2020; Li and Zhuge, 2021).

Therefore, this thesis explores using heterogeneous graphs to represent the source documents and incorporate these graphs into the summarization process. To achieve this, we propose to incorporate the heterogeneous graph representations into PLMs with multi-task training: heterogeneous graph compression and text summarization. Our model achieve the state-of-the-art performances in terms of ROUGE (Lin and Hovy, 2003) and BERTScore (Zhang et al., 2020b) on widely-used datasets including ARXIV (Cohan et al., 2018), MultiNews (Fabbri et al., 2019), and WCEP-100 (Ghalandari et al., 2020).

1.2.2 How to integrate opinionated information for opinion summarization?

This research question focuses on summarizing multiple documents which are mostly composed of opinionated information, and the summary is to merge and synthesize the overall opinions among the input documents. While summarization of ideational documents (e.g., news articles) aims to extract the most critical and objective facts or events from the documents, summarization of opinionated documents (e.g., product reviews) seeks to aggregate subjective user sentiments and opinions from the documents. The former demands factual accuracy and a focus on core information, while the latter is driven by the synthesis of attitudes and judgments, requiring the model to identify and categorize opinions.

Most of MDS datasets are based on ideational sources such as news articles (Fabbri et al., 2019), Wikipedia pages (Liu et al., 2018), or scientific papers (Cohan et al., 2018). These datasets generally do not provide any explicit cross-document relationships (e.g., whether one document contradicts another). For opinionated MDS datasets, there are few datasets and they mostly focus on product reviews (Brazinskas et al., 2021) or business reviews (Angelidis et al., 2021). The input documents for these opinion summarization datasets are short; typically one or several sentences. Like the ideational datasets, these datasets generally do not provide any information about the cross-document relationships either.

In this thesis, we focus on the scientific reviews and propose framing the generation of meta-reviews as a opinionated MDS task. We choose this domain for three reasons: (1) the input documents (i.e. paper reviews) are long (compared with product reviews); (2) there are rich inter-document relationships (e.g. disagreement between reviews); and (3) the meta-review generation process is complex and may not simply be the majority sentiment, as it should take into account the *strength of arguments* in the reviews. We first collect the source reviews and ground truth meta-reviews from scientific peer-review platforms. We derive cross-document relationships from the review scores (which capture disagreements/conflicts) and explicit conversational structure provided by these platforms. Next, we explore incorporating the explicit conversational structure among the source documents into the summarization model to improve the quality of generated meta-reviews. We implement a baseline model based on incorporating the conversational structure by manipulating attentions of the Transformer architecture. Our experiments results show that our simple baseline significantly outperforms other strong models. Manual analysis on the generated meta-reviews reveals that most models cannot recognize and resolve conflicts in the reviews, suggesting a promising avenue for future research.

Lastly, we propose an aspect-aware opinion consolidation framework that we hypothesize human meta-reviewers follow to improve generation quality and transparency of the summarization models. Our experiments show that prompting based on our framework generate better meta-reviews than other strong prompting approaches.

1.2.3 How to build grounded and transparent opinion summarisation systems that work across different domains?

While the thesis focuses on opinion summarization in the scientific domain for the second research question, we further explore using a similar aspect-aware framework to build a grounded and transparent opinion summarisation system that work across different domains, such as product and hotel reviews. In other words, we're asking the next question: can we develop a domain-general opinion summarization model.

There have been two general approaches proposed for abstractive summarization of reviews: end-to-end and pipeline-based approaches. The end-to-end approaches lack transparency due to its black-box nature (Beltagy et al., 2020; Liu and Lapata, 2019; OpenAI, 2023; Xiao et al., 2022). The pipeline-based approaches first extract information clusters in the format of sentences, and then generate summaries based on the clusters (Bhaskar et al., 2023b; Hosking et al., 2024). The pipeline-based approaches are typically designed for a specific domain. Although less opaque than end-to-end models, they are ultimately not completely transparent because their extracted clusters or intermediate steps have not been validated.

We achieve domain-general transparent opinion summarization with a pipeline approach using verified modules. We explore decomposing the summarization process across different domains following underlying aspects in opinion summarization, and investigate using the decomposition framework to improve the summary quality and the evaluation methodology in opinion summarization. We implement the pipeline by prompting LLMs. It first identifies aspect-related text fragments, then generate aspect-focused meta-review, and lastly combine all the aspect-focused meta-reviews to generate the final meta-review. This is different from our approach for the second research question although both are guided by review aspects. The length of the text fragments is dynamic, and that makes the model work for multiple domains and consider more opinion information such as justification when generating the final meta-review; in contrast, our approach for the second research question only considers sentiments and extracts sentiments based on the predefined format for scientific opinions before generating the final meta-review. Our experimental results show that this approach generates better meta-reviews in most domains domains with higher aspect coverage and more faithful generations than strong baselines while on par with our prior approach in the scientific domain. Additionally, we also find that our approach can generate intermediate steps which help human meta-reviewers to write better reviews in less time. This work ultimately shows that integration of reasoning traces into prompting benefit LLMs based MDS.

1.3 Thesis Structure

The remainder of the thesis is structured as follows. The second chapter focuses on the literature review in benchmark datasets, multi-document modelling techniques and evaluation metrics (Chapter 2). The work for the three research questions is then presented in the following three chapters, including ideational information integration (Chapter 3), scientific opinion summarization (Chapter 4) and domain-general opinion summarization (Chapter 5). The last chapter concludes the thesis with a summary of the work and future directions (Chapter 6).

Chapter 2

Literature Review

Following the research challenges in Chapter 1, this chapter reviews the situation of existing benchmark datasets for multi-document summarization (MDS) (Section 2.1), multi-document modelling approaches (Section 2.3), and evaluation metrics (Section 2.4). In addition, as the thesis experiments with language models, fundamental language modelling architectures will be explained, such as the Transformer architecture and large language models (LLMs) (Section 2.2).

2.1 Benchmark Dataset Construction

MDS is the task to automatically generate a summary of multiple documents. Given a set of m related input documents $\mathscr{D} = \{d_0, d_1, \ldots, d_m\}$ (i.e., a document cluster), MDS generates a textual summary $\hat{z} = \hat{w}_0, \hat{w}_1, \ldots, \hat{w}_T$ (composed of T words) that captures the essence of the documents.

There have been an increasing number of public datasets for MDS. We present the widely-used datasets in Table 2.1. These datasets are constructed using texts from various domains, and they contain different scales of samples, up to more than one million. Domains of the documents include biomedical articles (DeYoung et al., 2021), news articles (Fabbri et al., 2019; Ghalandari et al., 2020), academic papers (Lu et al., 2020), Wikipedia articles (Liu et al., 2018), product reviews (Brazinskas et al., 2021), business reviews (Angelidis et al.,

Dataset	Document Source	#Samples (train/val/test)	#Docs/Sample	#Words (in/out)
MS^2 (DeYoung et al., 2021)	Biomedical articles	14,188/2,021/1,667	24.0	6,974.3/61.2
WCEP (Ghalandari et al., 2020)	News articles	8,158/1,020/1,022	63.6	26,875.5/63.6
Multi-XScience (Lu et al., 2020)	Academic papers	30,369/5,066/5,093	4.4	691.5/105.1
Multi-News (Fabbri et al., 2019)	News articles	44,972/5,622/5,622	2.8	1768.6/216.7
WikiSum (Liu et al., 2018)	Wikipedia articles	1,579,360/38,144/38,205	40.0	2249.6/120.1
AmaSum (Brazinskas et al., 2021)	Product reviews	25,203/3,114/3,166	326.6	13,611.0/73.8
MetaTomatoes (Peper et al., 2024a)	Product reviews	0/0/212	67.9	1,582.9/168.7
FuseReviews (Slobodkin et al., 2024)	Business reviews	643/99/258	8.0	608.5/66.2
SPACE (Angelidis et al., 2021)	Business reviews	0/25/25	100	14,335.1/73.7

Table 2.1 An overview of MDS datasets. All these datasets are in English. (in: input, out: output)

2021), and scientific reviews (Li et al., 2023a). All these datasets are in English. The datasets are mostly based on two types of texts: ideational and opinionated documents. Based on the content type of the input documents, MDS is further categorized into ideational and opinionated MDS.

2.1.1 Ideational Documents

For ideational MDS, WikiSum (Liu et al., 2018) is the first large-scale dataset with more than million instances. The dataset uses the lead sections of Wikipedia pages as summaries and the input documents are the cited articles and another ten related articles from the Google Search. Ghalandari et al. (2020) develop the Wikipedia Current Events Portal (WCEP) dataset based on news events listed in the Wikipedia Current Events Portal[1]. WCEP uses the description of news events as the summary and linked external news articles as input documents. WCEP contains 10,200 instances and each instance contains one human-written summary and 235 articles as input documents on average. To retrieve more input documents in each cluster, WCEP includes similar articles in the Common Crawl News dataset[2]. Because of its training data size, these two datasets are popularly used in MDS research. However, both datasets have problems as they augment source documents with externally retrieved documents. As such, the summary may not be a faithful summary of the source documents.

[1] https://en.wikipedia.org/wiki/Portal:Current_events

[2] https://commoncrawl.org/2016/10/news-dataset-available/

Multi-News (Fabbri et al., 2019) is constructed based on summarization on news articles. The summary summarizes multiple news articles on the same event. The summaries are written by professionals. It is the first large-scale dataset for MDS with human-written summaries. While DUC 2004[3] and TAC 2011[4] also contain human-written summaries of news articles, they have only a few hundred instances and the input documents are also shorter.

Scientific papers have also been used to construct MDS datasets. Multi-XScience (Lu et al., 2020) is constructed using the paragraph of the related work section as the summary, and the cited articles' abstract are the input documents. However, the summary are not always reflective of the cited articles as the authors find that less than half of the statements are grounded by the input documents. MS^2 (DeYoung et al., 2021) uses biomedical papers as source documents for MDS. It is one of the first MDS datasets in the biomedical domain. The summaries are systematic literature reviews that synthesize results across a number of other medical studies on specific clinical questions. The input documents are the medical studies.

2.1.2 Opinionated Reviews

For opinion summarization, the input is usually composed of reviews, e.g., product reviews and business reviews. Source reviews for each instance present opinions on the same entity, and and they may comment on specific aspects (Hu and Liu, 2004). For example, hotel reviews may focus on aspects such as cleanliness, food, location, rooms, and service of hotels (Angelidis et al., 2021). The summarization task is to generate a summary (i.e., meta-review) that captures the overall opinions among the source reviews for these different aspects.

SPACE (Angelidis et al., 2021) is built using hotel reviews from TripAdvisor[5], in which the input is a cluster of customer reviews and the output is a human-written summary. It features aspect-based human-written summaries in the dataset, i.e. summaries that focus on a

[3] https://duc.nist.gov/duc2004/
[4] https://tac.nist.gov/2011/Summarization/
[5] https://www.tripadvisor.co.uk/

specific aspect. An aspect-based summary only summarize singular customer opinions on one aspect. AmaSum (Brazinskas et al., 2021) is a dataset on product reviews from four products such as electronic consumer products and sports shoes. In the dataset, summaries are written by professional reviewers and each summary has three components, verdicts, pros and cons, to emphasize the most important points about a product and focusing on fine-grained aspects. The input is a collection of reviews by customers from Amazon. FuseReviews (Slobodkin et al., 2024) is another opinion summarization dataset using business reviews as input documents. Different from the other datasets, in FuseReviews there are highlighted spans within the source reviews, and its human-written summary is to integrate the highlighted spans to generate a coherent and concise summary. MetaTomatoes (Peper et al., 2024b) is a dataset on movie reviews from RottenTomatoes[6]. The meta-reviews (i.e., summaries) are written by the RottenTomatoes editorial team and the input documents are the reviews. Compared with other opinion summarization datasets, MetaTomatoes has the longest input documents on average.

However, none of these datasets provide any explicit annotations on inter-document relationships (e.g., conflicts among source documents) although various inter-document relationships commonly exist in these real-world datasets. As MDS is to integrate information from multiple documents, summarization models must understand inter-document relationships among the source documents. Without detailed relationship information for input documents, it slows down the progress to investigate summarization models in understanding different multi-document relationships. This leads to under-explored research on inter-document relationship comprehension of summarization models. This motivated our research on constructing a dataset with more complex and explicit cross-document relationships in Chapter 3.

[6] https://www.rottentomatoes.com/

2.2 Fundamental Language Modelling

Before discussing the related work on how to model source documents for summarization and how to evaluate the generated summaries, this chapter reviews fundamental language modelling techniques from traditional language models including n-gram language models to large language models (LLMs). This is because that language models are used to model documents in modern MDS research. This thesis is also mostly based on LLMs.

Language models are originally used to predict upcoming words in a word sequence (Shannon, 1948). Formally, to estimate the probability of an upcoming word, w_m, a conditional probability, $\mathrm{p}(w_m|w_{1:m-1})$, is assigned to the word given the preceding $m-1$ words, $w_{1:m-1}$. Language models can be also used to assign a probability to an entire sequence. The joint probability of the sequence of words can be decomposed using the chain rule of probability, $\mathrm{p}(w_{1:m}) = \mathrm{p}(w_1) \cdot \mathrm{p}(w_2|w_1) \cdot \ldots \cdot \mathrm{p}(w_m|w_{1:m-1})$. To estimate the probability more effectively and efficiently, different modelling approaches have been proposed in the literature of language modelling.

2.2.1 Traditional Language Models

We first introduce the development of different language model architectures from n-gram language models to the Transformer architecture which are the foundation of today's language artificial intelligence (AI).

N-gram Language Models We can estimate $\mathrm{p}(w_m|w_{1:m-1})$ and $\mathrm{p}(w_{1:m})$ by counting frequencies of words and sequences if we had a large corpus. However, it is computationally infeasible to estimate the probabilities by counting the large corpus. This is because if m is large we need to take a large combination of words into account. The exponential combination may be a totally new word sequence that has never seen in any large corpus. For example, if the length of the word sequence is 50 and the vocabulary size is $2K$ there may be $(2K)^{50}$ different sequences, which is larger than any corpus and the entire web. Therefore, the n-gram model comes into play (Brown et al., 1990). Based on the Markov assumption, we can just estimate the probability of the next word based on only a few previous words instead of the entire proceeding word sequence. The n-gram model uses only $n-1$ words as the previous words, $\mathrm{p}(w_m|w_{1:m-1}) \approx \mathrm{p}(w_m|w_{m-n+1:m})$. However, the n-gram language model has its issues. It has to use a large number of parameters to maintain probabilities of different n grams while most of these probabilities cannot be estimated with maximum likelihood estimation (MLE) as they have never been seen in any corpus. For example even when $n = 3$ there are $(20K)^3$ parameters and not all of them are seen in the corpus. The second issue with n-gram language models is the short context assumption, which means they cannot model long-range dependencies in language.

Feedforward Neural Language Models To solve the parameter sparsity issue of n-gram language models, Bengio et al. (2003) introduced neural language models with word embeddings based on simple feedforward neural networks. The feedforward neural language model uses dense vectors as word representations (i.e., word embeddings) and a fully-connected feedforward neural network to predict the probability of upcoming words given the prior $n-1$ words. Formally, the feedforward language model is composed of different neural network layers, namely the input layer, the embedding layer, the hidden layer and the output layer. The input layer transforms the prior $n-1$ words into one-hot vectors, $\boldsymbol{X}$, where one element equal to 1 while all the other other elements are set to 0 in each vector. The embedding layer then transforms the one-hot vectors into real valued embeddings, $\boldsymbol{E} = \mathrm{embedding}(\boldsymbol{X})$. The entire process is described in Equation 2.1.

$$\boldsymbol{H} = \sigma(\boldsymbol{W} \cdot \boldsymbol{E} + \boldsymbol{b}) \tag{2.1}$$

$$\hat{\boldsymbol{y}} = \mathrm{softmax}(\boldsymbol{U} \cdot \boldsymbol{H}) \tag{2.2}$$

where, $\boldsymbol{H}$ is the output of the hidden layer, $\hat{\boldsymbol{y}}$ has $|V|$ elements and each element denotes the probability of the next word in the vocabulary V, σ is a non-linear transformation function, and $\boldsymbol{W}$, $\boldsymbol{b}$ and $\boldsymbol{U}$ are the parameter matrices for the model. The feedforward neural language model is trained with back-propagation optimization on a cross-entropy loss. For each word, the model is trained to minimize the difference between the predicted distribution $\hat{\boldsymbol{y}}$

and the gold distribution $\boldsymbol{y}$ which is actually the one-hot vector of the expected word. The feedforward neural language model uses only a small set of parameters with neural networks to model the probability of the next word and it can be generalize to unseen words because embeddings of semantically similar words are close to each other in the embedding space.

Recurrent Neural Language Models To make language models cover longer contexts, recurrent neural networks (RNNs) are introduced to directly model the sequential dependencies in human languages (Mikolov et al., 2010) with considering the entire context instead of n-1 words in the n-gram model and feedforward language model. RNNs are composed of layers of neural networks where the output value is directly dependent on its own earlier outputs as an input (Mikolov et al., 2010). The recurrent neural language model is also composed of an input layer, an embedding layer, an hidden layer and an output layer.

Different from the feedforward neural language models, recurrent neural language models process the input text word by word sequentially. The output layer of the recurrent neural language model $\boldsymbol{h}_t$ is not only based on the current word embedding w_t but also the hidden layer output $\boldsymbol{h}_{t-1}$ for the prior word w_{t-1}. The probability of the n-th word is predicted based on the hidden layer output for the $(n-1)$-th word. The modelling process is described in Equation 2.3.

$$\boldsymbol{h}_t = \mathrm{g}(\boldsymbol{U}_1 \cdot \boldsymbol{h}_{t-1} + \boldsymbol{W} \cdot \boldsymbol{e}_t) \tag{2.3}$$

$$\hat{\boldsymbol{y}}_n = \mathrm{softmax}(\boldsymbol{U}_2 \cdot \boldsymbol{h}_{n-1}) \tag{2.4}$$

where g is another non-linear activation function. Therefore, in principle recurrent neural language models can model variable-length texts and do not have the limited context problems that n-gram and feedforward neural language models have. Recurrent neural language models are also trained with minimizing the cross-entropy loss towards the gold words. However, it is quite difficult to train recurrent neural language models to predict the next word based on long input because of the problem of vanishing gradients in training. To solve the problem, long short-term memory (LSTM) network (Hochreiter and Schmidhuber, 1997) is introduced to use gates to control information flow to preserve gradients across longer time steps.

Text Tokenization The input of these language models is originally word sequences. The input text is just split into word-form units. However, it leads to a large vocabulary to represent all words in the world and there are a large number of words that are rare or unseen in the training data. Subword tokenization is introduced for machine translation and widely used in the entire community (Kudo and Richardson, 2018; Sennrich et al., 2016; Wu et al., 2016). For example, "text summarization" can be tokenized into <"text", "summar", "ization">. It reduces the vocabulary size by splitting words into smaller and reusable units. Because words share small units, subword tokenization makes it possible for language models to handle out-of-vocabulary words. Therefore, in practice we use tokens (i.e., subword units) in language models.

The Sequence-to-Sequence Model In text generation tasks like machine translation and text summarization, the generation is conditioned on a context (e.g., sentence in the source language in machine translation and source document in summarisation). The output text $\hat{Z}$ given the context X is modelled as follows.

$$\mathrm{p}(\hat{Z}|X) = \prod_{i=0}^{T} \mathrm{p}(\hat{w}_i|\hat{w}_0, \hat{w}_1, \ldots, \hat{w_{i-1}}, X) \tag{2.5}$$

where X and $\hat{Z} = \{w_0, w_1, ..., w_T\}$ are both token sequences. The sequence-to-sequence model is proposed to model this conditional probability with two recurrent neural language models (Cho et al., 2014; Sutskever et al., 2014). The model follows an encoder-decoder framework. Essentially, both the encoder and the decoder follow a recurrent neural language model architecture. The encoder takes the input sequence of tokens, X, and output a context vector which is expected to convey the essence of the input sequence. The decoder takes the context vector and it generates the output with an arbitrary length autoregressively. Formally,

$$\boldsymbol{h}_t^e = \mathrm{RNN}(\boldsymbol{h}_{t-1}^e, \boldsymbol{e}_t) \tag{2.6}$$

$$\boldsymbol{h}_t^d = \mathrm{RNN}(\boldsymbol{h}_{t-1}^d, \boldsymbol{e}_{t-1}, \boldsymbol{c}), \quad \boldsymbol{h}_0^d = \boldsymbol{c} \tag{2.7}$$

$$\hat{\boldsymbol{y}}_t = \mathrm{softmax}(\boldsymbol{h}_t^d) \tag{2.8}$$

where *RNN* denotes a recurrent neural networks, $\boldsymbol{h}_t^e$ is the hidden state of the t-th step in the encoder, $\boldsymbol{h}_t^d$ is the hidden state of the t-th step in the decoder, $\boldsymbol{c}$ is the context vector. The context vector is considered in generation of each token in the output sequence. The sequence-to-sequence model is trained based on the teacher forcing technique, which includes the ground truth token to predict the next token in training instead of the predicted one from the last step. However, in practice the context vector is the bottleneck of the encoder-decoder architecture. This is because the decoder has only access to the context vector which is represented with the last hidden state of the encoder and it may not have enough capacity to capture the full semantics of the source text. The attention mechanism is then proposed to make the decoding of tokens in the output text attend to all the hidden states from the source text directly (Bahdanau et al., 2015), instead of only considering the last hidden state of the encoder. We will discuss details of the attention mechanism next when we introduce the transformer architecture, which is now the de-facto standard architecture of building LLMs. The models developed in this thesis all use the Transformer architecture.

2.2.2 The Transformer Architecture

Because of the recurrent nature of recurrent neural language models, they can not be parallelized in training. The Transformer architecture is proposed to solve the problem of training efficiency by taking advantage of efficient matrix multiplication routines (Vaswani et al., 2017). Similar to the sequence-to-sequence model taking the sequence X as the input and $\hat{Z}$ as the output (Equation 2.5), the Transformer architecture also follows an encoder-decoder framework, shown in Figure 2.1. It has become a cornerstone of modern language generation (Zhao et al., 2023).

The encoder is composed of a stack of Transformer encoder blocks. Each encoder block is a multi-layer network and it is composed of a multi-head attention layer and a feedforward layer with layer normalization between them. The Transformer encoder blocks transform the input embeddings $\boldsymbol{E}$ of X into contextual representations $\boldsymbol{H}$ in vectors. To make the model sensitive to positions of words, position embeddings are added to the word embeddings. The decoder is composed of Transformer decoder blocks with the same number as the

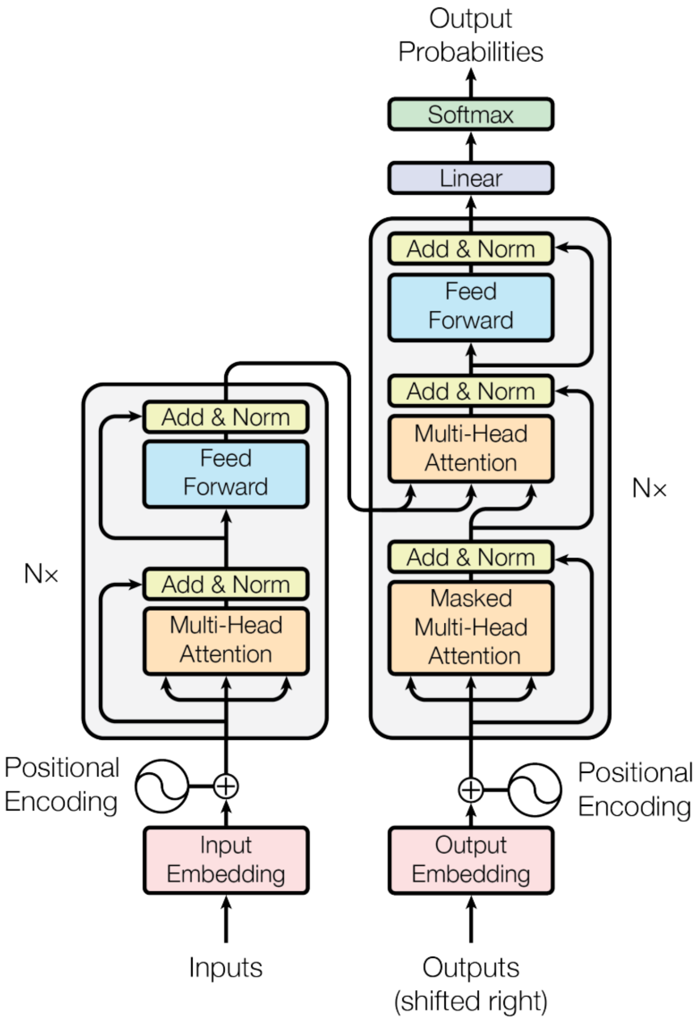


Fig. 2.1 The Transformer architecture with the encoder (the left part) and the decoder (the right part). The figure is from Vaswani et al. (2017).

Transformer encoder blocks. In each decoder block, besides the two layers in the encoder block, there is another layer to get the information from the encoder, called encoder-decoder multi-head attention layer. Finally, on the top of the decoder blocks, similar to the sequence-to-sequence model, the softmax function works on the output of the decoder blocks to predict the probability distribution over the vocabulary for each generated word. The Transformer architecture also follows an autoregressive generation process, and they are trained with the teacher forcing technique to minimize the cross-entropy loss between the predicted and gold sequences.

The most important part in the Transformer architecture is the multi-head attention and its two variants, the encoder-decoder multi-head attention and the masked multi-head attention. The multi-head attention is designed to learn the contextual representation of word meaning by weighting and combining the representations from appropriate other words in the same sequence or the other sequence. The attention mechanism is calculated based on a query, a key and a value and they are intuitively to transform the original embedding matrix with attention weights among each other. The attention transformation in the encoder block could be formalized as:

$$\boldsymbol{Z} = \mathrm{softmax}\left(\frac{\boldsymbol{Q}\cdot\boldsymbol{K}^{\top}}{\sqrt{d_k}}\right)\cdot\boldsymbol{V} \tag{2.9}$$

$$\boldsymbol{Q} = \boldsymbol{X}\cdot\boldsymbol{W}^{Q}; \quad \boldsymbol{K} = \boldsymbol{X}\cdot\boldsymbol{W}^{K}; \quad \boldsymbol{V} = \boldsymbol{X}\cdot\boldsymbol{W}^{V} \tag{2.10}$$

where $\boldsymbol{X}$ of shape $[m\times d]$ is the embedding matrix of the input sequence of m words, $\boldsymbol{Z}$ of shape $[m\times d_v]$ is the output of the attention transformation, $\boldsymbol{W}^{Q}$ of shape $[d\times d_k]$, $\boldsymbol{W}^{K}$ of shape $[d\times d_k]$, and $\boldsymbol{W}^{V}$ of shape $[d\times d_v]$ are projection matrices to get the query $\boldsymbol{Q}$ of shape $[m\times d_k]$, the key $\boldsymbol{K}$ of shape $[m\times d_k]$, and the value $\boldsymbol{V}$ of shape $[m\times d_v]$, respectively. This is only for one single head in the multi-head attention. In multi-head attention, the heads have different projection matrixes and the final output is a concatenation of the outputs from all attention heads. In the encoder-decoder multi-head attention, the calculation of the output from each head is similar to the Transformer encoder multi-head attention and the only difference is:

$$\boldsymbol{Q} = \boldsymbol{Y}\cdot\boldsymbol{W}^{Q} \tag{2.11}$$

where $\boldsymbol{Y}$ of shape $[n\times d]$ denotes the output sequence of n words. It means that the attention score is based on the similarity between the generated word sequence and the input word sequence (e.g., source sentence in translation and input texts in summarisation). For the masked multi-head attention in the decoder, to respect its autoregressive constraint, each word should only attend to prior words in generation. Therefore, the output of the attention in a single head is formalized as;

$$\boldsymbol{Z} = \mathrm{softmax}\left(\mathrm{mask}\left(\frac{\boldsymbol{Q} \cdot \boldsymbol{K}^{\top}}{\sqrt{d_k}}\right)\right) \cdot \boldsymbol{V} \tag{2.12}$$

where $\mathrm{mask}(\cdot)$ is a function that assigns infinity to the the element (i, j) in the result matrix $\boldsymbol{Q} \cdot \boldsymbol{K}^{\top} / \sqrt{d_k}$ if j is larger than i.

Compared with traditional language models, the Transformer architecture can model long-dependency structures in texts and have parallel processing for sequential data. The self-attention mechanism can directly connect every two tokens, better capturing longer contexts than traditional language models without gradient vanishing. In addition, the Transformer architecture can process all tokens in a sequence simultaneously, different from recurrent neural language models which recurrently process tokens in a sequence. The efficient parallelism makes it possible to scale Transformer-based language models. We will discuss in the next sections.

The encoder-decoder framework of these sequence-to-sequence models is originally introduced for translation, but we do not need to use both the encoder and decoder to build language models. In practice we can use either encoder, decoder or both, and we will discuss this further next section.

2.2.3 Pre-trained Language Models

Pre-training enables language models to learn linguistics and world knowledge from vast amounts of texts (Han et al., 2021). There are three types of Transformer-based pre-trained language models, using the original encoder-decoder architecture, or only the encoder or decoder. As the decoder-only and encoder-decoder pre-trained Transformers are the most two important variants for language generation while this thesis focuses on text summarization, we only discuss them here. We will first describe the pre-training techniques for these two types of Transformer models, and then how to fine-tune them on the downstream tasks.

Encoder-Decoder Pre-trained Transformers The encoder-decoder Transformers could be pre-trained with the objective of de-noising or predicting missing spans. T5 (Raffel et al., 2020) and BART (Lewis et al., 2020) are the most representative encoder-decoder pre-trained language models. In the training process, T5 unifies all the natural language processing tasks in to a text-to-text framework and learn to predict the consecutive spans in the input texts. BART is trained to reconstruct the original text based on the corrupted version. There are different types of corruption operations in different levels. For example, masking out one or two tokens in the original texts, masking one entity, and deleting tokens in the original texts. These models are developed for conditional text generation in general purpose, and they can be used in various tasks such as translation, question answering, and text summarization. There are also pre-trained language models tailored for specific tasks. For example, PEGASUS (Zhang et al., 2020a) is trained with the pre-training objective which forces the decoder to generate masked gap sentences in the decoding process, i.e., masking out several sentences from the source document and recovering them in order in the decoder. With this objective, the model is taught to identify and aggregate salient information in the source document during pre-training. This makes the model trained without reference summaries. The gap sentences are extracted from the source document in different strategies. The first strategy is simply select first sentences. The second strategy is to randomly select several sentences from the source document. The last strategy which is more effective is to select top ranked sentences according their importance. The importance is computed based on the ROUGE score between the sentence and the rest of the document. Although PEGASUS has good performance on abstractive summarization, it is not straightforward to use them for MDS, because the encoder is not designed for multi-document input. We will discuss more pre-trained language models tailored for MDS in Section 2.3.

Decoder-only Pre-trained Transformers The decoder-only Transformers are only based on the decoder of the Transformer architecture. The training objective is the same as that of recurrent neural language models. They are only trained to autoregressively generate text one token at a time by predicting the next token. These models has broad utility in NLP as most NLP tasks could be framed as a next word problem, e.g., for classification tasks we can predict the label as a token, and for conditional generation tasks such as machine translation the source sentence can be treated as prefix in the input. Such training scheme works very well for a large corpus as we do not have to add any annotated labels because we only need the natural sequence of words for supervision. In the training process, the weights in the Transformers are adjusted to minimize the average cross-entropy loss over the entire sequence. All the current large language models (LLMs) (discussed in Section 2.2.4) are all based on decoder-only pre-trained Transformers, such as Gemini 2.5 (Gemini Team, 2025), GPT-4.1 (OpenAI, 2025), Llama 4 (Llama Team, 2025), and Qwen2.5-1M (Yang et al., 2025). The training corpus of these pre-trained language models are mainly scraped from the web, e.g., C4 (Raffel et al., 2020), Pile (Gao et al., 2021), and Dolma (Soldaini et al., 2024), and the data is typically filtered for quality and safety to make the pre-trained language models have higher performance and safer.

Fine-tuning Pre-trained Transformers To make the pre-trained language models work on specific tasks or domains, these pre-trained models can be trained further on various downstream tasks including classification and generation tasks (Li et al., 2024a; Wang et al., 2023). For example, we could further train BART (Lewis et al., 2020) on machine translation by training it on specific machine translation data. The process of adapting pre-trained language models to new data is called fine-tuning, and it could be interpreted as a form of transfer learning. The pre-trained model has general linguistics and world knowledge while fine-tuning helps it learn domain-specific task knowledge. This is essentially to transfer the general knowledge learned from the pre-training stage to the more specific task-specific domain. Different from self-supervised pre-training, the objective of supervised fine-tuning is to generate the desired outputs for the specific tasks, e.g., text summarization. As such, this fine-tuning process requires labelled data for the task (e.g., for text summarisation we need the ground truth summary).

2.2.4 Large Language Models

By scaling the parameter size, the amount of pre-training corpus, or the amount of compute used for training, studies found that the decoder-only pre-trained Transformers continue to improve on task performance (Brown et al., 2020; Kaplan et al., 2020). Models with more parameters or trained on more data tend to have better performances than smaller models or models trained with less data. Models developed in these years tend to have at least several billions of parameters. For example, the Llama-3.1 series have various versions of open-source models including 8B, 70B, and 405B. They are developed with pre-training on large corpus including more than billions of tokens. These large models are named large language models (LLMs) in the research community.

Because LLMs are trained to predict next tokens, they can be prompted to do any text completion tasks (a.k.a prompting). For example, we could get an LLM to write a summary for an article by just prompting them with the additional term "tl;dr" (acronym for ‘too long; didn’t read’ which is used widely on internet forums and social media) following the texts of the article. We could also add any constraints in the prompt, e.g., the length of the generated summary.

Besides prompting, when the model gets larger in parameter size they demonstrate emerging capabilities of strong zero-shot and few-shot learning on various tasks with chain-of-thought prompting (CoT) (Wei et al., 2022) and in-context learning (ICL) (Brown et al., 2020; OpenAI, 2023).

In-context Learning In-context learning is one of the emergent capabilities of LLMs (Brown et al., 2020). We could prompt the model with some demonstration examples to show the model our expected input and output, and the model learns to follow the pattern behind the demonstrated examples. Intuitively, the prompting could be optimized with better quality of demonstration examples. It is called in-context learning as the model can learn to do a new task when it processes the prompt instead of learning via updating the parameters in pre-training and fine-tuning. For example, we could prompt the model with some examples with a specific format to make the model follow it although the model has never seen the format in training. We will not dive into the details of how to get better demonstration examples, as it is not the focus of the thesis.

Chain-of-Thought Prompting Chain-of-thought prompting (Wei et al., 2022) is another emergent capability of LLMs. We could just prompt the model with an additional instruction sentence, "think step by step", then the model could get better performance by generating plausible reasoning steps for each output. It improves the performance of LLMs on difficult reasoning tasks. Intuitively, CoT breaks difficult tasks into steps; similar to how humans solve complex problems. CoT can be combined with ICL and we could use demonstration examples which are composed of not only expected input and output, but also corresponding reasoning steps. This makes the model output similar reasoning steps to enhance their output quality.

Post-training and Model Alignment LLMs based on only the pre-training stage struggle to follow complex instructions though they have emergent capabilities. This is because these models are only trained to predict the next word in the pre-training stage. To improve their capabilities of following more complex textual instructions they are post-trained on diverse instruction data, a process also called instruction-tuning. Essentially, the models are further trained with instruction datasets which contains diverse textual tasks, e.g., Aya (Singh et al., 2024) and Flan (Longpre et al., 2023). In these instruction datasets, each sample is composed of an instruction and the expected output which is the response to the corresponding instruction. In addition, to make the models safe and aligned with the human need, they are further trained with the human preferences, which is called preference alignment (Christiano et al., 2017). The idea of preference alignment is to ensure that LLMs produce outputs that better match human needs, rather than simply predicting the most likely next word. The most widely used approach is based on reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022). In this process, the model is first equipped with a critic, often called a reward model, which is trained to evaluate how well a candidate response aligns with human preferences. To build this reward model, human annotators are typically asked to compare multiple responses generated by the base LLM and rank them from most to least preferred. The reward model then learns to predict these human judgments. Once the critic is trained, it acts as a surrogate for human feedback: instead of having humans evaluate every new response, the critic assigns a numerical reward to each response. This reward signal is then used within a reinforcement learning framework to further train the base LLM. Through this iterative process, the LLM gradually learns to produce responses that are not only grammatically correct or coherent, but also more aligned with human expectations. This method allows LLMs to scale human feedback efficiently and align more closely with human values without requiring direct human supervision for every generated response. There are many variants of RLHF proposed since then. We will not dive into details of the reinforcement learning algorithms as it is not the focus of the thesis.

2.3 Multi-document Modelling

How to model multiple documents is a long-standing problem in the area of MDS. Most approaches only focus on extractive summarization and directly use extracted text units as the final summary (Christensen et al., 2013a; Erkan and Radev, 2004; Lin and Hovy, 2002). These approaches only work on information extraction with salient text units, rather than information integration in a natural way similar to how humans do. Abstractive approaches aim to combine information extraction which is to get salient content and information integration which is to generate a human-like summary (Barzilay and McKeown, 2005; Li et al., 2020; Ma et al., 2020; Xiao et al., 2022). With the development of language modelling techniques from the n-gram model to large language models (LLMs) (Section 2.2), these language models can generate more and more fluent and coherent texts in these twenty years. MDS systems rely more on language models to generate plausible summaries (Ma et al., 2020). There are two approaches of using language models in the context of MDS. The first category is modelling the input documents as a long sequence by directly concatenating the documents and the model is trained to implicitly achieve information extraction and integration (Section 2.3.1). The other category is using explicit structures to represent the input documents and integrate the structural representations into the language generation process (Section 2.3.2).

2.3.1 Modelling as Long Sequences

The input documents can be simply concatenated into a long sequence, and be treated as a single document (Beltagy et al., 2020; Fabbri et al., 2019; Guo et al., 2022; Phang et al., 2022). The sequence-to-sequence language models are expected to implicitly learn to extract salient information and generate the final summary based on training on large-scale labelled data. This is difficult for traditional language models as a document collection may contain a number of documents and each of which may be long on its own and these models cannot handle it. Research in this direction only rises after the Transformer architecture (Beltagy et al., 2020). While general-purpose language models such as BART (Lewis et al., 2020) and T5 (Raffel et al., 2020) can be used for MDS, they can only handle 512 tokens in the input because of limitations of the computing infrastructure and model efficiency. As such, studies explore various methods to model long sequences using these language models. Modelling multi-document inputs as long sequences, the research focus is on developing long-context language model architectures to handle multiple documents (Beltagy et al., 2020; Guo et al., 2022; Phang et al., 2022), or conduct task-specific optimization based on large-scale data (Xiao et al., 2022). However, the generation process of these models is opaque because of the black-box nature of the sequence-to-sequence models and the outputs are not fully grounded to the input documents.

Long-Context Language Models To address the limitations of original Transformers which are constrained by their quadratic complexity in handling long input sequences (Vaswani et al., 2017), various studies have explored ways to expand the context length of Transformer-based language models. The sliding window technique with sparse attention is used to achieve efficient context scaling (Beltagy et al., 2020; Ratner et al., 2023; Sun et al., 2023), which reduces the calculation effort cost in the quadratic attentions. In addition, to extend the context length of LLMs a new position embedding strategy based on the rotary position encoding

(RoPE) (Su et al., 2024) is recently adopted (Peng et al., 2024; Zhu et al., 2024). However, the sparsity of the attention makes the models fail to capture the entire input sequences as some tokens in the sequence are ignored in the attention calculation. Besides the improvement on model architectures, advancements in efficient mechanisms (e.g., FlashAttention (Dao, 2024; Dao et al., 2022)) and parallelism techniques (e.g., ZeRo (Rajbhandari et al., 2020)) help reduce computational overhead. Benefiting from these techniques, Transformer-based language models can handle longer and longer inputs. Proprietary language models such as GPT-4.1 (OpenAI, 2023) and Gemini 2.5 Pro (Gemini Team, 2025) can even handle up to 1 million tokens.

Task-Specific Model Optimization To optimize language models for MDS, various fine-tuning objectives for language models are developed. For example, PRIMERA (Xiao et al., 2022) with the same model architecture as LED (Beltagy et al., 2020) is developed particularly for MDS. It is fine-tuned to generate pseudo summaries from the input documents. The pseudo summary is composed of text spans automatically extracted based on the entity salience from multiple documents. PRIMERA achieves better performances in terms of ROUGE (Lin and Hovy, 2003) on Multi-News, Multi-XScience than previous models (Zhang et al., 2020a). There are also approaches based on traditional sequence-to-sequence models where the encoder and decoder are recurrent neural networks. For example, to summarize multiple opinionated documents, MeanSum (Chu and Liu, 2019) is the first fully abstractive model for opinion summarization with unsupervised learning. The model is trained by reconstructing input documents with the objective to make the embedding of the generated summary similar to the average embedding of input documents. However, this model cannot handle long inputs and generations of this model have limited fluency and coherence.

2.3.2 Modelling with Explicit Structures

To effectively model the input documents for MDS, researchers explore to explicitly model the structural information to learn better representations of the multi-document inputs (Hosking et al., 2024; Liu and Lapata, 2019). Various studies model the multi-document inputs with linguistic structures (Liu and Lapata, 2019), content-based hierarchies (Hosking et al., 2024) or graph-base representations with various nodes and edges (Cui and Hu, 2021; Li et al., 2023b; Zhao et al., 2020). However, most of these models have an opaque generation process due to the black-box nature of sequence-to-sequence architectures, and their outputs are often not fully grounded in the input documents, leading to hallucinations.

Linguistic Structures Based on linguistic structures, Barzilay and McKeown (2005) propose a symbolic approach. They first identify fragments conveying common information by aligning constituents in the syntactic trees of the sentences, then combine the fragments with the transforming parsed tree into a fusion lattice, and finally uses a language model to generate texts from the lattice representation. However, this model has limited language capability and the input is short. To get the Transformer-based language models handle multi-document inputs, Liu and Lapata (2019) propose a hierarchical Transformer-based architecture following the discourse structure to capture cross-document relationships. Specifically, input documents are firstly split into paragraphs. The hierarchical Transformer encoder are then use to encode the top paragraphs to generate the summary with the Transformer decoder. The hierarchical Transformer encoder is based on the proposed inter-paragraph attention. After obtaining paragraph representations with a vanilla Transformer, the inter-paragraph attention mechanism is used to model the dependencies across multiple paragraphs. This model gets better performance than the original Transformer architecture on WikiSum in terms of human ratings. While the hierarchical Transformer architecture represent the input documents with paragraphs, MGSum (Jin et al., 2020) proposes a multi-granularity encoder based on semantic units of document collections, including documents, sentences, and words. With the multi-granularity interaction network, sentence selection and summary generation are trained in a unified architecture and it promotes capturing salient information of input documents. This model works much better than the hierarchical Transformer architecture based only on paragraphs on Multi-News in terms of ROUGE (Lin and Hovy, 2003).

Content-based Hierarchies The abstractive summarization can be achieved in a hybrid way by first extracting text units and then generating the summary based on extracted texts.

This pipeline may follow different hierarchical flows to get the final summary. For example, T-DMCA (Liu et al., 2018) first extracts a subset of paragraphs from input documents and concatenate the extracted paragraphs as the input to generate the summary with a decoder-only Transformer architecture. To make the Transformer decoder handle longer inputs, they modify the multi-head attention to reduce memory usage by limiting the dot products in the attention network. This work found that the decoder-only Transformer architecture work better than the encoder-decoder architecture and it shows promises to be extended to long sequences. To get LLMs summarize a large collection of reviews for opinion summarization (i.e., summarizing multiple opinionated documents), TCG (Bhaskar et al., 2023a) is also a pipeline approach with the hierarchical flow following topics rather than the surface form of texts. Sentences in the input reviews are first clustered based on the aspects closest to their topic; each cluster is then repeatedly chunked and summarized by GPT-3.5 (Brown et al., 2020) until the combined length falls below 35 sentences; the last step is using GPT-3.5 again to summarize sub-summaries from all clusters. This approach present a way of using GPT-3.5 with a limited context window size to summarize a large collection of reviews and it shows better performance on strong baselines on SPACE (Angelidis et al., 2021). HIRO (Hosking et al., 2024) also follows the extraction then abstractive generation approach. It first constructs hierarchical indexing by mapping sentence from reviews to a hierarchical discrete latent space, then identify sentence clusters that contain popular opinions, and generate the final summary by prompting an LLM with the central sentences in the identified clusters. This modular approach improves the controllability, flexibility, and interpretability of opinion summarization. These pipeline approaches make the summarization process more transparent and the generated summary more grounded. However, these approaches are not entirely transparent because the intermediate results cannot easily validated. These approaches extract text units in sentences or paragraphs, but it may be more promising to split input texts with flexible semantic boundaries.

Graph-based Representations Graphs can be used to model the input documents to better capture cross-document relationships among the input documents (Li et al., 2020; Li and

Model	Nodes	Edges
GraphSum (Li et al., 2020)	paragraphs	tf-idf similarity, topic distribution similarity, or discourse relations between paragraphs
SummPip (Zhao et al., 2020)	sentences	discourse relation and representation similarity between sentences
SLN (Li and Zhuge, 2021)	concepts, events	semantic links between concepts or events
TG-MultiSum (Cui and Hu, 2021)	words, topics, documents	topics contains words and documents contains topics
AF19 (Fan et al., 2019)	entities	knowledge graph relations
EMSum (Zhou et al., 2021)	entity clusters, paragraphs	paragraphs contains entities from each entity cluster

Table 2.2 Graphs with different nodes or edges to represent input documents.

Zhuge, 2021; Shah et al., 2021; Wang et al., 2020; Zhao et al., 2020). There is much effort exploring to integrate graph representations with language models to make the model develop better understanding of multi-document relationships. As language models are designed to process flat language sequences (Guo et al., 2022; Lewis et al., 2020; Phang et al., 2022; Zhang et al., 2020a), it is non-trivial to incorporate graphs into the language generation process. The developed models shown in Table 2.2 leverage various graphs to represent the input documents. These graphs are constructed based on discourse elements (Li et al., 2020; Zhao et al., 2020), topic modelling (Cui and Hu, 2021), or knowledge graphs (Fan et al., 2019; Zhou et al., 2021). GraphSum Li et al. (2020) constructs a similarity graph over the paragraphs. Edges of this paragraph graph are tf-idf similarities, topic distribution similarities, or discourse relations between paragraphs. The paragraphs from the input documents are first encoded by the Transformer layer to get paragraph representations. A graph-informed attention mechanism is then proposed to incorporate explicit graph representations into the encoding process. The graph-informed attention mechanism learns better inter-paragraph relationships by considering the explicit graphs. This model shows better performances than the normal Transformer and the hierarchical Transformer architecture in Liu and Lapata (2019). The discourse elements based graphs are also widely used in extractive summarization. SummPip (Zhao et al., 2020) represents input documents as a sentence graph. Edges are constructed based on approximate discourse graph (ADG) (Christensen et al., 2013b). Spectral clustering is then conducted on the sentence graph to get sentence clusters. The concatenation of the representative sentences from the clusters is the final extractive summary. Instead of discourse elements, Li and Zhuge (2021) represent the input documents with the proposed Semantic Link Network (SLN), which is a graph composed of concepts and events as nodes and their relations as edges. Concepts are noun phrases extracted from input documents based on dependency parsing, and their relations are extracted based on the phrases within the text between concepts. Events are extracted based on structured event information from text by pre-defined event schema, and their relations are inferred by sentence structures and discourse features in input documents. The summary is generated based on a compressed SLN with integer linear programming (ILP). TG-MultiSum (Cui and Hu, 2021) uses topics as bridge among different input documents and represents the input documents as a graph that comprises nodes of different granularity: word, topic, and document. The topics of input documents are predicted by a neural topic model, and each input document can be represented as a mixture of topic distributions. Edges in the graph consist of containing relationships (i.e., document-topic and topic-word). An adapted graph attention network (GAT) is then used to encode the graph to learn relationships among different units. To model the cross-document relationships, TG-MultiSum incorporates topic modelling into both the encoding and decoding phase of an encoder-decoder framework for MDS. The model generates summaries with a graph-to-sequence process. The decoder first decodes the topic and then generate summaries. There are also MDS approaches representing the input documents with knowledge graphs. For example, AF19 (Fan et al., 2019) represents the input documents as linearized knowledge graphs. EMSum (Zhou et al., 2021) encodes the cross-document relationships with the help of entities in documents and graph attention networks (GATs) (Velickovic et al., 2018) are used to learn the representations. However, although the constructed graphs in these approaches contains different types of nodes or edges, they are in fact homogeneous graphs in that the edges are modelled in a single type. It is interesting to explore using heterogeneous graph modelling which may have better expressiveness of complex relationships among the input documents which we explore in Chapter 3.

2.4 Generated Summary Evaluation

Quality evaluation for the machine-generated summary, $\hat{z}$, given the input text, $\mathscr{S}$, and the human-written summary (i.e., the ground truth), r, relies on automatic evaluation metrics. Because there are few automatic metrics specifically designed for MDS, the community typically uses metrics for generic language generation or single-document summarization (Gehrmann et al., 2023b). These evaluation metrics are based on text similarities in different representations (Lin and Hovy, 2003; Zhang et al., 2020b) (Section 2.4.1), fine-tuning pre-trained language models with human evaluation related data (Raffel et al., 2020; Zha et al., 2023; Zhong et al., 2022) (Section 2.4.2), or prompting large language models (LLMs) (Liu et al., 2023b) (Section 2.4.3). In addition, there are metrics measuring fine-grained quality of the generated summaries including fluency, coherence, consistency and relevance, not just the overall quality (Deng et al., 2021; Yuan et al., 2021; Zhong et al., 2022). The consistency and relevance are identified as two key aspects to characterize the content quality of generated summaries, while the fluency and coherence are focused on the language readability of the generated summaries. The consistency (a.k.a. factuality) dictates that the generated summary should only contain information consistent with the input documents. The relevance concerns how well the generated summary retains important information in the source documents.

2.4.1 Similarity-based Metrics

Similarity-based metrics mainly focus on measuring representation distances between the machine-generated and human-written summaries. There are different approaches to calculate similarities, including surface-form matching and embedding-based similarities.

ROUGE (Lin and Hovy, 2003), the most representative evaluation metric, simply computes the recall based on the lexical overlapping between the machine-generated summary and the human-written summary in different granularities. The lexical overlap is based on n-grams in the pair of compared summaries. If there are multiple human-written summaries for each instance, ROUGE can be adapted to calculation based on the union of n-grams in all ground truth summaries. It has been a widely used automatic evaluation metric for text summarization. However, ROUGE cannot capture semantic similarity of the machine-generated summaries as it only relies on the surface form of texts.

While surface-form matching based metrics cannot measure semantic similarity (Gehrmann et al., 2023b), similarity could be calculated based on contextual representations from off-the-shelf pre-trained language models to capture semantics. For example, BERTScore (Zhang et al., 2020b) calculates the overall quality score as a f-measure based on greedy similarity matching between contextual embeddings of the machine-generated summary words and the ground truth summary words from BERT (Reimers and Gurevych, 2019a). Intuitively, it measures the word-level similarity between the generated summary and the ground truth summary. In contrast, SUPERT (Gao et al., 2020) is an evaluation metric which does not require human-written summaries for MDS evaluation. It calculates Word Mover's Distance between the machine-generated summary and a pseudo summary to measure the relevance of the generated summary to the input. The pseudo summary is composed of top sentences and selected sentences by graph-based sentence extraction. Specifically, a graph is build to represent input documents and sentences are then clustered by a clustering algorithm, and the central sentence in each cluster is selected to build the pseudo summary. As an unsupervised metric, SUPERT has a good correlation results with human judgements. CTC (Deng et al., 2021) models evaluation as an information alignment task. Its variants CTC-E and CTC-D with different alignment functions are found to correlate well with human judgements on relevance and consistency, respectively. The consistency score is calculated as the alignment of the machine-generated summary to the input text, and the relevance score is calculated as the alignment of the machine-generated summary to both the ground truth summary and the input text. The alignment of the machine-generated summary to any text is calculated by averaging the word-level alignment scores. CTC-E calculates the word-level alignment score based on greedy similarity matching. CTC-E has much higher correlation with human judgements than BERTScore and SUPERT in relevance on SummEval (Fabbri et al., 2021). CTC-D has a different implementation of the word-level alignment score, and it will be introduced in next section. Although these metrics use contextual embeddings to capture semantic similarity in evaluation, they only consider one or two evaluation aspects and do not consider fluency or coherence. Different from BERTScore, SUPERT, and CTC-E, BARTScore (Yuan et al., 2021) evaluate the summary quality based on the likelihood of the machine-generated summary conditioned on the ground truth summary or the input text with BART (Lewis et al., 2020). The likelihood of the machine-generated summary conditioned on the input text is found empirically correlated well the quality aspects of coherence, fluency, and consistency. The likelihood conditioned on the human-written summary is used to calculate the score for relevance. Experiments on SummEval (Fabbri et al., 2021) show that BARTScore outperforms prior evaluation metrics, especially BERTScore.

2.4.2 Learning-based Metrics

Although similarity-based evaluation metrics have led to significant improvements in assessing fine-grained quality aspects, their correlation with human judgments particularly for content-related quality remains insufficient. To make evaluation metrics correlate better with human judgments, various studies build the evaluators by optimizing pre-trained language models with human judgements related labelled data. The evaluators are expected to directly learn human judgements. CTC-D (Deng et al., 2021) (i.e., a variant of CTC) trains a sequence tagging model to predict the probability of each word in the machine-generated summary to be aligned with the input text or the ground truth summary as the word-level alignment score, rather than based on the greedy similarity matching in CTC-E. The final scores are calculated in the same way of CTC-E (Section 2.4.1) but with a different alignment function. Instead of an alignment task, UniEval (Zhong et al., 2022) frames evaluation of machine-generated summaries as boolean question answering on the four quality aspects (e.g., is this a coherent summary to the document?), and train a T5 (Raffel et al., 2020) to predict the answer of

"Yes" or "No". To train the unified model, Zhong et al. (2022) use synthesized data for different quality aspects in training, and use data from other tasks like natural language inference (NLI) (Yin et al., 2021). To calculate scores with the trained model, they get the probability of the answer of "Yes" as the score for the corresponding aspects. UniEval gets much higher performance than all other metrics such as BERTScore, CTC, BARTScore on SummEval (Fabbri et al., 2021). Different from UniEval which only formulates evaluation as a discriminative task or BARTScore which only formulates evaluation as a generative task, T5Score (Qin et al., 2022) combines generative modelling and discriminative modelling to train the evaluator. The evaluator is based on mT5 (Xue et al., 2021) and it is trained to not only generate human-written summaries but also distinguish different quality of summaries with contrastive learning. They use a f-measure as the overall quality score of the generated summary. The precision is calculated as the probability of the generated summary conditioned on the input text and the recall is calculated as the probability of the human-written summary conditioned on the machine-generated summary. T5Score has marginally better correlation with human judgement than BARTScore on MultiSumm. To further improve the evaluation accuracy, there are also evaluation metrics which only focus on factuality or consistency of the machine-generated summary, such as SummaC (Laban et al., 2022), and AlignScore (Zha et al., 2023). SummaC (Laban et al., 2022) is based on sentence-level natural language inference (NLI). They first construct a NLI matrix based on the entailment probabilities between sentences from the input text and the machine-generated summary. The final consistency score is then calculated in two ways. One is based on greedy pair matching to the average of strongest support for each sentence in the machine-generated summary. The other one is to learn a convolution neural network to predict the score with synthetic data from FactCC (Kryscinski et al., 2020) which is a metric to predict sentence consistency. While SummaC has an improved correlation with human judgements on SummEval (Fabbri et al., 2021), evaluating the machine-generated summary against individual sentences in the input text can degrade metric performance as paragraph- and document-level semantic information is lost. Similar to CTC, AlignScore (Zha et al., 2023) also models consistency evaluation as information alignment but with a different alignment calculation.

To train the alignment model, they integrate and transform a large diversity of data sources from well-established tasks including NLI, question answering, paraphrasing, fact-checking, information retrieval, semantic similarity, and summarization into alignment data, and use the data to train RoBERTa (Liu et al., 2019) as the alignment model to predict the alignment for various tasks. Instead of directly using the alignment probability as the consistency score, they split the input text into chunks because of the limited context window size of RoBERTa and the machine-generated summary into sentences to get fine-grained evaluation. The final score is also based on greedy pair matching. They first get the alignment probability for each sentence in the machine-generated summary from the chunk that most strongly supports it, and then use the average value of all highest alignment probabilities as the factual consistency score. AlignScore shows better performance than all the other metrics and it is the state-of-the-art metric for consistency evaluation. It is a evaluation metric that this thesis relies on.

2.4.3 Prompting-based Metrics

As large language models (LLMs) have captured extensive world knowledge and prompting them has yielded strong performance across a variety of language tasks (Section 2.2.4) , some evaluation metrics for text summarization are directly based on prompting LLMs. These language models show the potential to not only generate high-quality summaries and also predict quality scores of any machine-generated summaries.

GPTScore (Fu et al., 2024) uses the likelihood of the machine-generated summary conditioned on the input text and aspect description from LLMs as the evaluation score for the corresponding quality aspect. Although it is similar to BARTScore of using likelihood as the score, GPTScore is based on in-context learning of LLMs and it is more controllable for fine-grained evaluation on more quality aspects. Specifically, different prompts are used for different evaluation targets, such as "does the generated text preserve the factual statements of the source text?" for the consistency of a summary to the input text. It shows a better performance on SummEval than ROUGE, BARTScore and BERTScore.

In contrast, G-Eval (Liu et al., 2023b) evaluates summaries with explicitly assigning evaluation scores instead of using the likelihood. The prompt is more complex than that in GPTScore, and it is based chain-of-thoughts (CoT) (Wei et al., 2022) (Section 2.2.4) and in-context learning (Brown et al., 2020) (Section 2.2.4). In the prompt, there are explicit definition of the evaluation aspect, its evaluation criteria or the score range for the aspect. The model needs to assign a score to the evaluated summary like human annotators. To get more accurate scores, G-Eval samples 20 scores for each instance and calculate the final score for the aspect as the mean of the valid scores. However, G-Eval is only validated on news summarization and dialogue generation datasets . It is still unclear how it performs for other summarization domains, such as opinion summarization. This is the reason why we design our own metrics based on LLMs in both Chapter 4 and Chapter 5 to evaluate the generated summaries in opinion summarization.

Chapter 3

Ideational Information Integration

Ideational documents are mainly composed of objective and factual texts. For example, (most) news articles and scientific publications are ideational documents, as they contain objective statements and report on facts.[1] To accurately understand what truly happens related to some interests (e.g., a specific news event and scientific development for a research topic), humans have to read all the related documents to integrate dispersed information from different sources. This is challenging because we struggle with limited reading speeds and brain capacity. To help readers digest the extensive ideational information, an increased number of computational models to automatically summarize the documents have been developed with the aim of delivering a short version of the text describing the salient facts of the source documents.[2]

The chapter addresses the first research question of the thesis: *how to integrate ideational information from multiple documents to generate better summaries*. The content of this chapter is based on the following publication.

- Miao Li, Jianzhong Qi, and Jey Han Lau. Compressed heterogeneous graph for abstractive multi-document summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 11, pp. 13085-13093. 2023.

[1] https://edition.cnn.com/

[2] Please note that there may not be only ideational information in news articles or scientific papers, while we only focus on ideational documents in this section. We will talk about a limitation related to this in the reflections in Chapter 3.2.

3.1 Heterogeneous Graph Compression

To generate summaries by integrating dispersed information from different ideational source documents, models need to understand the relationships among the sources. A cluster of ideational documents such as news articles are connected with underlying facts and they may have complex cross-document relationships. For example, among the input documents there may be contradiction, redundancy, and complementary information (Radev, 2000) and the summary needs to consider temporal relationships among the source documents. These complex cross-document relationships makes it non-trivial to produce an summary with accurate understanding of the input documents.

Researchers have started working on the problem since more than thirty years ago (Radev, 2000). As introduced in 2.3, pre-trained language models (PLMs) play an important role in language modelling in recent years and encoder-decoder PLMs are the backbone of the most performant text summarization models (Beltagy et al., 2020; Xiao et al., 2022; Zhang et al., 2020a). Most summarization models rely on fine-tuning general-purpose PLMs for text generation (Beltagy et al., 2020; Lewis et al., 2020; Raffel et al., 2020) or PLMs tailored for text summarization (Zhang et al., 2020a). To handle multi-document inputs, these models concatenate multiple documents into a flat sequence and are expected to somehow learn MDS. However, these models produce plausible but unfaithful summaries (Xiao et al., 2022).

To incorporate cross-document relationships into the summarization process to improve the quality of generated summaries, in this chapter we propose to represent source documents with heterogeneous graphs and integrate it to an encoder-decoder PLM. Although Cui and Hu (2021); Jin et al. (2020); Li et al. (2020); Li and Zhuge (2021) have used graphs to represent source documents for MDS, their constructed graphs are homogeneous (i.e., nodes or edges in the graphs are of a single type) even though heterogeneous graphs have greater expression capability on underlying relationships. Our heterogeneous graph is composed of different types of nodes and edges. Specifically, we use words, sentences and documents as the three levels of nodes. At the word level, we have two types of weighted edges, including cosine similarities of embeddings between any two noun words and between any two adjacent words. Edges between sentences are weighted based on the cosine similarity of their sentence embeddings from SentenceBERT (Reimers and Gurevych, 2019b). We also have edges to represent compositional relationships among words, sentences, and documents. Specifically, the words will be connected to the sentences they belong to and the sentences will be connected to the document they belong to (with the edge weight as 1 if there is connection otherwise 0). The model is expected to learn multi-document relationships based on the representation of heterogeneous graphs

To integrate the graphs with PLMs, we borrow the idea of heterogeneous graph compression which is to compress a heterogeneous graph into a smaller graph with only important nodes and edges. Intuitively, our PLMs are not only trained to maximize the probabilities of the ground truth summaries but also make the compressed graph similar to the graph constructed based on the ground truth summary. The model is trained with multi-task objectives in an end-to-end manner: the cross entropy based loss between the generated summary and human-written reference summary for text generation, and embedding distance based loss between the compressed graph and the original heterogeneous graph. In the encoding process, the heterogeneous graphs are constructed based on the original PLM encoder and a graph encoder based on the proposed multi-channel graph attention networks that considers the different types of edges in graph neural networks is used to compress the heterogeneous graph into a smaller graph. In the decoding process, taking the compressed graph as input the original PLM decoder generates a summary that capture the salient information from the compressed graph.

Compressed Heterogeneous Graph for Abstractive Multi-Document Summarization

Miao Li, Jianzhong Qi, Jey Han Lau
School of Computing and Information Systems,
The University of Melbourne
miao4@student.unimelb.edu.au, {jianzhong.qi, laujh}@unimelb.edu.au

Abstract

Multi-document summarization (MDS) aims to generate a summary for a number of related documents. We propose HGSUM — an MDS model that extends an encoder-decoder architecture to incorporate a *heterogeneous* graph to represent different semantic units (e.g., words and sentences) of the documents. This contrasts with existing MDS models which do not consider different edge types of graphs and as such do not capture the diversity of relationships in the documents. To preserve only key information and relationships of the documents in the heterogeneous graph, HGSUM uses graph pooling to compress the input graph. And to guide HGSUM to learn the compression, we introduce an additional objective that maximizes the similarity between the compressed graph and the graph constructed from the ground-truth summary during training. HGSUM is trained end-to-end with the graph similarity and standard cross-entropy objectives. Experimental results over MULTI-NEWS, WCEP-100, and ARXIV show that HGSUM outperforms state-of-the-art MDS models. The code for our model and experiments is available at: https://github.com/oaimli/HGSum.

Introduction

Multi-document summarization (MDS) aims to automatically generate a concise and informative summary for a cluster of topically related source documents (Ma et al. 2020; Radev, Hovy, and McKeown 2002). It has a wide range of applications such as creating news digests (Fabbri et al. 2019), product review summaries (Gerani et al. 2014), and summaries for scientific literature (Moro et al. 2022; Otmakhova et al. 2022). Our work targets *abstractive* MDS, which generates summaries with words that do not necessarily come from the source documents, resembling the summarization process of human beings.

State-of-the-art text summarization models use *pretrained language models* (PLMs) including both general-purpose PLMs for text generation (Beltagy, Peters, and Cohan 2020; Lewis et al. 2020) and PLMs designed for text summarization (Zhang et al. 2020a; Xiao et al. 2022). When applied to the abstractive MDS task, these models take a flat concatenation of the (multiple) source documents, which may not capture cross-document relationships such as contradiction, redundancy, or complementary information very well (Radev 2000). Ma et al. (2020) argue that explicit modeling of cross-document relationships can potentially improve the quality of summaries. Following this, several recent studies (Li et al. 2020; Jin, Wang, and Wan 2020; Cui and Hu 2021) explore graphs to model source documents to improve abstractive MDS. However, these graphs are *homogeneous* in that the nodes or edges are not distinguished for different semantic units (e.g., words, sentences, and paragraphs) in the encoding process. This means these MDS models cannot capture the diverse cross-document relationships among different types of semantic units.

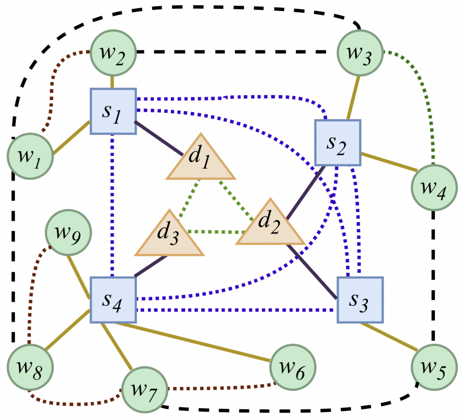


Figure 1: The structure of the heterogeneous graph given three documents in a document cluster: The orange triangles denote document nodes d, the blue quadrates denote sentence nodes s, the green circles denote word nodes w, and the line (or curve) segments between nodes denote edges. A detailed description of the graph is in the Preliminaries.

In this paper, we propose HGSUM — an MDS model that extends an encoder-decoder architecture to incorporate a heterogeneous graph to better capture the interaction between different semantic units in the documents. HGSUM's heterogeneous graph has different types of nodes and edges to model words, sentences, and documents, as shown in Figure 1. To facilitate HGSUM to learn cross-document relationships, we construct edges between sentences *across documents* based on the similarity of their sentence embeddings. We also explore compressing the graph with graph pooling to preserve only salient information (i.e., nodes and edges) that is helpful for summarization, before feeding signals from the compressed graph to the text decoder to generate the final summary. To guide HGSUM to learn this compression, we introduce an auxiliary objective that maximizes the similarity between the compressed graph and the graph derived from the ground-truth summary, in addition to the standard cross-entropy objective during training.

There are several challenges that we face. First, it is non-trivial to encode heterogeneous graphs with existing graph neural networks, as different types of nodes and edges should not be processed by the same function. To address this challenge, we propose multi-channel graph attention networks to encode heterogeneous graphs. Second, there are few graph compression or pooling methods proposed for heterogeneous graphs. Inspired by Lee, Lee, and Kang (2019), we introduce a compression method based on self-attentions to condense the heterogeneous graph. One novelty of our method is that it uses *soft masking* so that it does not break the differentiability of the network, allowing us to train HGSUM in an end-to-end manner.

To summarize, our contributions are given as follows:

- We propose HGSUM, an MDS model that extends the encoder-decoder architecture to incorporate a compressed graph to model the input documents. The graph is a heterogeneous graph that captures the diversity of semantic relationships in the documents, and it is compressed with a pooling method that helps preserve the most salient information for summarization.
- HGSUM is trained with two objectives that maximize the likelihood of generating the ground-truth summary and the similarity between the compressed graph and the graph constructed from the ground-truth summary.
- Experimental results over multiple datasets show that HGSUM outperforms state-of-the-art MDS models.

Preliminaries

Given a set of m related source documents $\mathcal{D} = \{d_0, d_1, \dots, d_m\}$ (i.e., a document cluster), our aim is to generate a text summary $\hat{z} = \hat{w_0}, \hat{w_1}, \dots, \hat{w_T}$ (composed of T words) that captures the essence of the source documents. As mentioned earlier, we generate the summary in an abstractive fashion, i.e., words in the generated summary can be words that are not found in the source documents. The generation of each word in the summary is modeled as:

$$\mathrm{p}(\hat{z}|\mathcal{D}) = \prod_{i=0}^{T} \mathrm{p}(\hat{w_i}|\mathcal{D}, \hat{w_0}, \hat{w_1}, \dots, \hat{w_{i-1}}) \tag{1}$$

As heterogeneous graphs explicitly represent relationships among different semantic units (documents, sentences, and words), we construct a heterogeneous graph to represent a cluster of documents. We next explain how we construct the heterogeneous graph.

Heterogeneous Graph Construction

We denote the heterogeneous graph constructed to represent a cluster of documents as $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$, where $\mathcal{V}$ represents the set of nodes in the graph, and $\mathcal{E}$ the set of edges. As the example in Figure 1 shows, there are three types of nodes and six types of edges in $\mathcal{G}$. Specifically, $\mathcal{V} = \mathcal{V}_d \cup \mathcal{V}_s \cup \mathcal{V}_w$, where $\mathcal{V}_d$ is a set of document nodes: every document in the cluster corresponds to a node in $\mathcal{V}_d$ (orange triangles in Figure 1); $\mathcal{V}_s$ is a set of sentence nodes: every sentence in the documents corresponds to a node in $\mathcal{V}_s$ (blue quadrates in Figure 1); and $\mathcal{V}_t$ is a set of word nodes[1]: every word in the sentences corresponds to a node in $\mathcal{V}_w$ (green circles in Figure 1).

We next define the edges, which are all undirected:

- The sets $\mathcal{E}_{we}$ and $\mathcal{E}_{wo}$ contain edges between word nodes (dash and dot lines between word nodes in Figure 1). Every edge in $\mathcal{E}_{we}$ connects two nodes corresponding to noun words (identified based on a dependency parser[2]).[3] The weight of an edge for a word pair in $\mathcal{E}_{we}$ is the cosine similarity of their embeddings. We use GloVe (Pennington, Socher, and Manning 2014) as the static word embeddings in this work. Edges in $\mathcal{E}_{wo}$, on the other hand, connect the nodes corresponding to every adjacent word pairs in a sentence. All edges in $\mathcal{E}_{wo}$ have a weight of 1.0.
- The set $\mathcal{E}_{ss}$ contains edges that connect every pair of sentences (dot lines between sentence nodes in Figure 1). The weight of an edge for a pair of sentences is the cosine similarity of their pre-trained sentence embeddings. We use Sentence-BERT (Reimers and Gurevych 2019) to compute the sentence embeddings, which is pre-trained based on the natural language inference task (Bowman et al. 2015).
- The set $\mathcal{E}_{dd}$ contains edges between document nodes (dot lines between document nodes in Figure 1). Every document is connected to all other documents in the cluster, and their edges are weighted using their n-gram overlap in terms of the average F1 value of ROUGE-1, ROUGE-2, and ROUGE-L (Lin and Hovy 2003).
- The sets $\mathcal{E}_{ds}$ and $\mathcal{E}_{st}$ contain edges that connect a document with its sentences (solid lines between document nodes and sentence nodes in Figure 1) and edges that connect a sentence with its words (solid lines between sentence nodes and words nodes in Figure 1). These edges are designed to preserve the hierarchical document-sentence and sentence-word structures. All the edge weights in these sets are set to 1.0.

To summarize, we have $\mathcal{E} = \mathcal{E}_{we} \cup \mathcal{E}_{wo} \cup \mathcal{E}_{ss} \cup \mathcal{E}_{dd} \cup \mathcal{E}_{ds} \cup \mathcal{E}_{sw}$. These edges collectively create a connected graph over all three types of nodes (words, sentences, and documents). Note that the choice of pre-trained word/sentence embeddings is flexible in our architecture, and in future work it would be interesting to explore other pre-trained embeddings.

[1] Technically, these are *subword* nodes since we use subword tokenization, although most nodes map to full words in practice.

[2] https://spacy.io/

[3] Note that nodes that do not map to a full word will not have this type of edge, since they cannot be a noun.

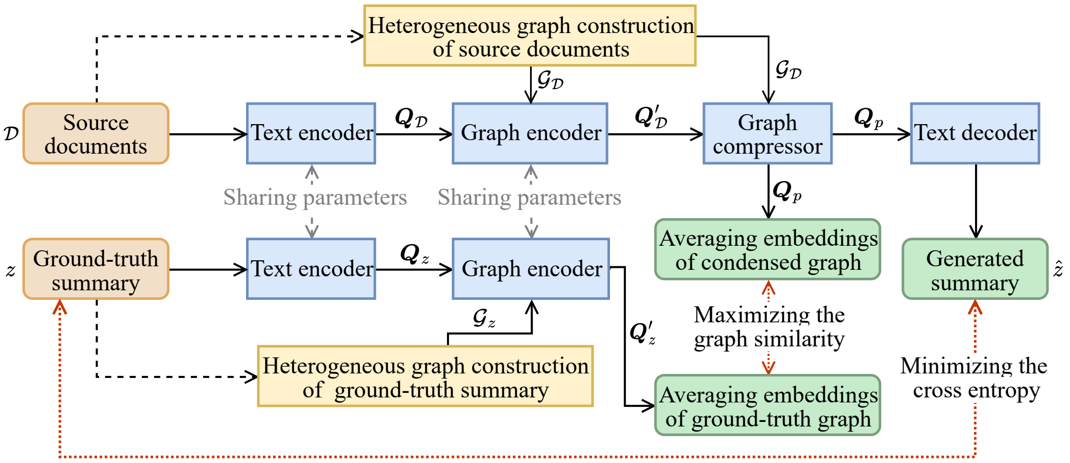

Heterogeneous graph construction of source documents
$\mathcal{G}_{\mathcal{D}}$
$\mathcal{D}$
Source documents
Text encoder
$Q_{\mathcal{D}}$
Graph encoder
$Q'_{\mathcal{D}}$
Graph compressor
Q_p
Text decoder
Sharing parameters
Sharing parameters
Averaging embeddings of condensed graph
Generated summary
$\hat{z}$
z
Ground-truth summary
Text encoder
Q_z
Graph encoder
$\mathcal{G}_z$
Q'_z
Maximizing the graph similarity
Minimizing the cross entropy
Heterogeneous graph construction of ground-truth summary
Averaging embeddings of ground-truth graph

Figure 2: The HGSUM architecture: There are four main components: (1) text encoder (initialised using PRIMERA weights); (2) graph encoder; (3) graph compressor; and (4) text decoder (initialised using PRIMERA weights).

The HGSUM Model

At its core, HGSUM extends a text encoder-decoder architecture (PRIMERA; Xiao et al. (2022)) to incorporate information from a compressed heterogeneous graph derived from the input source documents, as presented in Figure 2. HGSUM has four main components: (1) text encoder (initialized using PRIMERA weights), (2) graph encoder, (3) graph compressor, and (4) text decoder (initialized using PRIMERA weights).

During training, we first generate two heterogeneous graphs $\mathcal{G}_{\mathcal{D}}$ and $\mathcal{G}_z$ based on the input source documents $\mathcal{D}$ and the ground-truth summary z, respectively, following the graph construction procedure described in the previous section. The text of input source documents $\mathcal{D}$ and ground-truth summary z is processed by the text encoder to obtain contextual word embeddings $Q_{\mathcal{D}}$ and Q_z, respectively. These contextual word embeddings are then used by the graph encoder as the initial node embeddings of $\mathcal{G}_{\mathcal{D}}$ and $\mathcal{G}_z$, respectively. After processed by the graph encoder, we have the graph encodings $Q'_{\mathcal{D}}$ and Q'_z respectively for the source documents and the ground-truth summary.[4] The graph encoding of the source documents ($Q'_{\mathcal{D}}$) will be further processed by the graph compressor to produce compressed graph encoding Q_p, and this will be used by the text decoder to generate the final summary $\hat{z}$. To train HGSUM, we minimize the cross entropy between the ground-truth summary z and generated summary $\hat{z}$ and maximize the similarity between the compressed graph encoding (Q_p) and ground-truth summary graph encoding (Q'_z).

Once the model is trained, we only use the text and graph encoders to encode the input source documents, the graph compressor to compress the document graph, and the text decoder to decode the summary, without using any ground-truth summary as input. We next detail these components.

Text Encoder

The text encoder follows the encoder architecture of PRIMERA — which uses the sparse attention of longformer (Beltagy, Peters, and Cohan 2020) to accommodate long text input — and is initialized with PRIMERA weights:

$$Q_{\mathcal{D}} = \mathrm{longformer}(\mathcal{D}) \tag{2}$$
$$Q_z = \mathrm{longformer}(z) \tag{3}$$

The text encoder takes as input a concatenated string containing all the words from the documents, and it produces contextualized embeddings for these words as the output ($Q_{\mathcal{D}}$ for source documents and Q_z for the ground-truth summary). Note that we use special delimiters ⟨*sent-sep*⟩ and ⟨*doc-sep*⟩ to mark sentence and document boundaries, which allows us to extract sentence and document embeddings that we use as the initial sentence and document node embeddings in the graph encoder.

Graph Encoder

The graph encoder is responsible for learning node embeddings for the document graph $\mathcal{G}_{\mathcal{D}}$ and the ground-truth summary graph $\mathcal{G}_z$. We explain how the graph encoder works for the document graph below, but the same principle works for processing the ground-truth summary graph.

Node embeddings for the heterogeneous graph $\mathcal{G}_{\mathcal{D}}$ represent the words, sentences, and documents, and they are initialized using the contextual embeddings learned from the text encoder ($Q_{\mathcal{D}}$). As standard graph neural networks (GNNs) based on message passing cannot be applied to the heterogeneous graphs directly, we propose *multi-channel graph attention networks* (MGAT) inspired by graph attention networks (GAT; Velickovic et al. (2018)) to encode the heterogeneous graph.

Similar to GAT, MGAT is a multi-layer graph network. Intuitively, in each layer, MGAT aggregates embeddings of different channels (i.e., edge types) for each node. The com-

[4] By graph encoding we mean the collective node embeddings in the graph.

putation of the l-th layer of MGAT is given as follows:

$$\boldsymbol{h}_i^{(l+1)} = \boldsymbol{U}\boldsymbol{H}_i^{(l)} \tag{4}$$

$$\boldsymbol{H}_i^{(l)} = \|_{c=1}^{C} \boldsymbol{h}_i^{(l),c} \tag{5}$$

where $\boldsymbol{h}_i^{(l+1)}$ is the output embedding of node i in the l-th layer, $\|$ is the concatenation operation, C is the number of channels (which equals to the number of edge types in the heterogeneous graph, six in our case), and $\boldsymbol{U}$ is the shared transformation matrix for different nodes. Intuitively, $\boldsymbol{h}_i^{(l),c}$ represents the embedding of node i in the c-th channel at the l-th layer, and $\boldsymbol{H}_i^{(l)}$ is the concatenation of node embeddings from all channels for node i in the l-th layer. Note that the input node embeddings of the first layer of any channel are the output contextual embeddings (words, sentences, and documents) of the text encoder, i.e., $\boldsymbol{h}_i^{(0)} = \boldsymbol{q}_i$ where $\boldsymbol{q}_i \in \boldsymbol{Q}_{\mathcal{D}}$. The graph encoding, $\boldsymbol{Q}'_{\mathcal{D}}$, consists of all updated node embeddings from the final layer, i.e., $\boldsymbol{Q}'_{\mathcal{D}} = \|_i \boldsymbol{h}_i^{(L)}$.

To compute $\boldsymbol{h}_i^{(l),c}$ in each channel:

$$\boldsymbol{h}_i^{(l),c} = \|_{m=1}^{M} \sigma\Big(\sum_{j \in \mathcal{N}_i^c} \alpha_{ij}^{m,c} \boldsymbol{W}^{m,c} \boldsymbol{h}_j^{(l),c} \Big) \tag{6}$$

where M is the number of attention heads. We can now see that $\boldsymbol{h}_i^{(l),c}$ is the concatenated representation of M independent attention heads with different weight matrices $\boldsymbol{W}^{m,c}$ and normalized attention weights $\alpha_{ij}^{m,c}$, with the latter computed as follows:

$$\alpha_{ij}^{m,c} = \frac{\exp(d_{ij}^{m,c})}{\sum_{k \in \mathcal{N}_i^c} \exp(d_{ik}^{m,c})} \tag{7}$$

where $\mathcal{N}_i^c$ denotes the set of nodes connected to node i by an edge of type c. The attention coefficient $d_{ij}^{m,c}$ represents the correlation between nodes, and is learned as follows:

$$d_{ij}^{m,c} = \sigma\big(e_{ij} \cdot \boldsymbol{w}_{m,c}^{\top} [\boldsymbol{W}^{m,c}\boldsymbol{h}_i^{(l),c} \| \boldsymbol{W}^{m,c}\boldsymbol{h}_j^{(l),c}]\big) \tag{8}$$

where e_{ij} is the edge weight between node i and node j (defined in the Preliminaries section).

To summarize, MGAT computes node embeddings by attending to neighbouring nodes just like GAT, but it does this for each edge type independently and then concatenates them together to produce the final node embeddings, and it repeats this for multiple layers/iterations to learn higher order connections. We note that HGSUM has only one graph encoder, which is used to process both the source document graph $\mathcal{G}_{\mathcal{D}}$ to produce $\boldsymbol{Q}'_{\mathcal{D}}$ and the ground-truth summary graph $\mathcal{G}_z$ to produce $\boldsymbol{Q}'_z$.

Graph Compressor

Given $\mathcal{G}_{\mathcal{D}}$ and $\boldsymbol{Q}'_{\mathcal{D}}$ from the graph encoder, the graph compressor aims to "compress" the graph by selecting a subset of salient nodes and edges. Here we focus on filtering the sentence nodes, because we want to identify key sentences that help generate the summary. After the compression, all selected sentence nodes *and* their linked word and document nodes represent the compressed graph and their embeddings will be used by the text decoder for summary generation.

The graph compressor is inspired by Lee, Lee, and Kang (2019), and it works by computing the attention scores for all sentence nodes, filtering out nodes with the lowest scores, and then masking the rest using their attention scores. Firstly, attention scores of the sentence nodes are calculated based on the updated node embeddings from our proposed graph encoder $\mathrm{MGAT}(\boldsymbol{Q}_{\mathcal{D}}, \mathcal{G}_{\mathcal{D}})$:

$$\boldsymbol{t} = \mathrm{softmax}(\mathrm{MGAT}(\boldsymbol{Q}_{\mathcal{D}}, \mathcal{G}_{\mathcal{D}}) \cdot \boldsymbol{r}) \tag{9}$$

where $\boldsymbol{r}$ is the only trainable parameter of the graph compressor which transforms the updated node embedding into a scalar. Then, based on these scores, we select sentence nodes with the highest scores:

$$\mathcal{I}_s = \text{top-k}(\boldsymbol{t}, k, \mathcal{G}_{\mathcal{D}}) \tag{10}$$

$$\mathcal{I} = \mathrm{extend}(\mathcal{I}_s, \mathcal{G}_{\mathcal{D}}) \tag{11}$$

where top-k is a function that selects top-ranked sentence nodes in $\mathcal{G}_{\mathcal{D}}$ based on $\boldsymbol{t}$, $k \in (0, 1]$ is a hyper-parameter that determines the ratio of sentence nodes to be kept, $\mathcal{I}_s$ is the set of selected sentence nodes, and extend is a function that extends the selected sentence nodes in $\mathcal{I}_s$ to include word and document nodes that they link to (and so $\mathcal{I}$ includes word, sentence and document nodes). Lastly, we mask all the selected nodes using their attention scores, producing the encoding of the compressed graph, $\boldsymbol{Q}_p$:

$$\boldsymbol{Q}_p = \|_i^{\mathcal{I}} \boldsymbol{q}'_i \times \boldsymbol{t}_i, \boldsymbol{q}'_i \in \boldsymbol{Q}'_{\mathcal{D}}. \tag{12}$$

Text Decoder

The text decoder follows the same architecture as a decoder Transformer (which uses masked attention to prevent attention to future words), is initialized with PRIMERA weights, and takes $\boldsymbol{Q}_p$ as input to generate the summary:

$$\hat{z} = \mathrm{transformer}(\boldsymbol{Q}_p) \tag{13}$$

Note that the node embeddings in $\boldsymbol{Q}_p$ retain the original word index in the source documents $\mathcal{D}$, and as such positional embeddings are added to them following standard transformer architecture.

Multi-Task Training

HGSUM is trained with two objectives: maximizing the likelihood of generating the ground-truth summary z and the graph similarity between the compressed graph encoding $\boldsymbol{Q}_p$ and ground-truth summary graph encoding $\boldsymbol{Q}'_z$.

To maximize the likelihood of generating the ground-truth summary, we minimize the cross entropy over the ground-truth summary and the generated summary with conventional teacher forcing.

$$\mathcal{L}_{ce} = -\frac{1}{T}\sum_{i=1}^{T} w_i \log \hat{w}_i \tag{14}$$

where w_i is the i-th word in the ground-truth summary, while $\hat{w}_i$ is the i-th word in the generated summary.

To maximize the graph similarity, we compute the cosine similarity of the average node embeddings from the compressed graph and the ground-truth summary graph:

$$\mathcal{L}_{gs} = -\,\mathrm{sim}(\mathrm{avg}(\boldsymbol{Q}_p), \mathrm{avg}(\boldsymbol{Q}'_z)) \tag{15}$$

Dataset	#c	#d/c	#w/d	#w/s
MULTI-NEWS	56,216	2.79	690.97	241.61
WCEP-100	10,200	63.38	439.24	30.53
ARXIV	215,913	5.63	978.17	251.07

Table 1: Dataset statistics. "c" = cluster; "d" = document; "w" = word; and "s" = summary. "#" denotes "the number of" and "/" denotes "in each".

Model	#parameters	Len-in	Len-out
PEGASUS	568M	1,024	512
LED	459M	16,384	512
PRIMERA	447M	4,096	512
MGSum	129M	2,000	400
GraphSum	463M	4,050	300
HGSUM	501M	4,096	512

Table 2: Model parameter sizes. Len-in and Len-out denote the maximum lengths of the model input and the model output, respectively.

The final loss function of HGSUM is the sum of $\mathcal{L}_{ce}$ and $\mathcal{L}_{gs}$ weighted by hyper-parameter $\beta \in (0, 1)$.

$$\mathcal{L} = \beta\mathcal{L}_{ce} + (1 - \beta)\mathcal{L}_{gs} \tag{16}$$

Experiments

We test our proposed model HGSUM and compare it against state-of-the-art abstractive MDS models over several datasets. We also report the results of an ablation study to show the effectiveness of the components of HGSUM.

Experimental Setup

Datasets We use MULTI-NEWS (Fabbri et al. 2019), WCEP-100 (Ghalandari et al. 2020), and ARXIV (Cohan et al. 2018) as benchmark English datasets. These datasets come from different domains including news, Wikipedia, and scientific domains. MULTI-NEWS contains clusters of news articles plus a summary corresponding to each cluster written by professional editors. WCEP-100 contains human-written summaries of different news events from Wikipedia. In ARXIV, each cluster corresponds to a research paper in the scientific domain, where the paper abstract is used as the summary, while sections of the paper are used as the source documents in each cluster. Table 1 summarizes statistics of these datasets.

Competitors We compare our model with two groups of state-of-the-art abstractive MDS models: *PLM-based* and *graph-based*. (1) The PLM-based models include **PEGASUS** (Zhang et al. 2020a), **LED** (Beltagy, Peters, and Cohan 2020), and **PRIMERA** (Xiao et al. 2022). LED is a general-purpose PLM that introduces the longformer architecture which uses sparse self-attention to allow it to process much longer input than previous models. LED is pre-trained by reconstructing documents from their corrupted input in the same way as BART (Lewis et al. 2020). In contrast, PEGASUS and PRIMERA are pre-trained models designed for summarization (the former for single-document and the latter multi-document summarization). Specifically, PEGASUS is pre-trained by generating pseudo summaries for documents, where the pseudo summaries are composed of gap sentences extracted from a document based on ROUGE scores. PRIMERA is similarly pre-trained to generate pseudo summaries, but their pseudo summaries are extracted based on the salience of entities which correspond to their document frequency. For these PLM-based models, we take their off-the-shelf models and fine-tune them on our datasets. We follow the standard approach where we concatenate documents from the same cluster to form a long and flat input string. (2) For the graph-based models, we compare against **MGSum** (Jin, Wang, and Wan 2020)[5] and **GraphSum** (Li et al. 2020)[6]. To model cross-document relationships in MDS, MGSum (Jin, Wang, and Wan 2020) uses a three-level hierarchical graph to represent source documents, including different levels of nodes (documents, sentences, and words). It learns semantics with a multi-level interaction network. Although there are different types of nodes in this hierarchical graph, all of its edges are of the same type (i.e., it is a homogeneous graph).[7] GraphSum (Li et al. 2020) uses a similarity graph over paragraphs to capture cross-document relationships, and it uses pre-trained RoBERTa (Liu et al. 2019) as its encoder. Just like MGSum, its graph is homogeneous.

Implementation Details For the PLMs, we use the large version of the models which roughly have the same number of parameters (Table 2).[8] For the graph-based models, we use open-source code from the original authors and train them on our datasets, following their recommended hyper-parameters and configurations. As Table 2 shows, most models are trained to generate a maximum length of 512 subwords ("Len-out") for the summary (exception: MGSum and GraphSum where we follow the original output length). Note though that the maximum input lengths ("Len-in") of these models range from 1K-16K subwords, depending on the architecture of their encoder.

For HGSUM, the text encoder and decoder are initialized with PRIMERA weights. To alleviate overfitting, we apply label smoothing during training with a smoothing factor of 0.1. We use beam search decoding with beam width 5 to generate the summary. The hyper-parameter β is set to 0.5 to balance two loss functions. All other hyper-parameters are tuned based on the development set.

All experiments are run on Intel(R) Xeon(R) Gold 6326 CPU @ 2.90GHz with NVIDIA Tesla A100 GPU (40G).

Overall Results

We report the average F1 of ROUGE-1 (R-1), ROUGE-2 (R-2) and ROUGE-L (R-L) (Lin and Hovy 2003). Note that we

[5] https://github.com/zhongxia96/MGSum

[6] https://github.com/PaddlePaddle/Research/tree/master/NLP

[7] For fair comparison we use the abstractive variant of MGSum.

[8] PLM-based models are implemented using the HuggingFace library: https://huggingface.co/

Model	Multi-News			WCEP-100			Arxiv		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
PEGASUS	47.70	18.36	43.62	42.43	17.33	32.35	44.21	16.95	38.87
LED	47.68	19.72	43.83	43.05	20.94	34.99	46.50	18.96	41.87
PRIMERA	49.40	20.51	45.35	43.11	**21.85**	35.89	47.24	20.24	42.61
MGSum	45.63	16.71	40.92	38.88	14.22	23.37	40.58	11.22	29.93
GraphSum	45.71	17.12	41.99	39.56	14.38	29.41	42.98	16.55	37.01
HGSum (our model)	**50.64**†	**21.69**†	**45.90**†	**44.21**†	21.81	**36.21**†	**49.32**†	**21.30**†	**44.50**†
Performance gain	+2.51%	+5.75%	+1.21%	+2.55%	-0.18%	+0.89%	+4.40%	+5.24%	+4.44%

Table 3: Model performance on summarizing Multi-News, WCEP-100, and Arxiv in terms of F1 of ROUGE scores. The best performance results are in boldface, while the second best is underlined. †: significantly better than others (p-value < 0.05).

Doc 1	*... Parents are risking their babies' health because of a surge in the popularity of swaddling ...*
Doc 2	*There has been a recent resurgence of swaddling because of ...*
Doc 3	*... Swaddling babies is on the rise: Add it to the long list of mixed messages new parents get about infant care ...*
Generated summary	*The trend of swaddling babies is on the rise, but an orthopaedic surgeon ... is warning parents against the practice.*

Table 4: An example of a generated summary in Multi-News by HGSum.

Model	**R-1**	**R-2**	**R-L**	**BScore**
HGSum	50.64	21.69	45.90	87.38
w/o MGAT	48.87	20.32	43.21	87.08
w/o graph compressor	49.00	20.38	45.01	86.92
w/o multi-task training	48.10	20.30	44.24	86.85

Table 5: Results of ablation study on Multi-News.

Initialized by	**R-1**	**R-2**	**R-L**	**BScore**
random weights	18.99	27.86	16.88	79.32
LED	48.36	19.99	44.25	86.73
PRIMERA	50.64	21.69	45.90	87.38

Table 6: Summarization results of HGSum with different initialization on Multi-News.

use the summary-level R-L,[9] and each summary is split into sentences using NLTK[10].

Table 3 reports the performance of all models over all datasets. HGSum outperforms most of the benchmark systems, demonstrating the effectiveness of incorporating a compressed heterogeneous graph for text summarization. Interestingly, the PLMs (PEGASUS, LED, PRIMERA, and HGSum) also seem to be consistently better than graph-based models (MGSum and GraphSum). This shows that using graph-based document representations does not necessarily lead to better MDS results, thus confirming the advantage of our heterogeneous graph-based model design. We give an example of generated summary by HGSum in Multi-News in Table 4.

Ablation Study

To show the effectiveness of the HGSum components, we conduct an ablation study and compare it with three model variants: (1) HGSum **w/o MGAT**, which replaces MGAT with the vanilla GAT model that treats all graph nodes and edges as being the same type, (2) HGSum **w/o graph compressor**, which drops the graph compressor from HGSum and uses the output from the graph encoder directly as the input for the text decoder, and (3) HGSum **w/o multi-task training**, which replaces the multi-task objective using only the cross entropy objective.

For the ablation results, we also present the performance in terms of BERTScore ("BScore"; Zhang et al. (2020b)), which measures the semantic similarity between the ground-truth and generated summary based on BERT embeddings. Table 5 shows the ablation results on the test set of Multi-News.[11] We see that removing the heterogeneous graph encoder, graph compressor, or the multi-task objective result in a performance drop over all metrics, confirming the effectiveness of these components. In particular, dropping the multi-task objective leads to the largest degradation in model performance, suggesting that this auxiliary task is essential to help HGSum learn how to compress the graph for summarization.

More Analysis

Impact of Text Encoder and Decoder Initialization Our text encoder and decoder can be initialized by any pre-trained Transformer models. Here we make a comparison

[9] We note that prior studies use a mixture of summary-level and sentence-level R-L, and for more details about their differences, we refer the reader to: https://pypi.org/project/rouge-score/

[10] https://www.nltk.org/

[11] We found similar results for different datasets, and present only Multi-News here in light of space.

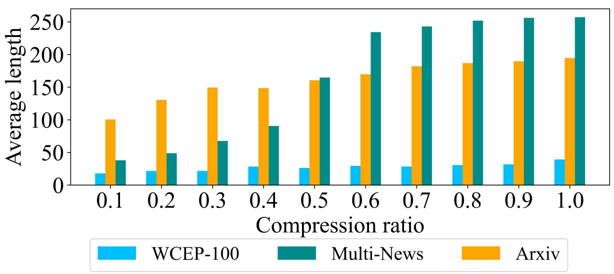


Figure 3: Average lengths of generated summaries for different datasets when the compression ratio k is set to different values.

on initialization using PRIMERA, the large version of LED and random weights. Table 6 shows results using such initialization strategies on the test set of MULTI-NEWS. We see that initialization with random weights has much worse performance than initialization using pre-trained PLMs, which is expected. Using PRIMERA leads to better empirical performance than using the LED, consistent with prior findings.

Impact of the Graph Compression Ratio k The hyperparameter k in the heterogeneous graph pooling is to control the proportion of sentence nodes to be retained in the compressed graph. To understand how much k affects the generated summary length, we present average lengths of generated summaries for different datasets when the compression ration k is set to different values in Figure 3. Interestingly, we see that larger k generally produces longer summary, and this effect is strongest for MULTI-NEWS.

Related Work

Abstractive Multi-Document Summarization

PLM-Based Models Recent PLM-based models have shown strong performance for abstractive text summarization tasks. These models follow a Transformer-based (Vaswani et al. 2017) encoder-decoder architecture. For example, general-purpose PLMs such as T5 (Raffel et al. 2020), BART (Lewis et al. 2020), and LED (Beltagy, Peters, and Cohan 2020) can be fine-tuned for abstractive text summarization. PEGASUS (Zhang et al. 2020a) is a strong PLM-based model pre-trained with an objective that predicts gap sentences as a pseudo summary. These models can be used for MDS by concatenating the source documents into a single document. PRIMERA (Xiao et al. 2022) has the same architecture as LED, but is designed for MDS specifically in that it is pre-trained to generate pseudo summaries — text spans that are automatically extracted based on the entity salience. Although these models show impressive performances and can even handle zero-shot cases, they use a flat concatenation of the input documents, which limits their capability in learning the cross-document relationships among different semantic units.

Graph-Based Models Although graphs are commonly used to boost text summarization (Wu et al. 2021b; You et al. 2022; Song and King 2022), there are only a handful of models which have been proposed to use graphs to encode the documents in abstractive MDS (Li et al. 2020; Jin, Wang, and Wan 2020; Li and Zhuge 2021; Cui and Hu 2021). Most of these models only leverage homogeneous graphs as they do not consider different edge types of graphs. For example, MGSum (Jin, Wang, and Wan 2020) constructs a three-level (i.e., document, sentence, and word levels) hierarchical graph and learns semantics with a multi-level interaction network. GraphSum (Li et al. 2020) constructs a similarity graph over the paragraphs. It learns a graph representation for the paragraphs and uses a hierarchical graph attention mechanism to guide the summary generation process. The graphs constructed in these models are in fact homogeneous, in that GraphSum only consider paragraph nodes, and MGSum uses the same edge type to connect the graph nodes.

Graph Neural Networks

Graph Modeling GNNs have yielded strong performance for modeling documents (Wu et al. 2021a), e.g., to model relationships among text spans for MDS. Graph convolutional networks (GCN; Kipf and Welling (2017)) and graph attention networks (GAT; Velickovic et al. (2018)) are two representative GNN models, which are frequently used in modeling graph-structured data composed of nodes and edges. GAT is based on the attention mechanism (Vaswani et al. 2017), while GCN is based on Laplacian transformation on the adjacency matrix. Another difference between these two is that edge weights of GCNs (i.e., the adjacency matrix) are fixed in training but those of GAT (i.e., the attentions) can be updated, although both of them perform message passing (Gilmer et al. 2017) on graphs.

Graph Pooling Graph pooling (Liu et al. 2022) aggregates node embeddings to obtain compressed graph representations. Existing graph pooling methods can be largely grouped into two categories: *global pooling* and *hierarchical pooling*. Global pooling generates the graph representation with a mean- or sum-pooling over the node embeddings. This method does not preserve the hierarchical structure of graphs. Hierarchical pooling, in contrast, considers the graph structure by compressing an input graph into smaller graphs iteratively, through node clustering (Bianchi, Grattarola, and Alippi 2020) or node dropping (Lee, Lee, and Kang 2019). Our graph compressor follows the idea of the hierarchical pooling, and condenses the graph by removing nodes to generate a small-sized graph.

Conclusion

We propose HGSUM, an extended encoder-decoder model that builds on PLMs to incorporate a compressed heterogeneous graph for abstractive multi-document summarization. HGSUM is novel in that it captures the heterogeneity between words, sentences, and document units in the constructed graph for source documents, and it also learns to compress the heterogeneous graph by 'mimicking' the ground-truth summary graph during training. Experimental results over multiple datasets show that HGSUM outperforms current state-of-the-art MDS systems.

Acknowledgements

We would like to thank the anonymous reviewers for their great comments. This work is supported by China Scholarship Council (CSC).

References

Beltagy, I.; Peters, M. E.; and Cohan, A. 2020. Longformer: The Long-Document Transformer. *CoRR*, abs/2004.05150.

Bianchi, F. M.; Grattarola, D.; and Alippi, C. 2020. Spectral Clustering with Graph Neural Networks for Graph Pooling. In *ICML*, 874–883.

Bowman, S. R.; Angeli, G.; Potts, C.; and Manning, C. D. 2015. A large annotated corpus for learning natural language inference. In *EMNLP*, 632–642.

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 *NAACL-HLT*, 615–621.

Cui, P.; and Hu, L. 2021. Topic-Guided Abstractive Multi-Document Summarization. In *Findings of EMNLP*, 1463–1472.

Fabbri, A. R.; Li, I.; She, T.; Li, S.; and Radev, D. R. 2019. Multi-News: A Large-Scale Multi-Document Summarization Dataset and Abstractive Hierarchical Model. In *ACL*, 1074–1084.

Gerani, S.; Mehdad, Y.; Carenini, G.; Ng, R. T.; and Nejat, B. 2014. Abstractive Summarization of Product Reviews Using Discourse Structure. In *EMNLP*, 1602–1613.

Ghalandari, D. G.; Hokamp, C.; Pham, N. T.; Glover, J.; and Ifrim, G. 2020. A Large-Scale Multi-Document Summarization Dataset from the Wikipedia Current Events Portal. In *ACL*, 1302–1308.

Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; and Dahl, G. E. 2017. Neural Message Passing for Quantum Chemistry. In *ICML*, 1263–1272.

Jin, H.; Wang, T.; and Wan, X. 2020. Multi-Granularity Interaction Network for Extractive and Abstractive Multi-Document Summarization. In *ACL*, 6244–6254.

Kipf, T. N.; and Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *ICLR*.

Lee, J.; Lee, I.; and Kang, J. 2019. Self-Attention Graph Pooling. In *ICML*, 3734–3743.

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 *ACL*, 7871–7880.

Li, W.; Xiao, X.; Liu, J.; Wu, H.; Wang, H.; and Du, J. 2020. Leveraging Graph to Improve Abstractive Multi-Document Summarization. In *ACL*, 6232–6243.

Li, W.; and Zhuge, H. 2021. Abstractive Multi-Document Summarization Based on Semantic Link Network. *TKDE*, 33(1): 43–54.

Lin, C.; and Hovy, E. H. 2003. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In *HLT-NAACL*, 71–78.

Liu, C.; Zhan, Y.; Li, C.; Du, B.; Wu, J.; Hu, W.; Liu, T.; and Tao, D. 2022. Graph Pooling for Graph Neural Networks: Progress, Challenges, and Opportunities. *CoRR*, abs/2204.07321.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Ma, C.; Zhang, W. E.; Guo, M.; Wang, H.; and Sheng, Q. Z. 2020. Multi-document Summarization via Deep Learning Techniques: A Survey. *CoRR*, abs/2011.04843.

Moro, G.; Ragazzi, L.; Valgimigli, L.; and Freddi, D. 2022. Discriminative Marginalized Probabilistic Neural Method for Multi-Document Summarization of Medical Literature. In *ACL*, 180–189.

Otmakhova, Y.; Verspoor, K.; Baldwin, T.; and Lau, J. H. 2022. The Patient Is More Dead than Alive: Exploring the Current State of the Multi-document Summarisation of the Biomedical Literature. In *ACL*, 5098–5111.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global Vectors for Word Representation. In *EMNLP*, 1532–1543.

Radev, D. R. 2000. A Common Theory of Information Fusion from Multiple Text Sources Step One: Cross-Document Structure. In *SIGDIAL Workshop*, 74–83.

Radev, D. R.; Hovy, E. H.; and McKeown, K. R. 2002. Introduction to the Special Issue on Summarization. *Computational Linguistics*, 28(4): 399–408.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *JMLR*, 21: 140:1–140:67.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *EMNLP-IJCNLP*, 3980–3990.

Song, Z.; and King, I. 2022. Hierarchical Heterogeneous Graph Attention Network for Syntax-Aware Summarization. In *AAAI*, 11340–11348.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *NeurIPS*, 5998–6008.

Velickovic, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; and Bengio, Y. 2018. Graph Attention Networks. In *ICLR*.

Wu, L.; Chen, Y.; Shen, K.; Guo, X.; Gao, H.; Li, S.; Pei, J.; and Long, B. 2021a. Graph Neural Networks for Natural Language Processing: A Survey. *CoRR*, abs/2106.06090.

Wu, W.; Li, W.; Xiao, X.; Liu, J.; Cao, Z.; Li, S.; Wu, H.; and Wang, H. 2021b. BASS: Boosting Abstractive Summarization with Unified Semantic Graph. In *ACL*, 6052–6067.

Xiao, W.; Beltagy, I.; Carenini, G.; and Cohan, A. 2022. PRIMERA: Pyramid-based Masked Sentence Pre-training for Multi-document Summarization. In *ACL*, 5245–5263.

You, J.; Li, D.; Kamigaito, H.; Funakoshi, K.; and Okumura, M. 2022. Joint Learning-based Heterogeneous Graph Attention Network for Timeline Summarization. In *NAACL*, 4091–4104.

Zhang, J.; Zhao, Y.; Saleh, M.; and Liu, P. J. 2020a. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. In *ICML*, 11328–11339.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2020b. BERTScore: Evaluating Text Generation with BERT. In *ICLR*.

3.2 Reflections

In this work, we propose a MDS model which encodes source documents as heterogeneous graphs, and it is trained with a multi-task objective borrowing the idea of heterogeneous graph compression. Based on our experiments on multiple widely-used datasets, including ARXIV (Cohan et al., 2018), MultiNews (Fabbri et al., 2019) and WCEP-100 (Ghalandari et al., 2020), our model achieves the state-of-the-art performances in terms of ROUGE (Lin and Hovy, 2003) and BERTScore (Zhang et al., 2020b). This suggests that representing source documents with heterogeneous graphs is promising to improve the quality of generated summaries for MDS. Specifically, although BERTScore is a suboptimal metric capturing semantic similarity between the generated summary and the reference summary (introduced in Chapter 2), better performances in terms of BERTScore indicate that summaries generated by our model are semantically closer to the reference summaries and more faithful to input documents.

When we did the study in 2023, text summarization was mainly based on encoder-decoder pre-trained language models. However, LLMs have since emerged, such as GPT-4 (OpenAI, 2023) and DeepSeek-V3 (DeepSeek-AI, 2024) as discussed in Chapter 2. It remains a question how we can incorporate explicit graph representations into decoder-only LLMs and whether it would improve the performance of them on MDS. We could perhaps inject textual representation of the heterogeneous graph into the prompt to inject explicit cross-document relationships into the model (e.g., by describing the graph with a JSON). This might make the model more sensitive to those described relationships in prompts.

Although our proposed approach achieves the state-of-the-art performance on the experimental datasets, there are limitations for the work in modelling methodology, experimental data, and evaluation metrics.

For modelling, although the proposed heterogeneous graph compression has improved the quality of generated summaries by better capturing cross-document relationships, it is time-consuming to construct the graphs for large-scale documents. This is because we have to get embeddings for sentences in all documents and calculate cosine similarities among them to build the graphs. It is important to find a more efficient way to represent the multi-document input without losing cross-document relationship information. Also, we could incorporate other structural information into our heterogeneous graphs. For example, we could extract the document structure as articles all have multiple sections and use sections as another type of nodes in our graph. This will make the model aware of the document structures. In addition, although our approach can generate better summaries with the heterogeneous graphs, our graphs cannot still capture some complex cross-document relationships. For example, our model may struggle in understanding the contradiction based on negation because our graphs are constructed based on similarity among words and sentences and the embeddings may not be able to distinguish the negation texts in its embedding space. Therefore, we should have better datasets which contain meta information about various cross-document relationships, such as what the cross-document relationships are and where they happen in the input text. Only if we have this kind of data can we understand more about where and why these summarization models fail to solve the cross-document relationships.

For the experimental data, our experiments are only on English datasets, and we should evaluate our approach on datasets in more languages to understand whether these results translate to other languages. That said, our models should be easily adaptable to other languages assuming the backbone pre-trained language models and embedding models work on those languages. Additionally, in our experimented datasets, there may not be only ideational information as in news articles there are editorial opinions. As such, an important preprocessing step is to filter these opinion-driven articles, although it may not be trivial to do this at scale (since it involves detecting opinion-driven articles vs. ideational documents).

Lastly, for evaluation metrics, as the development of text generation there are more evaluation metrics that directly assess the quality of generated texts such as G-Eval (Liu et al., 2023b) and AlignScore (Zha et al., 2023) (Chapter 2.4), and these metrics focus more on meaning of texts instead of the surface form. It would be interesting to conduct fine-grained evaluation on our generated summaries on the experimental datasets with these recent advanced metrics to further confirm that our approach improves the performance of MDS, especially in faithfulness of the generated summaries. In addition, we only evaluate the performance of our model based on the ground truth summaries. It would also be interesting to conduct more detailed evaluation on how the models handle specific cross-document relationships. For example, if we get a cluster of documents with conflicts (i.e., information from one article that contradicts another article), we should analyse the model behaviours in handling these conflicts and potentially also other cross-document relationships. However, there is no MDS datasets with annotated cross-document relationships. As such, we will construct a benchmark summarization dataset with explicit cross-document relationships in Chapter 4.

Chapter 4

Scientific Opinion Summarization

We have investigated multi-document summarization (MDS) over ideational documents in Chapter 3. While ideational documents primarily focus on presenting ideas, concepts, or facts, opinionated documents primarily convey viewpoints, preferences, evaluations, or subjective stances on a particular topic. Reviews as opinionated documents are omnipresent in the digital world, providing invaluable insights into products (Brazinskas et al., 2021), businesses (Angelidis et al., 2021), and scientific articles. For example, in scientific peer reviewing the reviews provide opinions to accept or reject a research paper. MDS over ideational documents is to summarize the salient factual information among the documents, while summarizing over opinionated documents is to extract the 'overall' opinions.[1] As reviewed in Chapter 2.1, most existing MDS datasets are based on ideational documents and none of them provide explicit cross-document relationships which hinders the development of research in the domain.

This chapter aims to address the second research question of the thesis: *how to integrate opinionated information for opinion summarization*. It is challenging because models have to possess the capability to aggregate opinions from different perspectives, even when the opinions may contradict with each other.

[1] We use 'meta-review' and 'summary' interchangeably in opinion summarization.

We found that peer-review platforms such as OpenReview[2] is promising for facilitating research on MDS over opinionated documents. Their meta-reviewers have to understand reviews from the reviewers and write a meta-review to aggregate their opinions to express an overall opinion for each research paper. These meta-reviews are written based on the reviews and the conversations between reviewers and the author. The task of scientific opinion summarization requires models to understand the conversational structure among the reviews and the complex cross-document relationships especially the conflicts among the reviews. It is a good test bed for MDS systems. In addition, the developed opinion summarization models for the scientific domain can be used to potentially aid human meta-reviewers by automatically generating a first draft of their meta-reviews.

Therefore, we use those peer-review platforms to construct a MDS dataset, and build summarization models using it. While this chapter focuses on opinion summarization in the scientific domain, we will revisit this to explore opinion summarisation methods that work across domains in Chapter 5.

The content of this chapter is based on the following two publications. We first present the task of scientific meta-review generation with a benchmark dataset in Chapter 4.1, and then further investigate the sentiment consolidation capabilities of existing models on scientific reviews in Chapter 4.2.

- **Miao Li**, Eduard Hovy, and Jey Han Lau. 2023. Summarizing Multiple Documents with Conversational Structure for Meta-Review Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7089–7112, Singapore. Association for Computational Linguistics.

- **Miao Li**, Jey Han Lau, and Eduard Hovy. 2024. A Sentiment Consolidation Framework for Meta-Review Generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10158–10177, Bangkok, Thailand. Association for Computational Linguistics.

[2] https://openreview.net/

4.1 Scientific Meta-Review Generation

Scientific meta-review generation is the task to automatically write a meta-review in an abstractive fashion to summarize opinionated information in the scientific peer-review process. Meta-review generation in other domains such as products (Brazinskas et al., 2021), businesses (Angelidis et al., 2021) only contain meta-reviews (i.e., summaries) and their corresponding source documents without any explicit cross-document relationships (that is, the source documents are a list of reviews without any explicit relationships or structure between them). However, scientific meta-reviews in the computer science domain on OpenReview are usually written based on official reviews (written by assigned reviewers), public reviews (written by public users), and the multi-turn conversations between assigned reviewers, public users and the paper author. By using OpenReview data, we could get data with explicit conversational structures. Although there are existing scientific meta-review generation datasets (Bhatia et al., 2020; Shen et al., 2022), their source documents are only composed of official reviews without considering the multi-turn conversations.

We fist develop a new summarization dataset from OpenReview. In our dataset, the meta-review is the summary, and there are seven types of source documents: (1) official reviews (reviews by assigned reviewers); (2) public reviews (comments by the public users); (3) author comments (an overall response by paper authors); (4) official responses (responses by assigned reviewers); (5) public responses (responses by public users); and (6) author responses within a thread. Because the reviews also contain ratings, we use that information to investigate instances with and without conflicts to understand how summarization models resolve conflicts in source documents. We assess the quality of this new OpenReview summarisation dataset in terms of abstractiveness and faithfulness. For abstractiveness, we calculate the percentages of unigrams, bigrams, and trigrams of summaries that exist in summaries but not in any source documents. To assess the faithfulness of our summaries, we conduct human evaluation by highlighting text spans in the summary that can be semantically anchored to the source documents.

As a baseline model, we develop a meta-review generation model based on pre-trained language models to investigate whether the conversational structure of the source documents can be used to improve the quality of generated meta-reviews. As the original Transformer architecture cannot explicitly capture the cross-document structural information, we modified pre-trained encoder-decoder language models with a relationship-aware sparse attention mechanism to incorporate the conversational structure as part of the encoding process. We fine-tune the model with multi-task learning utilizing the metadata information to additionally predict source document types, review ratings/confidences and the paper acceptance outcome in addition to the next-word prediction objective using human-written meta-reviews.

Summarizing Multiple Documents with Conversational Structure for Meta-Review Generation

Miao Li[1] and **Eduard Hovy**[1,2] and **Jey Han Lau**[1]
[1]School of Computing and Information Systems, The University of Melbourne
[2]Language Technologies Institute, Carnegie Mellon University
miao4@student.unimelb.edu.au, {eduard.hovy, laujh}@unimelb.edu.au

Abstract

We present **PEERSUM**, a novel dataset for generating meta-reviews of scientific papers. The meta-reviews can be interpreted as abstractive summaries of reviews, multi-turn discussions and the paper abstract. These source documents have rich inter-document relationships with an explicit hierarchical conversational structure, cross-references and (occasionally) conflicting information. To introduce the structural inductive bias into pre-trained language models, we introduce **RAMMER** (Relationship-aware Multi-task Meta-review Generator), a model that uses sparse attention based on the conversational structure and a multi-task training objective that predicts metadata features (e.g., review ratings). Our experimental results show that RAMMER outperforms other strong baseline models in terms of a suite of automatic evaluation metrics. Further analyses, however, reveal that RAMMER and other models struggle to handle conflicts in source documents of PEERSUM, suggesting meta-review generation is a challenging task and a promising avenue for further research.[1]

1 Introduction

Text summarization systems need to recognize internal relationships among source texts and effectively aggregate and process information from them to generate high-quality summaries (El-Kassas et al., 2021). It is particularly challenging in multi-document summarization (MDS) due to the complexity of the relationships among (semi-)parallel source documents (Ma et al., 2020). However, existing MDS datasets do not provide explicit inter-document relationships among the source documents (Liu et al., 2018; Fabbri et al., 2019; Ghalandari et al., 2020; Lu et al., 2020) although inter-document relationships may also exist in nature and should be considered in methodology (Fabbri et al., 2019). This makes it hard to research inter-document relationship comprehension for information integration and aggregation in abstractive text summarization.

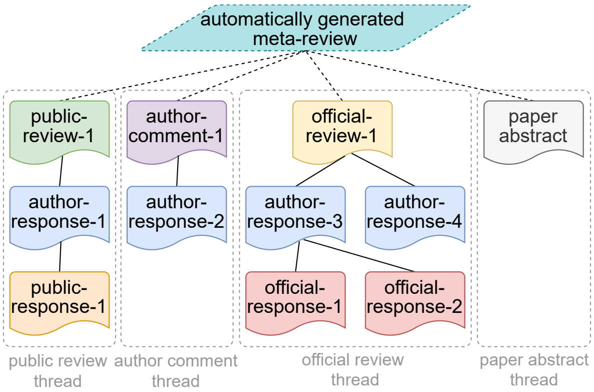


Figure 1: An illustration of the hierarchical conversational structure that PEERSUM features.

To enable this, we introduce PEERSUM, an MDS dataset for automatic meta-review generation. We formulate the creation of meta-reviews as an abstractive MDS task as the meta-reviewer needs to comprehend and carefully summarize information from individual reviews, multi-turn discussions between authors and reviewers and the paper abstract. From an application perspective, generating draft meta-reviews could serve to reduce the workload of meta-reviewers, as meta-reviewing is a highly time-consuming process for many scientific publication venues.

PEERSUM features a hierarchical conversational structure among the source documents which includes the reviews, responses and the paper abstract in different threads as shown in Figure 1. It has several distinct advantages over existing MDS datasets: (1) we show that the meta-reviews are largely faithful to the corresponding source documents despite being highly abstractive; (2) the source documents have rich inter-document relationships with an explicit conversational structure;

[1] The dataset and code are available at https://github.com/oaimli/PeerSum

Findings of the Association for Computational Linguistics: EMNLP 2023, pages 7089–7112
December 6-10, 2023 ©2023 Association for Computational Linguistics

(3) the source documents occasionally feature *conflicts* which the meta-review needs to handle as reviewers may have disagreement on reviewing a scientific paper, and we explicitly provide indicators of conflict relationships along with the dataset; and (4) it has a rich set of metadata, such as review rating/confidence and paper acceptance outcome — the latter which can be used for assessing the quality of automatically generated meta-reviews. These make PEERSUM serve as a probe that allows us to understand how machines can reason, aggregate and summarise potentially conflicting opinions.

However, there is limited study on abstractive MDS methods that can recognize relationships among source documents. The most promising approaches are based on graph neural networks (Li et al., 2020, 2023), but they introduce additional trainable parameters, and it is hard to find effective ways to construct graphs to represent source documents. To make pre-trained language models have the comprehension ability of complex relationships among source documents for MDS, we propose RAMMER, which uses *relationship-aware attention manipulation* — a lightweight approach to introduce an inductive bias into pre-trained language models to capture the hierarchical conversational structure in the source documents. Concretely, RAMMER replaces the full attention mechanism of Transformer (Vaswani et al., 2017) with sparse attention that follows a particular relationship in the conversational structure (e.g., the parent-child relation). To further improve the quality of generated meta-reviews by utilising the metadata information, RAMMER is trained with a multi-task objective to additionally predict source document types, review ratings/confidences and the paper acceptance outcome.

We conduct experiments to compare the performance of RAMMER with a number of baseline models over automatic evaluation metrics including the proposed evaluation metric based on predicting the paper acceptance outcome and human evaluation. We found that RAMMER performs strongly, demonstrating the benefits of incorporating of the conversational structure and the metadata. Further analyses on instances with conflicting source documents, however, reveal that it still struggle to recognise and resolve these conflicts, suggesting that meta-review generation is a challenging task and promising direction for future work.

2 Related Work

2.1 MDS Datasets

There are a few popular MDS datasets for abstractive summarization in these years, such as WCEP (Ghalandari et al., 2020), Multi-News (Fabbri et al., 2019), Multi-XScience (Lu et al., 2020), and WikiSum (Liu et al., 2018) from news, scientific and Wikipedia domains. Multi-XScience is constructed using the related work section of scientific papers, and takes a paragraph of related work as a summary for the abstracts of its cited papers. Although the summaries are highly abstractive, they are not always reflective of the cited papers — this is attested by the authors' finding that less than half of the statements in the summary are grounded by their source documents. WikiSum and WCEP have a similar problem as they augment source documents with retrieved documents and as such they may only be loosely related to the summary. Notably, none of the source documents in these datasets provides any explicit structure of inter-document relationships or conflicting information, although different inter-document relationships may exist among source documents in these datasets (Ma et al., 2020). This leads to underexplored research on inter-document relationship comprehension of abstractive summarization models. In the peer-review domain, Shen et al. (2022); Wu et al. (2022) developed datasets for meta-review generation. However, they only consider official reviews, or their datasets do not feature the rich hierarchical conversational structure that PEERSUM has.

2.2 Structural Inductive Bias for Summarization

Transformer-based pre-trained language models (PLMs) (Lewis et al., 2020; Zhang et al., 2020a; Guo et al., 2022; Phang et al., 2022) are the predominant approach in abstractive text summarization. However, it is challenging to incorporate structural information into the input as Transformer is designed to process flat text sequences. As such, most studies for MDS treat the input documents as a long flat string (via concatenation) without any explicit inter-document relationships (Xiao et al., 2022; Guo et al., 2022; Phang et al., 2022). To take into account the structural information, most work uses graph neural networks (Li et al., 2020; Jin et al., 2020; Cui and Hu, 2021; Li et al., 2023) but it is difficult to construct effective graphs to

Features	ICLR	NeurIPS
#samples	9,835	5,158
#official-review-thread/cluster	3.51	3.67
#author-comment-thread/cluster	0.59	0.01
#public-review-thread/cluster	0.22	0.00
#paper-abstract-thread/cluster	1.0	1.0

Table 1: PEERSUM statistics.

represent multiple documents and they introduce additional parameters to the pre-trained language models. Attention manipulation is one approach to introduce structural inductive bias without increasing the model size substantially. Studies that take this direction, however, by and large focus on incorporating syntax structure of sentences or internal structure of single documents (Bai et al., 2021; Cao and Wang, 2022) rather than higher level inter-document discourse structure. RAMMER is inspired by these works, and the novelty is that it uses attention manipulation to capture broader inter-document relationships.

3 The PEERSUM Dataset

3.1 Dataset Construction

PEERSUM is constructed using peer-review data scraped from OpenReview[2] for two international conferences in computer science: ICLR and NeurIPS. As meta-reviewers are supposed to follow the meta-reviewer guidelines[3] with comprehending and carefully summarizing information shown in the peer-reviewing web page (the example shown in Appendix A), and we observe from example meta-reviews as shown in Table 3 that meta-reviewers are complying with the guidelines, we collate the paper abstract, official/public reviews and multi-turn discussions as the source documents, and use the meta-review as the summary. We note that there may be private discussion among the reviewers and meta-reviewer which may influence the meta-review. However, our understanding is that reviewers are advised to amend their reviews if such a discussion changes their initial opinion. For this reason, we believe the meta-review is reflective of the (observable) reviews, discussions and the paper abstract, and this is empirically validated in Section 3.3.

A meta-review (summary) and its corresponding *source documents* (i.e., reviews, discussions and the paper abstract) form a sample in PEERSUM.[4] The source documents has an explicit tree-like conversational structure,[5] as illustrated in Figure 1 (a real example is presented in Appendix A). In total, PEERSUM contains 14,993 samples (train/validation/test: 11,995/1,499/1,499) for ICLR 2018–2022 and NeurIPS 2021–2022; see Table 1 for some statistics. To summarise, PEERSUM has seven types of source documents (shown in different colors in Figure 1): (1) official reviews (reviews by assigned reviewers); (2) public reviews (comments by public users); (3) author comments (an overall response by paper authors); (4) official responses; (5) public responses; (6) author responses within a thread; and (7) the paper abstract. It also features some metadata for each sample: (1) paper acceptance outcome (accept or reject); and (2) a rating (1–10) and confidence (1–5) for each official review.

To compare PEERSUM with other MDS datasets, we present some statistics on sample size and document length for PEERSUM and several other MDS datasets in Table 2.

We next present some analyses to understand the degree of conflicts in the source documents, and abstractiveness and faithfulness in the summaries.

3.2 Conflicts in Source Documents

One interesting aspect of PEERSUM is that source documents are not only featuring explicit hierarchical conversational relationships but also presenting conflicting information or viewpoint occasionally such as conflicting sentiments shown in Table 4. We extract *conflicts* among source documents based on review ratings in different official reviews. Denoting CF for samples with conflicts where at least one pair of official reviews that have a rating difference ≥ 4 (otherwise Non-CF), we found that 13.6% of the dataset are CF samples. The meta-reviews for these instances will need to handle these conflicts. In our experiments (Section 5) we present some results to show whether summarization systems are able to recognize and resolve conflicts in these difficult cases.

[2] https://openreview.net/

[3] https://iclr.cc/Conferences/2022/ACGuide, https://nips.cc/Conferences/2022/AC-Guidelines

[4] Henceforth we use the terms *summary* and *meta-review* interchangeably in the context of discussion of PEERSUM.

[5] The average tree height/width = 3.63/5.31.

Metric	PEERSUM	WikiSum	Multi-News	WCEP	Multi-XScience
Domain	Peer-review	Wikipedia	News	News	Scientific
#Samples	15,983	1,655,709	56,216	10,200	40,528
#Documents/Sample	10.48	40	2.79	63.38	4.45
#Sentences/Document	19.66	2.85	30.40	18.24	7.10
#Tokens/Document	397.32	54.54	690.97	439.24	172.90
#Sentence/Summary	6.51	5.17	10.12	1.44	5.06
#Tokens/Summary	142.74	121.20	241.61	30.53	116.41

Table 2: Statistics of PEERSUM and other MDS datasets.

M1	"This meta-review is written after considering the reviews, the authors' responses, the discussion, and the paper itself."
M2	"... the authors made substantial improvements during the discussion phase ..."
M3	"... but the bar for introducing yet another variant of memory-augmented neural nets has been significantly raised, which is a sentiment shared by the reviewers. the author's response had not swayed the reviewers' opinion, and i am sticking to the reviewers' decisions. ..."

Table 3: Three example meta-reviews (M1, M2, and M3) of meta-review sentences to show that the meta-reviewer is trying to comprehend and carefully summarize information from the paper, the individual reviews, and multi-turn discussions between paper authors and reviewers.

P1	S1: The approach proposed in the paper seems to be a small incremental change on top of the previous GNN pre-train work. *The novelty aspect is low*. S2: The main contribution is *the novel pre-training strategy* introduced. The work has *potential high impact* in the research area...
P2	S1: Introduction section is *not well-written*. S2: This paper is *well written* and looks correct.

Table 4: Two example pairs (P1 and P2) of contradictory sentiments between official reviewers for two scientific papers, and italic texts are conflicts between the two sentences (S1 and S2).

Dataset	Unigram	Bigram	Trigram
PEERSUM	28.28	82.31	92.95
WikiSum	22.75	63.55	79.34
Multi-News	23.49	66.10	82.01
WCEP	5.25	37.62	65.27
Multi-XScience	**44.09**	**86.54**	**96.40**

Table 5: Percentage of novel n-grams in the summaries of different datasets.

3.3 Abstractiveness and Faithfulness of Summaries

Abstractiveness — the degree that a summary contains novel word choices and paraphrases — is an important quality for MDS datasets. Following Fabbri et al. (2019) and Ghalandari et al. (2020), we preprocess source documents and summaries with lemmatisation and stop-word removal, and calculate the percentage of unigrams, bigrams, and trigrams in the summaries that are not found in the source documents and present the results in Table 5. We see that PEERSUM summaries are highly abstractive, particularly for bigrams and trigrams. Although Multi-XScience is the most abstractive, as discussed in Section 2.1 the summaries are not always reflective of the content of source documents.

To understand whether the summaries in PEERSUM are faithful to the source documents, i.e., whether the statements/assertions in the meta-review are grounded in the source documents, we perform manual analysis to validate this. We recruit 10 volunteers to annotate 60 samples (25 `Non-CF` and 35 `CF`) to highlight text spans in the summary that can be semantically anchored to the source documents (full instructions for the task is given in Appendix C).[6] Based on the results in Table 6, we can see that for samples with non-conflicting reviews (first row), almost 80% of the words in the meta-reviews are grounded in the source documents. Although this percentage drops to 72% when we are looking at the more difficult cases with conflicting reviews (second row), our analysis reveals that the meta-reviews are by and large faithful, indicating that they function as a good summary of the reviews, discussions and the paper abstract.

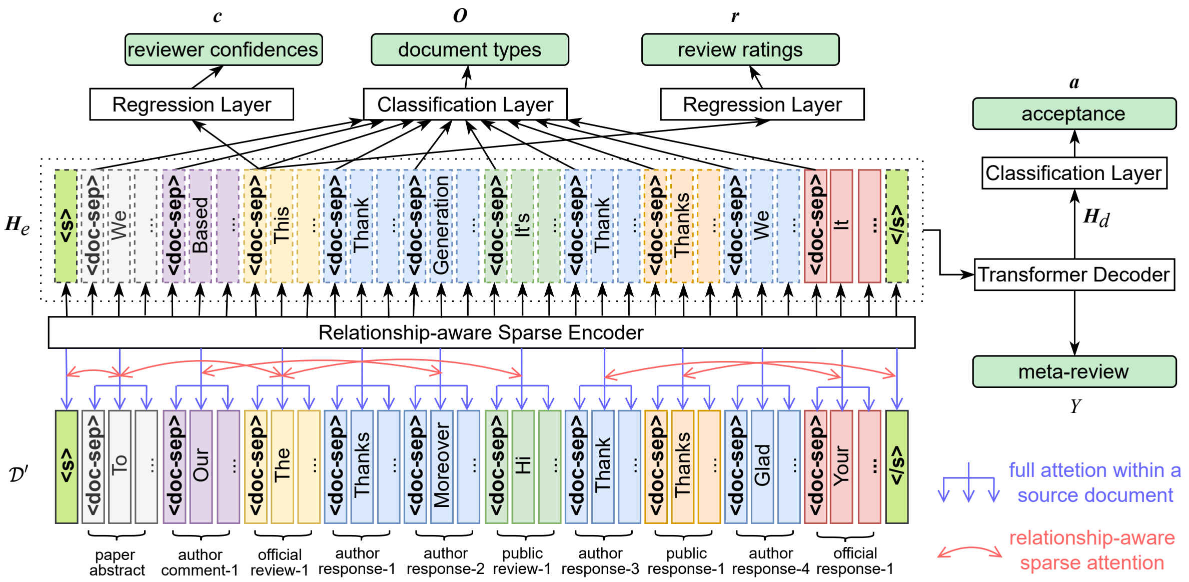

c
o
r
reviewer confidences
document types
review ratings
Regression Layer
Classification Layer
Regression Layer
a
acceptance
Classification Layer
H_d
Transformer Decoder
meta-review
Y
H_e
<s> <doc-sep> We … <doc-sep> Based … <doc-sep> This … <doc-sep> Thank … <doc-sep> Generation … <doc-sep> It's … <doc-sep> Thank … <doc-sep> Thanks … <doc-sep> We … <doc-sep> It … </s>
Relationship-aware Sparse Encoder
$\mathcal{D}'$
<s> <doc-sep> To … <doc-sep> Our … <doc-sep> The … <doc-sep> Thanks … <doc-sep> Moreover … <doc-sep> Hi … <doc-sep> Thank … <doc-sep> Thanks … <doc-sep> Glad … <doc-sep> Your … </s>
paper abstract
author comment-1
official review-1
author response-1
author response-2
public review-1
author response-3
public response-1
author response-4
official response-1
full attetion within a source document
relationship-aware sparse attention

Figure 2: There are six main components in the RAMMER architecture: (1) a relationship-aware sparse encoder to encode source documents; (2) a vanilla Transformer decoder to generate meta-reviews; (3) two different regression layers to predict reviewer confidences and review ratings; (4) two classification layers to predict the type of each source document and the paper acceptance outcome. There are two different types of attention mechanisms: 'full attention within a document' denotes that there are attention calculation between tokens within a document and 'relationship-aware sparse attention' denotes that there are attention calculation between tokens in documents only when there is a connection between the two documents in the corresponding tree-like hierarchical structure.

Data	#Samples	Mean Variance	Anchored Words (%)
Non-CF	25	0.717	79.67%
CF	35	6.668	72.74%

Table 6: Percentage of words in the meta-review grounded in the source documents in CF and Non-CF samples. "Mean Variance" denotes the average of rating variance of official reviews.

4 The RAMMER Model

We now describe RAMMER, a meta-review generation model that captures the conversational structure in the source documents (Section 4.1) and uses a multi-task objective to leverage metadata information (Section 4.2). RAMMER is built on an encoder-decoder PLM to automatically generate a summary/meta-review Y from a cluster of source documents $\mathcal{D}$; its overall architecture is presented in Figure 2. The input to RAMMER is the concatenation of all source documents ($\mathcal{D}$) and we insert a delimiter <doc-sep> to denote the start of each document.[7]

4.1 Relationship-Aware Sparse Attention

To explicitly incorporate hierarchical relationships among source documents into the pretrained Transformer model, we propose an encoder with relationship-aware sparse attention (RSAttn), which improves the summarization performance with the introduction of structural inductive bias. The main idea is to use sparse attention by considering hierarchical conversational relationships among source documents.

Based on the tree-like hierarchical conversational structure and the nature of meta-review generation, we extract seven types of relationships which are represented as matrices (an element is 1 if one document is connected to another, else 0):

- $\boldsymbol{R}_1$, *ancestor-1* which captures the parent asymmetric relationship and the attention from the parent document towards to the current one;
- $\boldsymbol{R}_2$, *ancestor-all* which captures the ancestor asymmetric relationship as the ancestor documents would provide context for the current one;
- $\boldsymbol{R}_3$, *descendant-1* which captures child asymmetric relationship and the attention from the child document towards to the current one;
- $\boldsymbol{R}_4$, *descendant-all* which captures descen-

[6] All volunteers are PhD students who major in computer science and are familiar with peer-reviewing.

[7] For PLMs that do not have <doc-sep> in their tokenizers we use </s> instead.

dant asymmetric relationship as sometimes concerns would be addressed after the discussion in descendant documents;

- $\boldsymbol{R}_5$, *siblings* which captures the sibling symmetric relationship as usually reviewers or the paper authors use sibling documents to provide more complementary information;
- $\boldsymbol{R}_6$, *document-self* which captures the full self-attention among each individual document as token representations are learned based on a rich context within the document;
- $\boldsymbol{R}_7$, *same-thread* which captures the symmetric relationship among documents which are in the same thread (source documents in each grey dashed rectangle in Figure 1) as usually documents in the same thread are talking about the same content.

Next, we use a weighted combination of these relationship matrices to mask out connections to those source documents not included in any relationships for each source document and scale the attention weights. The output of each head in RSAttn in the l-th layer is calculated as:

$$\boldsymbol{H}_l = \text{softmax}(\frac{\boldsymbol{QK}^T \odot \sum_j \beta_j \cdot \boldsymbol{R}_j^\dagger}{\sqrt{d_k}})\boldsymbol{V}, \quad (1)$$

where $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ are representations after the non-linear transformation of $\boldsymbol{H}_{l-1}$, the output of the previous layer, or $\boldsymbol{X}$, the output of the embedding layer from the input $\mathcal{D}'$ with delimiter tokens; $\boldsymbol{\beta}$ is a very small-scale trainable balancing weight vector for different relationships initialized with a uniform distribution, and different heads have different $\boldsymbol{\beta}$, as different heads in each layer may focus on different relationships; $\boldsymbol{R}_j^\dagger$ is automatically extended from $\boldsymbol{R}_j$ (if an element of $\boldsymbol{R}_{j,p,q}$ is 1, elements of $\boldsymbol{R}_j^\dagger$ from tokens of the p-th document to tokens of the q-th one are 1, else 0.).

To reduce memory consumption, we implement masking with matrix block multiplication instead of whole attention masking matrices, which means that we only calculate attention weights between every two documents that have at least one relation. This makes the model work for long source documents without substantially increasing computation complexity.

4.2 Multi-Task Learning

To utilise metadata information in PEERSUM — review rating, review confidence, paper acceptance outcome and source document type (Section 3.1) — we train RAMMER on four auxiliary tasks. We use the output embeddings from the encoder to predict review ratings/confidences and source document types, and the output embeddings from the decoder to predict the paper acceptance outcome. Formally, the overall training objective is:

$$\mathcal{L} = \alpha_g \mathcal{L}_g + \alpha_c \mathcal{L}_c + \alpha_r \mathcal{L}_r + \alpha_o \mathcal{L}_o + \alpha_a \mathcal{L}_a \quad (2)$$

where α is used to balance different objectives, $\mathcal{L}_g$ the standard cross-entropy loss for text generation based on the reference meta-review, $\{\mathcal{L}_c, \mathcal{L}_r\}$ the mean squared error for predicting the review confidence and review rating respectively, and $\{\mathcal{L}_o, \mathcal{L}_a\}$ the cross-entropy loss for predicting the paper acceptance outcome and the source document type respectively. Next, we describe more details about auxiliary tasks for the encoder and the decoder.

4.2.1 Encoder Auxiliary Tasks

We use $\mathcal{I}^d$ to denote the set of indices containing the special delimiters in the input. Auxiliary objectives of multi-task learning for the encoder are then based on the embeddings of these delimiters. Denoting the output embeddings produced by RAMMER's encoder as $\boldsymbol{H}_e$, we use two regression layers to predict the review confidences $\hat{\boldsymbol{c}}$ and review ratings $\hat{\boldsymbol{r}}$ respectively:

$$\hat{\boldsymbol{c}}_i = \text{sigmoid}(\text{MLP}(\boldsymbol{H}_e^{\mathcal{I}_i^d})), \quad (3)$$

$$\hat{\boldsymbol{r}}_i = \text{sigmoid}(\text{MLP}(\boldsymbol{H}_e^{\mathcal{I}_i^d}) \quad (4)$$

where $\mathcal{I}_i^d$ denotes the index of the delimiter token of the i-th official review, and $\boldsymbol{H}_e^{\mathcal{I}_i^d}$ denotes the corresponding embedding. $\mathcal{L}_c$ and $\mathcal{L}_r$ are then computed as:

$$\mathcal{L}_c = \text{mse}(\hat{\boldsymbol{c}}, \boldsymbol{c}), \quad \mathcal{L}_r = \text{mse}(\hat{\boldsymbol{r}}, \boldsymbol{r}), \quad (5)$$

where mse denotes mean squared error, and $\boldsymbol{c}$ and $\boldsymbol{r}$ denote the *normalised* ($[0-1]$) vector of ground truth review confidences and the review ratings, respectively.

To predict the types of source documents we apply a classification layer on the contextual embeddings of its delimiter tokens. The predicted classification distribution $\hat{\boldsymbol{O}}_j$ of the j-th source document is computed as follows:

$$\hat{\boldsymbol{O}}_j = \text{softmax}(\text{MLP}(\boldsymbol{H}_e^{\mathcal{I}_j^d})), \quad (6)$$

where $\boldsymbol{H}_e^{\mathcal{I}_j^d}$ denotes the embedding of the j-th source document and $\mathcal{I}_j^d$ is the corresponding index

Initialization	R-L	BERTS	ACC
BART	27.51	15.57	0.738
PRIMERA	29.30	13.24	0.745
LED	30.31	17.35	0.759

Table 7: RAMMER performance when initialized with different pre-trained language models.

in $\mathcal{I}^d$. The total loss for predicting all the document types, $\mathcal{L}_o$, is:

$$\mathcal{L}_o = \frac{1}{|\mathcal{I}^d|}\sum_{j=1}^{|\mathcal{I}^d|}\text{cross-entropy}(\boldsymbol{O}_j, \hat{\boldsymbol{O}}_j), \qquad (7)$$

where $\boldsymbol{O}_j$ is the one-hot embedding of the ground truth document type of the j-th source document.

4.2.2 Decoder Auxiliary Tasks

There is only one auxiliary objective for the decoder, to predict the paper acceptance outcome (accept vs. reject):

$$\hat{\boldsymbol{a}} = \text{MLP}(\text{mean}(\boldsymbol{H}_d)), \qquad (8)$$

$$\mathcal{L}_a = \text{cross-entropy}(\hat{\boldsymbol{a}}, \boldsymbol{a}), \qquad (9)$$

where $\boldsymbol{H}_d$ is the output embeddings from the last layer of the decoder and $\boldsymbol{a}$ is the one-hot embedding of the ground truth paper acceptance.

5 Experiments

5.1 Experimental Setup

We compare RAMMER with a suite of strong abstractive text summarization models.[8] We have three groups of models that target different types of summarization:[9] (1) short single-document: **BART** (Lewis et al., 2020) and **PEGASUS** (Zhang et al., 2020a); (2) long single-document: **LED** (Beltagy et al., 2020) and **PegasusX** (Phang et al., 2022); and (3) multi-document: **PRIMERA** (Xiao et al., 2022). We use the large variant for these models (which have a similar number of parameters). We fine-tune these models on PEERSUM using the default recommended hyper-parameter settings. All models have the same maximum output tokens (512), but they feature different budgets of the maximum input length. Given a sequence length budget, for each sample we divide the budget by the total number of source documents to get the maximum length permitted for each document and truncate each document based on that length. During training of RAMMER, we use a batch size of 128 with gradient accumulation and label smoothing of 0.1 (Müller et al., 2019). We tune RAMMER's α (Section 4.2) using the validation partition and the optimal configuration is: $\alpha_g = 2, \alpha_c = 2, \alpha_r = 1, \alpha_o = 1, \alpha_a = 2$, indicating that all metadata benefit the final performance and the reviewer confidence and paper acceptance outcome are the more important features. We present more details on training and hyper-parameter configuration in Appendix B.

5.2 Automatic Evaluation on Generated Meta-Reviews

We evaluate the quality of generated meta-reviews with metrics including ROUGE (Lin and Hovy, 2003),[10] BERTScore (Zhang et al., 2020b)[11] and UniEval (Zhong et al., 2022)[12]. ROUGE and BERTScore measure the lexical overlap between the generated and ground truth summary, but the former uses surface word forms and latter contextual embeddings. UniEval achieves fine-grained evaluation for abstractive summarization and it is based on framing evaluation of text generation as a boolean question answering task. As faithfulness and informativeness are more important to summarization, we only report the evaluation results of "consistency" and "relevance" from UniEval, respectively.

In addition to these metrics, we introduce another evaluation metric (ACC) based on the metadata of PEERSUM. It is an alternative reference-free metric that measures how well generated meta-reviews are consistent with the ground truth meta-reviews. To this end, we first fine-tune a BERT-based classifier using *ground truth* meta-reviews and paper acceptance outcomes, and then use this classifier to predict the paper acceptance outcome using *generated* meta-reviews. The idea is that if the generated meta-review is consistent with the ground truth meta-review, the predicted paper acceptance outcome should match the ground truth paper acceptance outcome.

As RAMMER can use any encoder-decoder pre-trained models as the backbone, we first present *val-*

[8] All experiments are run on 4 NVIDIA 80G A100 GPUs.

[9] We fine-tune these pre-trained models on PEERSUM with the Huggingface library (https://huggingface.co/).

[10] For ROUGE-L, we use the summary-level version 'RougeLsum' from https://pypi.org/project/rouge-score/.

[11] Following Koto et al. (2020), we use F1 metrics of ROUGE and BERTScore.

[12] https://github.com/maszhongming/UniEval

Model(#Params)	Test Data	R-L↑	BERTS↑	UniEval-Con↑	UniEval-Rel↑	ACC↑
BART (406M)	Non-CF	27.50	16.61	72.97	79.87	0.728
PEGASUS (568M)	Non-CF	27.24	14.75	74.52	80.78	0.725
PRIMERA (447M)	Non-CF	28.70	12.67	68.56	82.33	0.725
LED (459M)	Non-CF	29.52	16.59	70.98	82.97	0.748
PegasusX (568M)	Non-CF	29.65	17.36	73.44	82.24	0.745
RAMMER (459M)	Non-CF	**30.39***	**17.42***	**75.07***	**83.84***	**0.768**
BART (406M)	CF	26.84	14.89	71.85	78.74	0.683
PEGASUS (568M)	CF	26.77	13.66	73.12	79.49	0.649
PRIMERA (447M)	CF	29.13	12.33	66.85	81.70	0.639
LED (459M)	CF	29.19	15.32	70.04	82.82	0.698
PegasusX (568M)	CF	**29.30**	15.69	71.33	81.30	0.707
RAMMER (459M)	CF	29.19	**15.88***	**73.21***	**83.15***	**0.724**
RAMMER (459M)	CF ∪ Non-CF	30.23	17.21	74.82	83.75	0.762
w/o RSAttn (406M)	CF ∪ Non-CF	29.67	16.88	71.36	83.01	0.758
w/o multi-task (406M)	CF ∪ Non-CF	30.27	17.01	72.99	83.57	0.749

Table 8: Performance of summarization models over PEERSUM in terms of ROUGE-L F1 (R-L), BERTScore F1 (BERTS), UniEval consistency (UniEval-Con) and relevance (UniEval-Rel) and paper outcome (ACC). Higher value means better performance for all metrics. Results of ROUGE-1 and ROUGE-2 which are not present are consistent with that of ROUGE-L. *: significantly better than others in the same group (p-value < 0.05).

idation results for RAMMER where it is initialised with BART, PRIMERA and LED with different maximum input lengths (1,024, 4,096 and 4,096, respectively) in Table 7. We see consistently that the LED variant performs better than the other two, and this helps us choose the LED variant as the backbone of RAMMER. This also indicates that our idea of RSAttn and multi-task learning with metadata can also work on other pre-trained language models.

We next compare RAMMER with the baseline text summarization models on the *test* set in Table 8[13]. Here we also break the test partition into CF and Non-CF samples. Broadly speaking, summarisation performance across all metrics for CF is lower than that of Non-CF, confirming our suspicion that the CF instances are more difficult to summarise. The disparity is especially significant for ACC and BERTScore, suggesting that these two are perhaps the better metrics for evaluate the quality of generated meta-reviews. Comparing RAMMER with the baselines, it is encouraging to see that it is consistently better (exception: R-L results of most models on CF samples are more or less the same). This demonstrates the benefits of incorporating the conversational structure and metadata in the source documents into pre-trained language models. To better understand the impact of RAMMER's sparse attention (RSAttn; Section 4.1) and multi-task objective (Section 4.2), we also present two RAMMER ablation variants (the last three rows in Table 8). It is an open question which method has more impact, as even though most metrics (R-L, BERTScore and UniEval) seem to indicate RSAttn is the winner, ACC — which we believe is the most reliable metric — appear to suggest otherwise. That said, we can see they complement with each other and as such incorporating both produces the best performance.

5.3 Human Evaluation on Conflict Recognition and Resolution

To dive deeper into understanding how well these summarization models recognize and resolve conflicting information in source documents, we conduct a human evaluation.

We randomly select 40 CF samples and recruit two volunteers[14]. We ask them to first assess whether each *ground truth* meta-review **recognises** conflicts, i.e., whether the meta-review discusses or mentions conflicting information/viewpoints that are in the official reviews. For each sample, the volunteers are presented with all the source documents and are asked to make a binary judgement about conflict recognition. We found that 23 out of 40 ground truth meta-reviews have successfully

[13] Some random examples and corresponding model generations are present in Appendix D.

[14] Both volunteers major in computer science and are familiar with peer-reviewing.

Model	Recognition	Resolution
PRIMERA	3/23	2/23
LED	4/23	4/23
PegasusX	5/23	5/23
RAMMER	8/23	3/23

Table 9: Performances of summarization models on conflict recognition and resolution for CF samples.

done this, and we next focus on assessing generated meta-reviews for these remaining 23 samples.

For these 23 samples, we ask the volunteers to assess conflict recognition for *generated* meta-reviews. Additionally, we also ask them to judge (binary judgement) whether the generated meta-review **resolves** the conflicts in a similar manner consistent with the ground truth meta-review. Conflict recognition and revolution results for RAMMER and three other baselines are presented in Table 9. In terms of recognition, relatively speaking RAMMER does better than the baselines which is encouraging, but ultimately it still fails to recognise conflicts in majority of the samples. When it comes to conflict resolution, all the models perform very poorly, indicating the challenging nature of resolving conflicts in source documents of PEERSUM.

6 Conclusion

We introduce PEERSUM, an MDS dataset for meta-review generation. PEERSUM is unique in that the summaries (meta-reviews) are grounded in the source documents despite being highly abstractive, it has a rich set of metadata and explicit inter-document structure, and it features explicit conflicting information in source documents that the summaries have to handle. In terms of modelling, we propose RAMMER, an approach that extends Transformer-based pre-trained encoder-decoder models to capture inter-document relationships (through the sparse attention) and metadata information (through the multi-task objective). Although RAMMER is designed for meta-review generation here, our approach of manipulating attention to incorporate the input structure can be easily adapted to other tasks where the input has inter-document relationships. Compared with baselines over a suite of automatic metrics and human evaluation, we found that RAMMER performs favourably, outperforming most strong baselines consistently. That said, when we assess how well RAMMER does for situations where there are conflicting information/viewpoints in the source documents, the outlook is less encouraging. We found that RAMMER fail to recognise and resolve these conflicts in its meta-reviews in the vast majority of cases, suggesting this is a challenging problem and promising avenue for further research.

Limitations

Our work frames meta-review generation as an MDS problem, but one could argue that writing a meta-review requires not just summarising key points from the reviews, discussions and the paper abstract but also wisdom from the meta-reviewer to judge opinions. We do not disagree, and to understand the extent to which the meta-review can be "generated" based on the source documents we conduct human assessment (Section 3.3) to validate this. While the results are encouraging (as we found that most of the content in the meta-reviews are grounded in the source documents) the approach we took is a simple one, and the assessment task can be further improved by decomposing it into subtasks that are more objective (e.g., by explicitly asking annotators to link statements in the meta-reviews to sentences in the source documents).

In the age of ChatGPT and large language models, there is also a lack of inclusion of larger models for comparison. We do not believe it makes sense to include closed-source models such as ChatGPT for comparison (as it is very possible that they have been trained on OpenReview data), but it could be interesting to experiment with large open-source models such as OPT (Zhang et al., 2022), LLaMA (Touvron et al., 2023) or Falcon (Almazrouei et al., 2023). We contend, however, the results we present constitute preliminary results, and that it could be promising direction to explore how RAMMER's RSAttn can be adapted for large autoregressive models.

Lastly, we only consider explicit conversational structure in this paper. As our results show, incorporating such structure only helps to recognise conflicts to some degree but not for resolving them. It would be fascinating to test if incorporating *implicit* structure, such as argument and discourse links, would help. This is not explored in this paper, but it would not be difficult to adapt our methods to incorporate these structures.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

Jiangang Bai, Yujing Wang, Yiren Chen, Yaming Yang, Jing Bai, Jing Yu, and Yunhai Tong. 2021. Syntax-bert: Improving pre-trained transformers with syntax trees. In *EACL*, pages 3011–3020.

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

Shuyang Cao and Lu Wang. 2022. HIBRIDS: Attention with hierarchical biases for structure-aware long document summarization. In *ACL*.

Peng Cui and Le Hu. 2021. Topic-guided abstractive multi-document summarization. In *Findings of EMNLP*, pages 1463–1472.

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

Alexander R. Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir R. Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *ACL*, pages 1074–1084.

Demian Gholipour Ghalandari, Chris Hokamp, Nghia The Pham, John Glover, and Georgiana Ifrim. 2020. A large-scale multi-document summarization dataset from the wikipedia current events portal. In *ACL*, pages 1302–1308.

Mandy Guo, Joshua Ainslie, David C. Uthus, Santiago Ontañón, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2022. Longt5: Efficient text-to-text transformer for long sequences. In *Findings of NAACL*, pages 724–736.

Hanqi Jin, Tianming Wang, and Xiaojun Wan. 2020. Multi-granularity interaction network for extractive and abstractive multi-document summarization. In *ACL*, pages 6244–6254.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2020. FFCI: A framework for interpretable automatic evaluation of summarization. *CoRR*, abs/2011.13662.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*, pages 7871–7880.

Miao Li, Jianzhong Qi, and Jey Han Lau. 2023. Compressed heterogeneous graph for abstractive multi-document summarization. In *AAAI*.

Wei Li, Xinyan Xiao, Jiachen Liu, Hua Wu, Haifeng Wang, and Junping Du. 2020. Leveraging graph to improve abstractive multi-document summarization. In *ACL*, pages 6232–6243.

Chin-Yew Lin and Eduard H. Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *HLT-NAACL*, pages 71–78.

Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. In *ICLR*.

Yao Lu, Yue Dong, and Laurent Charlin. 2020. Multi-xscience: A large-scale dataset for extreme multi-document summarization of scientific articles. In *EMNLP*, pages 8068–8074.

Congbo Ma, Wei Emma Zhang, Mingyu Guo, Hu Wang, and Quan Z. Sheng. 2020. Multi-document summarization via deep learning techniques: A survey. *CoRR*, abs/2011.04843.

Rafael Müller, Simon Kornblith, and Geoffrey E. Hinton. 2019. When does label smoothing help? In *NeurIPS*, pages 4696–4705.

Jason Phang, Yao Zhao, and Peter J. Liu. 2022. Investigating efficiently extending transformers for long input summarization. *CoRR*, abs/2208.04347.

Chenhui Shen, Liying Cheng, Ran Zhou, Lidong Bing, Yang You, and Luo Si. 2022. Mred: A meta-review dataset for structure-controllable text generation. In *Findings of ACL*, pages 2521–2535.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*, pages 5998–6008.

Po-Cheng Wu, An-Zi Yen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2022. Incorporating peer reviews and rebuttal counter-arguments for meta-review generation. In *CIKM*, pages 2189–2198.

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2022. PRIMERA: pyramid-based masked sentence pre-training for multi-document summarization. In *ACL*, pages 5245–5263.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020a. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In *ICML*, pages 11328–11339.

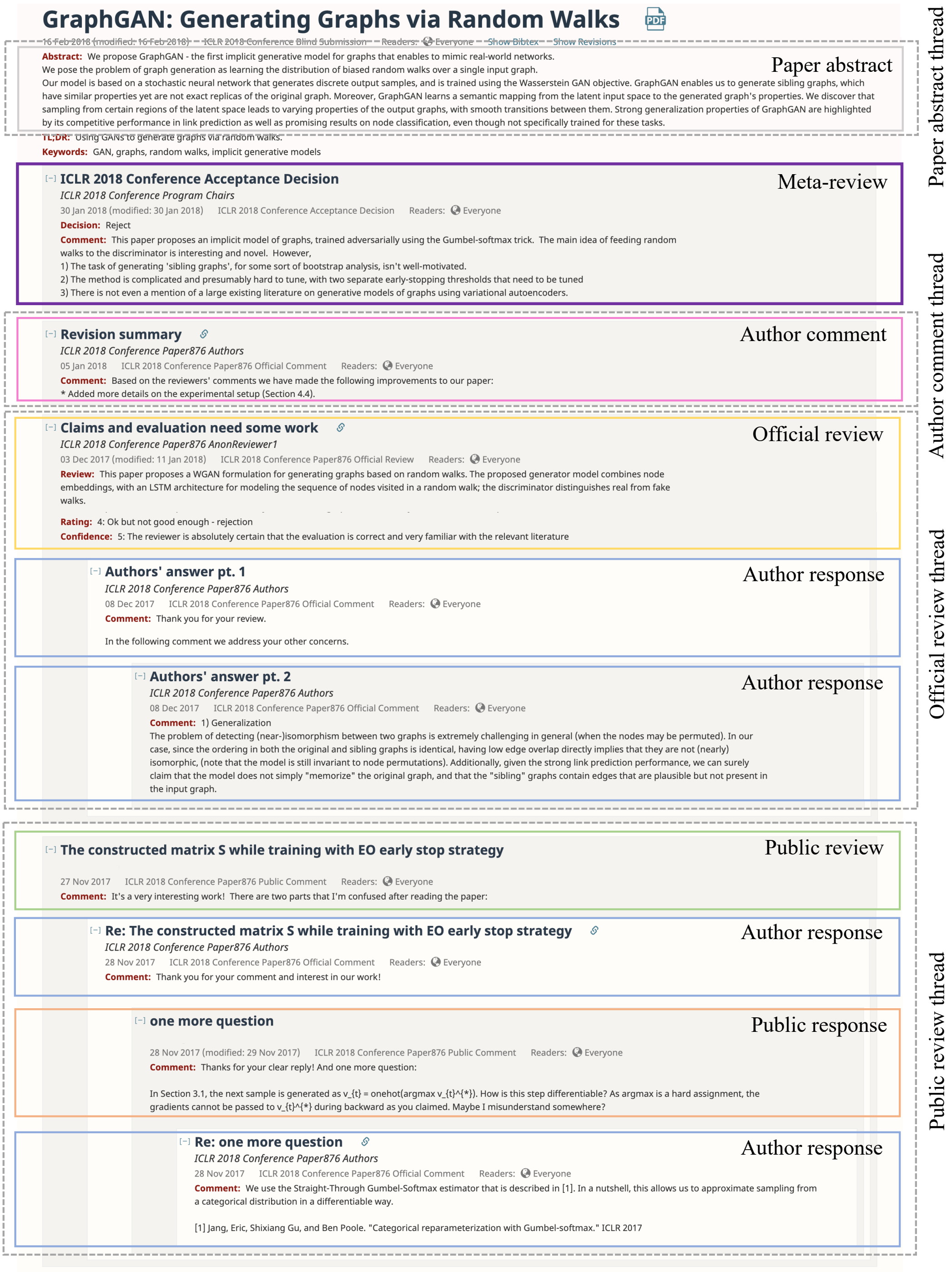
Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: open pre-trained transformer language models. *CoRR*, abs/2205.01068.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with BERT. In *ICLR*.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multidimensional evaluator for text generation. In *EMNLP*, pages 2023–2038.

A Appendix: Different types of source documents in PeerSum

We present a real PeerSum example with annotation in Figure 3. Please note that we randomly select this example from PeerSum and we have removed the author names of the paper and reviewer names, and the content in this real example is not the same as in the synthesized example in Figure 1, while both of the two feature hierarchical inter-document relationships. As shown in Figure 3, automatic meta-review generation is aiming to generate the meta-review automatically based on the paper abstract, official and public reviews and the multi-turn discussions.



GraphGAN: Generating Graphs via Random Walks

16 Feb 2018 (modified: 16 Feb 2018) ICLR 2018 Conference Blind Submission Readers: Everyone Show Bibtex Show Revisions

Abstract: We propose GraphGAN - the first implicit generative model for graphs that enables to mimic real-world networks.
We pose the problem of graph generation as learning the distribution of biased random walks over a single input graph.
Our model is based on a stochastic neural network that generates discrete output samples, and is trained using the Wasserstein GAN objective. GraphGAN enables us to generate sibling graphs, which have similar properties yet are not exact replicas of the original graph. Moreover, GraphGAN learns a semantic mapping from the latent input space to the generated graph's properties. We discover that sampling from certain regions of the latent space leads to varying properties of the output graphs, with smooth transitions between them. Strong generalization properties of GraphGAN are highlighted by its competitive performance in link prediction as well as promising results on node classification, even though not specifically trained for these tasks.

TL;DR: Using GANs to generate graphs via random walks.

Keywords: GAN, graphs, random walks, implicit generative models

[−] **ICLR 2018 Conference Acceptance Decision**

ICLR 2018 Conference Program Chairs

30 Jan 2018 (modified: 30 Jan 2018) ICLR 2018 Conference Acceptance Decision Readers: Everyone

Decision: Reject

Comment: This paper proposes an implicit model of graphs, trained adversarially using the Gumbel-softmax trick. The main idea of feeding random walks to the discriminator is interesting and novel. However,
1) The task of generating 'sibling graphs', for some sort of bootstrap analysis, isn't well-motivated.
2) The method is complicated and presumably hard to tune, with two separate early-stopping thresholds that need to be tuned
3) There is not even a mention of a large existing literature on generative models of graphs using variational autoencoders.

[−] **Revision summary**

ICLR 2018 Conference Paper876 Authors

05 Jan 2018 ICLR 2018 Conference Paper876 Official Comment Readers: Everyone

Comment: Based on the reviewers' comments we have made the following improvements to our paper:
* Added more details on the experimental setup (Section 4.4).

[−] **Claims and evaluation need some work**

ICLR 2018 Conference Paper876 AnonReviewer1

03 Dec 2017 (modified: 11 Jan 2018) ICLR 2018 Conference Paper876 Official Review Readers: Everyone

Review: This paper proposes a WGAN formulation for generating graphs based on random walks. The proposed generator model combines node embeddings, with an LSTM architecture for modeling the sequence of nodes visited in a random walk; the discriminator distinguishes real from fake walks.

Rating: 4: Ok but not good enough - rejection

Confidence: 5: The reviewer is absolutely certain that the evaluation is correct and very familiar with the relevant literature

[−] **Authors' answer pt. 1**

ICLR 2018 Conference Paper876 Authors

08 Dec 2017 ICLR 2018 Conference Paper876 Official Comment Readers: Everyone

Comment: Thank you for your review.

In the following comment we address your other concerns.

[−] **Authors' answer pt. 2**

ICLR 2018 Conference Paper876 Authors

08 Dec 2017 ICLR 2018 Conference Paper876 Official Comment Readers: Everyone

Comment: 1) Generalization
The problem of detecting (near-)isomorphism between two graphs is extremely challenging in general (when the nodes may be permuted). In our case, since the ordering in both the original and sibling graphs is identical, having low edge overlap directly implies that they are not (nearly) isomorphic, (note that the model is still invariant to node permutations). Additionally, given the strong link prediction performance, we can surely claim that the model does not simply "memorize" the original graph, and that the "sibling" graphs contain edges that are plausible but not present in the input graph.

[−] **The constructed matrix S while training with EO early stop strategy**

27 Nov 2017 ICLR 2018 Conference Paper876 Public Comment Readers: Everyone

Comment: It's a very interesting work! There are two parts that I'm confused after reading the paper:

[−] **Re: The constructed matrix S while training with EO early stop strategy**

ICLR 2018 Conference Paper876 Authors

28 Nov 2017 ICLR 2018 Conference Paper876 Official Comment Readers: Everyone

Comment: Thank you for your comment and interest in our work!

[−] **one more question**

28 Nov 2017 (modified: 29 Nov 2017) ICLR 2018 Conference Paper876 Public Comment Readers: Everyone

Comment: Thanks for your clear reply! And one more question:

In Section 3.1, the next sample is generated as v_{t} = onehot(argmax v_{t}^{*}). How is this step differentiable? As argmax is a hard assignment, the gradients cannot be passed to v_{t}^{*} during backward as you claimed. Maybe I misunderstand somewhere?

[−] **Re: one more question**

ICLR 2018 Conference Paper876 Authors

28 Nov 2017 ICLR 2018 Conference Paper876 Official Comment Readers: Everyone

Comment: We use the Straight-Through Gumbel-Softmax estimator that is described in [1]. In a nutshell, this allows us to approximate sampling from a categorical distribution in a differentiable way.

[1] Jang, Eric, Shixiang Gu, and Ben Poole. "Categorical reparameterization with Gumbel-softmax." ICLR 2017

Figure 3: A set of source documents and the corresponding meta-review for a scientific paper in PeerSum.

B Appendix: Hyper-parameters for fine-tuning pre-trained text summarization models

We present all hyper-parameters in Table 10 for all models.

Model	**Max-len(in/out)**	**optimizer**	**lr**	**warm up**	**scheduler**	**batch size**	**beam size**	**length penalty**
BART	1,024/512	Adafactor	3e-5	0k	constant schedule	64	5	1.0
PEGASUS	1,024/512	Adafactor	5e-5	0k	square root decay	256	8	0.8
PRIMERA	4,096/512	Adam	3e-5	0.5k	linear decay	16	5	1.0
LED	4,096/512	Adam	3e-5	0.2k	linear decay	32	5	0.8
PegasusX	4,096/512	Adafactor	8e-4	0k	constant schedule	64	1	1.0
RAMMER	4,096/512	Adafactor	5e-5	0.2k	linear decay	128	5	1.0

Table 10: Hyper-parameters for all models in experiments.

C Appendix: Instructions for annotation of PEERSUM quality

Welcome to the annotation project for PeerSum. Please have a careful read of the project introduction and task instructions and finish the tasks in the separate document.

Introduction of the project:

To enhance the capabilities of multi-document summarization systems we present PeerSum, a novel dataset for automatically generating meta-reviews of scientific papers based on reviews, multi-turn discussions and the paper abstract in the peer-reviewing process in https://openreview.net/. In the reviewing process, all assigned reviewers and public users can give comments to each paper, and then the author of the paper might respond to those comments. There may be a couple of rounds of discussions or rebuttals during the reviewing process. In the end, the meta-reviewer will write a summary of these comments and discussions, to support their final decision on the paper acceptance. Usually, meta-reviewers are supposed to write the meta-review based on summarizing all reviews, discussions, and the paper abstract, but they may sometimes draw on some external knowledge which is not present in the source documents, such as their own knowledge in the field, reading of the full paper beyond the paper abstract.

The objective of the annotation task is to assess whether the statements/assertions in meta-reviews are exclusively drawn from the reviews, discussion and the paper abstract which are the source documents in PeerSum. Annotators are expected to help highlight the statements/assertions in meta-reviews that can be drawn from the source documents. Highlighted texts will be heavily dependent on source documents and mainly talk about information that is present in the source documents, while they will not be heavily dependent on meta-reviewer's judgements, the meta-reviewer's own knowledge in the field, reading of the full paper beyond the paper abstract, or any other external knowledge relative to the source documents. Please note that if assertions have very light judgement from meta-reviewers but the content are mostly drawn from source documents, we will prefer to highlight these assertions, as these assertions usually cover much about critical information in the source documents.

Instructions for the task:

Each of you will get 6 samples in total. For each sample:

- Please carefully read the source documents including the paper abstract, reviews by different reviewers, and discussions between reviewers and the author (all responses) in the linked OpenReview page.
- Please read the meta-review which is the same as the section of Paper Decision in the corresponding OpenReview link, and highlight all assertions or statements (which may be a clause, a sentence, or a paragraph) which draws knowledge solely from source documents with the colour of blue.

Annotation examples:

Please also carefully read the following two examples of annotation tasks. We also prepare explanations for unhighlighted or highlighted texts following each example, but you do not need to write explanations when annotating.

Example one

Source Documents:

Link to OpenReview: https://github.com/oaimli/PeerSum/blob/main/examples/Hygy01StvH.pdf

Meta-review:

The reviewers have pointed out several major deficiencies of the paper, which the authors decided not to address.

Example two

Source Documents:

Link to OpenReview: https://github.com/oaimli/PeerSum/blob/main/examples/H1DkN7ZCZ.pdf

Meta-review:

Authors present a method for representing DNA sequence reads as one-hot encoded vectors, with genomic context (expected original human sequence), read sequence, and CIGAR string (match operation encoding) concatenated as a single input into the framework. Method is developed on 5 lung cancer patients and 4 melanoma patients. Pros: - The approach to feature encoding and network construction for task seems new. – The target task is important and may carry significant benefit for healthcare and disease screening. Cons: - The number of patients involved in the study is exceedingly small. Though many samples were drawn from these patients, pattern discovery may not be generalizable across larger populations. Though the difficulty in acquiring this type of data is noted. – (Significant) Reviewer asked for use of public benchmark dataset, for which authors have declined to use since the benchmark was not targeted toward task of ultra-low VAFs. However, perhaps authors could have sourced genetic data from these recommended public repositories to create synthetic scenarios, which would enable the broader research community to directly compare against the methods presented here. The use of only private datasets is concerning regarding the future impact of this work. – (Significant) The concatenation of the rows is slightly confusing. It is unclear why these were concatenated along the column dimension, rather than being input as multiple channels. This question doesn't seem to be addressed in the paper. Given the pros and cons, the committee recommends this interesting paper for workshop.

Explanations:

- "However, perhaps authors could have sourced genetic data from these recommended public repositories to create synthetic scenarios," is highlighted, because this assertion is logically based on recommended public repositories and synthetic scenarios which are from source documents.
- "which would enable the broader research community to directly compare against the methods presented here. The use of only private datasets is concerning regarding the future impact of this work." is not highlighted, because this is heavily based on meta-reviewer's own experience in the field or suggestion about impact of the paper.
- "This question doesn't seem to be addressed in the paper." is not highlighted, because it is a meta-reviewer's own judgement about the paper.
- In "Given the pros and cons, the committee recommends this interesting paper for workshop." which is not highlighted, there is external knowledge about the workshop information.

Example three

Source Documents:

Link to OpenReview: https://github.com/oaimli/PeerSum/blob/main/examples/ZeE81SFTsl.pdf

Meta-review:

Dear authors, I apologize to the authors for insufficient discussion in the discussion period. Thanks for carefully responding to reviewers. Nevertheless, I have read the paper as well, and the situation is clear to me (even without further discussion). I will not summarize what the paper is about, but will instead mention some of the key issues. 1) The proposed idea is simple, and in fact, it has been known to me for a number of years. I did not think it was worth publishing. This on its own is not a reason for rejection, but I wanted to mention this anyway to convey the idea that I consider this work very incremental. 2) The idea is not supported by any convergence theory. Hence, it remains a heuristic, which the authors admit. In such a case, the paper should be judged by its practical performance, novelty and efficacy of ideas, and the strength of the empirical results, rather than on the theory. However, these parts of the paper remain lacking compared to the standard one would expect from an ICLR paper. 3) Several elements of the ideas behind this work existed in the literature already (e.g., adaptive quantization, time-varying quantization, ...). Reviewers have noticed this. 4) The authors compare to fixed / non-adaptive quantization strategies which have already been surpassed in subsequent work. Indeed, QSGD was developed 4 years ago. The quantizers of Horvath et al in the natural compression/natural dithering family have exponentially better variance for any given number of levels. This baseline, which does not use any adaptivity, should be better, I believe, to what the author propose. If not, a comparison is needed. 5) FedAvg is not the theoretical nor practical SOTA method for the problem the authors are solving. Faster and more communication efficient methods exist. For example, method based on error feedback (e.g., the works of Stich, Koloskova and others), MARINA method (Gorbunov et al), SCAFFOLD (Karimireddy et al) and so on. All can be combined with quantization. 6) The reviewer who assigned this paper score 8 was least confident. I did not find any comments in the review of this reviewer that would sufficiently justify the high score. The review was brief and not very informative to me as the AC. All other reviewers were inclined to reject the paper. 7) There are issues in the mathematics – although the mathematics is simple and not the key of the paper. This needs to be thoroughly revised. Some answers were given in author response. 8) Why should expected variance be a good measure? Did you try to break this measure? That is, did you try to construct problems for which this measure would work worse than the worst case variance? Because of the above, and additional reasons mentioned in the reviewers, I have no other option but to reject the paper. Area Chair

Explanations:

- In this meta-review, the meta-reviewer write it based on own reading of the full paper. In this kind of cases, meta-reviewers draw on external knowledge, but some of the assertions are still based on source documents, such as “All other reviewers were inclined to reject the paper”.
- “Dear authors, I apologize to the authors for insufficient discussion in the discussion period. Thanks for carefully responding to reviewers.” is not highlighted, because this is coordination words and some own judgements from the meta-reviewer.
- “Nevertheless, I have read the paper as well, and the situation is clear to me (even without further discussion).” is not highlighted, because this is based on meta-reviewer’s own reading of the full paper.
- “I will not summarize what the paper is about, but will instead mention some of the key issues.” is not highlighted, because this is coordination words from the meta-reviewer.
- “1) The proposed idea is simple, and in fact, it has been known to me for a number of years. I did not think it was worth publishing. This on its own is not a reason for rejection, but I wanted to mention this anyway to convey the idea that I consider this work very incremental.” is not highlighted, because this is based on the meta-reviewer’s own experience.
- “In such a case, the paper should be judged by its practical performance, novelty and efficacy of ideas, and the strength of the empirical results, rather than on the theory. However, these parts of the paper remain lacking compared to the standard one would expect from an ICLR paper.” is not highlighted, because this is the meta-reviewer’s experience about the standard of ICLR.
- “which have already been surpassed in subsequent work. Indeed, QSGD was developed 4 years ago. The quantizers of Horvath et al in the natural compression/natural dithering family have exponentially better variance for any given number of levels. This baseline, which does not use any adaptivity, should be better, I believe, to what the author propose. If not, a comparison is needed.” is not highlighted, because this is based on the meta-reviewer’s experience in the field.
- “5) FedAvg is not the theoretical nor practical SOTA method for the problem the authors are solving. Faster and more communication efficient methods exist. For example, method based on error feedback (e.g., the works of Stich, Koloskova and others), MARINA method (Gorbunov et al), SCAFFOLD (Karimireddy et al) and so on. All can be combined with quantization.” is not

highlighted, because this is based on the meta-reviewer's experience in the field.

- "I did not find any comments in the review of this reviewer that would sufficiently justify the high score. The review was brief and not very informative to me as the AC." is not highlighted, because this is meta-reviewer's judgement on the review.
- "7) There are issues in the mathematics – although the mathematics is simple and not the key of the paper. This needs to be thoroughly revised.", this is not highlighted because it is based on meta-reviewer's reading of the full paper.
- "8) Why should expected variance be a good measure? Did you try to break this measure? That is, did you try to construct problems for which this measure would work worse than the worst case variance? Because of the above, and", this is not highlighted because it is based on the meta-reviewer's own knowledge in the field.
- "I have no other option but to reject the paper. Area Chair" is not highlighted, as this is the meta-reviewer's judgement on the paper.

D Appendix: Generated meta-reviews for PEERSUM by different models

We present five groups of example meta-reviews generated by fully-supervised PRIMERA, LED, PegasusX, and RAMMER in Table 11 with the input of varying lengths, 1,024, 4,096, and 4,096, respectively, and also ROUGE scores measuring the quality of generated meta-reviews in comparison to the ground truth one. These examples are randomly selected from the test set of PEERSUM. It is clear to see that although RAMMER outperforms other strong baseline models in terms of evaluation metrics in Table 8, the quality of generated meta-reviews still needs to be improved. This further confirms our claim that PEERSUM is a really challenging dataset.

Example 1, `https://github.com/oaimli/PeerSum/blob/main/examples/e1.pdf`

Ground truth meta-review	Understanding neural networks once they have been trained is a big open problem for machine learning. This manuscript designed graph theoretic and information theoretic measures aimed at helping us understand community structure and function in trained networks. In particular, they measure community structure (modularity) and entropy for trained networks and related these to the performance of the networks. The manuscript runs experiments with fully connected networks on problems such as MNIST and CIFAR. Both community structure and entropy measures are shown to correlate (Spearman and Pearson correlation coefficients) with performance metrics in the networks studied. Reviewers tended to agree that the paper was well written and motivated by an interesting and timely question (understanding trained networks). However, on the whole, most of the reviewers believe that the manuscript is too preliminary for publication at ICLR and I agree. A central issue cited by most of the reviewers is that the experiments are performed on small/toy models for small tasks and under particular hyperparameter regimes. It is therefore unclear to what extent the results would generalize to other situations. E.g. would the results hold for larger dataset or for convolutional neural networks? Connected to this complaint, reviewers worry that there is not enough connection to the literature and baseline methods that could be used to predict performance given measures of trained network activity. Even allowing that the observed correlations are true and generalizable, are these measures better than those covered elsewhere in the literature? Additionally problematic, the measures are not theoretically justified either. Thus, we are missing both reasoned arguments for the metrics and robust quantification beyond a limitted experimental setting. One reviewer, Xmnm, is compelled by the work and recommends acceptance. However, they do not present a compelling case for acceptance, and even repeat several of the concerns raised by other reviewers. In sum, the work is on an interesting subject and timely, but needs further work to be ready for publication.
PRIMERA	This paper proposes two methods to analyze the behavior of neurons in neural nets. The main idea is to study the neuron activation patterns of classification models and explore if the performance can be explained through neurons' activation behavior. The authors propose two approaches: one that models neurons' activation behavior as a graph and examines whether the neurons form meaningful communities, and the other that examines the predictability of neurons' behavior using entropy. The paper is interesting, easy to read, and very well-structured. Furthermore, the subject of the paper is very relevant: the exploration of tools to increase the understanding and explainability of neural models. However, several major concerns are obvious: (1) The contribution of this paper is not properly compared against existing methods, making it difficult to tell their significance. (2) The scalability of this work is suspicious. There 's no guarantee that proposed activation analysis methods will also show the same correlation tendency in other neural networks with more complicated architectures than multi-layer perceptron networks. (3) Experiments on other various hyperparameters/conditions that can affect neuron configurations or learning ability of the neural network(e.g., pruning) could add more validities to their 'comprehensive experimental study'. (4) The suggested explanatory methodologies and approaches were pretty persuasive to explain the network behavior in microscopic context. Nevertheless, the scalability and practicality of the work should be examined carefully. (ROUGE-1: 0.37184, ROUGE-2: 0.04348, ROUGE-L: 0.34657)

LED	This paper proposes two new metrics to analyze the behavior of neurons in neural nets. The main idea is to study the neuron activationpatterns of classification models and explore if the performance can be explained through neurons' activation behavior. The authors propose two approaches: one that models neurons' activation behavior as a graph and examines whether the neurons form meaningful communities, and the other that examines the predictability of neurons' behavior using entropy. The reviewers agree that this is an interesting paper and the authors have done a good job in their rebuttal. However, the paper is not ready for publication at ICLR due to the following reasons: 1. There are many existing approaches in explaining deep learning models, such as turning deep models into decision trees in "Distilling a Neural Network Into a Soft Decision Tree". Then complexity measurement of the decision tree can be a baseline. It would be interesting to see how the techniques presented in this paper compare to other techniques already present in the literature. (ROUGE-1: 0.36032, ROUGE-2: 0.07317, ROUGE-L: 0.34008)
PegasusX	The paper proposes two methods to analyze the behavior of neurons in neural nets. The main idea is to study the neuron activation patterns of classification models and explore if the performance can be explained through neurons' activation behavior. The authors propose two approaches: one that models neurons' Activation behavior as a graph and examines whether the neurons form meaningful communities, and the other that examines the predictability of neurons' behavior using entropy. The paper is interesting, easy to read, and well-structured. However, the paper is below the acceptance threshold. The reviewers raised several concerns, including the lack of a thorough literature survey, the lack and limited experimental validation, and concerns about the scalability of the proposed methods. The AC agrees with the reviewers and recommends rejection. (ROUGE-1: 0.32743, ROUGE-2: 0.03111, ROUGE-L: 0.29646)
RAMMER	The paper proposes two methods to analyze the behavior of neurons in neural nets. The authors propose two approaches: one that models neurons' activation behavior as a graph and examines whether the neurons form meaningful communities, and the other that examines the predictability of neurons' behavior using entropy. They found that entropy is negatively correlated withtraining accuracy, modularity is correlated with training accuracy, and that the number of well defined neural communities increase with accuracy. The reviewers agree that the paper is interesting, easy to read, and very well-structured. The main technical contribution of this paper comes from explaining the neural classifiers by combining the graph-theoretic and information- theoretic approaches. The reviewers pointed out that the entropy based performance metric is a novel contribution, which sets our work apart from previous researches. However, the scalability and practicality of the work should be examined carefully. (ROUGE-1: 0.35319, ROUGE-2: 0.06410, ROUGE-L: 0.34043)
Example 2, `https://github.com/oaimli/PeerSum/blob/main/examples/e2.pdf`	

Ground truth meta-review	This paper studies the method to achieve the batch size-invariant for policy gradient algorithms (PPO, PPG). The paper achieves this by decoupling the proximal policy from the behavior policy. Empirical results show that the methods are somewhat effective at providing batch size invariance. After reading the authors' feedback, the reviewer discussed the paper and they did not reach a consensus. On the one hand, the rebuttal made some reviewers change their minds who appreciated the explanations provided by the authors and the new Figure that better highlights the batch size invariance property. On the other hand, some reviewers think that there is still significant work to be done to get this paper ready for publication. In particular, it is necessary to improve the theoretical analysis and the evaluation of the empirical results. I encourage the authors to follow the reviewers' suggestions while they will update their paper for a new submission.
PRIMERA	This paper proposes a new method to deal with batch size-invariance for policy optimization algorithms such as PPO and PPG. The key insight is to decouple the proximal policy (used for controlling policy updates) from the behavior policy (used for off-policy corrections). Our experiments help explain why these algorithms work, and additionally show how they can make more efficient use of stale data. The reviewers agree that this paper studies an interesting problem of batch size invariance. The writing is clear, and the method is easy to understand. The experimental results are sufficient to demonstrate the effectiveness of the proposed method. However, the reviewers also agree that the paper could be improved in terms of clarity of presentation and the design of the experiments. In particular, it is important to provide a more rigorous analysis of our ablations, showing that the effects of all of the ablations are statistically significant at the 0.1 (ROUGE-1: 0.48780, ROUGE-2: 0.17439, ROUGE-L: 0.46612)
LED	This paper proposes a method to achieve batch size-invariance for policy optimization. The reviewers agree that this is an interesting problem and the paper is well-written. However, the reviewers also agree that the paper lacks theoretical justification, and the claim of the batch size invariant property is vague and weak. The authors are encouraged to improve the paper and resubmit to a future venue. (ROUGE-1: 0.41284, ROUGE-2: 0.21296, ROUGE-L: 0.40367)
PegasusX	The paper proposes a method to achieve batch size invariance for policy optimization algorithms such as PPO and PPG. The paper achieves this by decoupling the proximal policy from the behavior policy. The experiments demonstrate the effectiveness of the method. The reviewers were unanimous in their opinion that the paper is not ready for publication at ICLR. The main concerns were around the clarity of the presentation and the experimental design. The authors have addressed some of the concerns in their rebuttal, but the paper still needs significant work before it can be accepted. (ROUGE-1: 0.52846, ROUGE-2: 0.23770, ROUGE-L: 0.50407)

RAMMER	The paper proposes a method to deal with batch size-invariance for policy optimization algorithms such as PPO and PPG. The paper achieves this by decoupling the proximal policy from the behavior policy. The authors also provide two new algorithm variants to achieve the batch size invariance property. The reviewers agree that the paper is clearly written, is simple to read, and flows well. The experimental results are sufficient to demonstrate the effectiveness of the proposed method. However, the theoretical analysis on how this method could be batch-invARIant in the surrogate loss or final performance is not sufficient. The reviewers also point out the lack of theoretical justification, especially the use of the EWMA. (ROUGE-1: 0.54682, ROUGE-2: 0.24151, ROUGE-L: 0.50187)
Example 3, `https://github.com/oaimli/PeerSum/blob/main/examples/e3.pdf`	
Ground truth meta-review	This paper reveals that popular data poisoning systems, Fawkes and LowKey, fail to effectively protect user privacy in facial recognition. The methods to defend against poisoning attacks are quite simple—you can either adaptively tune the face recognition models or just wait for more advanced facial recognition systems. Given these "disappointed" findings from the technical solution side, this paper further argues that legislation may be the only viable solution to prevent abuses of facial recognition. Overall, all the reviewers highly appreciate the comprehensive and rigorous evaluations provided in this paper and enjoy reading it. The biggest concern is raised by the Reviewer 6s7m, given this work fails to discuss/compare to previous works on Facial identity anonymizing and the technical contribution is incremental. During the discussion period, all other reviewers reach a consensus that 1) facial identity anonymizing is not relevant; and 2) this work make enough contributions and is worthy to be heard by the general community; the Reviewer 6s7m still hold the opposite opinion, but is okay for accepting this paper anyway. In the final version, the authors should include all the clarification provided in the discussion period.

PRIMERA	This paper studies the effect of data poisoning in face recognition and the relation to the defense techniques. Two solutions of defense are given: oblivious trainer and adaptive trainer. The claim is that, any existing poisoning methods cannot protect the privacy of users in the face images. The paper shows that perturbing facial images does not offer long term security; future systems can still recognize the once perturbed image(s). The paper is well written and easy to follow. The topic discussed in this paper is of significant importance as the rapid development of deep learning techniques nowadays also poses great threats to people's privacy, especially for the face data. Extensive experiments have been conducted to statistically ascertain the authors' claims. However, the technical novelty is very limited. Rather than fancying legislative alternative, a research paper needs to propose technical solution. The original evaluations performed by the Fawkes and LowKey papers were deemed sufficient by reviewers at the time. We believe our paper thus makes an important contribution in demonstrating how such evaluations may have overestimated the protections offered to users. The design of defense strategies that achieve high robustness and accuracy, in Section 3.5. These strategies demonstrate that robustness need not come at a cost in accuracy in all situations, as is often claimed in the adversarial examples literature. Overall, this is a facial privacy analysis with insightful claims, but the presentation and the discussion is very confusing. Thus, we are not able to agree this argument whether is reasonable and solid. The final rating will depend on the authors' feedback. (ROUGE-1: 0.41410, ROUGE-2: 0.07522, ROUGE-L: 0.38326)
LED	This paper studies the effect of data poisoning in face recognition and the relation to the defense techniques. It shows that the methods fail over time and they do not generalize to future attacks. Two solutions of defense are given: oblivious trainer and adaptive trainer. Both methods successfully defeat two poisoning attack baselines, raising awareness on the inefficiency of the poisoning-based identity protection mechanism. The paper is well written and easy to follow. The topic discussed in this paper is of significant importance as the rapid development of deep learning techniques nowadays also poses great threats to people's privacy, especially for the face data. It is in interesting paper and has value for the community to consider longer term security measures for facial recognition privacy concerns. The experimental results are within expectation and little technical insight is gained. (ROUGE-1: 0.40606, ROUGE-2: 0.06707, ROUGE-L: 0.35152)
PegasusX	The paper presents a new defense against data poisoning attacks. The paper is well written and the analysis is thorough. The authors also provide a discussion of the defense and the attack dynamics. The main concern is that the technical novelty is limited. The rebuttal did not convince the reviewers. The AC agrees with the reviewers and recommends acceptance. (ROUGE-1: 0.29719, ROUGE-2: 0.06478, ROUGE-L: 0.28112)

RAMMER	The paper studies the problem of data poisoning in face recognition and the relation to the defense techniques. The paper shows that perturbing facial images does not offer long term security; future systems can still recognize the once perturbed image(s). The paper also shows that an "oblivious" model trainer can simply wait for future developments in computer vision to nullify the protection of pictures collected in the past. The paper is well-written and easy to read. The authors have examined several strategies to enable modern face recognition models to defense attacks from widely used data poisoning methods. However, the technical contribution from this paper is limited as little new insight is provided. The experimental results are within expectation and little technical insight is gained. The main conclusion is that people should not rely on technical solutions to protect users privacy and legislation actions are what is actually needed. (ROUGE-1: 0.42353, ROUGE-2: 0.08284, ROUGE-L: 0.39412)
Example 4, `https://github.com/oaimli/PeerSum/blob/main/examples/e4.pdf`	
Ground truth meta-review	This paper presents a method which selects feasible data augmentations suitable for contrastive time series representation learning. The topic in this paper is timely and interesting. One of 4 reviewers did not complete the review, not responding to a few reminders. So, one emergency reviewer, who is an expert in meta-learning was added. While there is one review that strongly supports this work, two reviews remained unsupportive after the discussion period ended. I appreciate the authors for making efforts in responding to reviewers' comments. However, after the discussion period, most of reviewers had concerns in this work, pointing out that the technical correctness needs further justification and experiments should be improved. While the idea is interesting, the paper is not ready for the publication at the current stage. I encourage to resubmit the paper after addressing these concerns.
PRIMERA	This paper proposes an information-aware approach to representation learning for time series. The formulation focuses on how to obtain effective data augmentations and addresses the underlying problem from information-theoretic viewpoints, leading to the two optimization criteria, namely, high fidelity and high variety. The experimental results on several time series datasets for forecasting and classification show improvements over the methods in comparison. All reviewers agree that this is a very good submission that touches upon a kind of neglected and less fancy area of data augmentation for time-series - authors propose a very neat solution that improves upon the state of the art and demonstrates good performance across datasets and various ablation studies. (ROUGE-1: 0.31621, ROUGE-2: 0.03187, ROUGE-L: 0.26877)
LED	This paper proposes a new approach to data augmentation for time series representation learning. The reviewers agree that the paper is well-written and the proposed approach is novel and interesting. However, there are concerns about the novelty of the approach and the significance of the results. The authors are encouraged to revise the paper based on the reviewers' comments and resubmit to a future venue. (ROUGE-1: 0.41951, ROUGE-2: 0.15764, ROUGE-L: 0.40000)

PegasusX	The paper proposes a new data augmentation approach based on information theory, a meta learning approach and an approach to select optimal data augmentation for contrastive learning. The paper is well written and the proposed approach is interesting. However, the experimental results are not convincing. The authors have provided a detailed response to the reviews, but the paper still needs more work before publication. (ROUGE-1: 0.34483, ROUGE-2: 0.09950, ROUGE-L: 0.32512)
RAMMER	This paper proposes a meta-learning approach for data augmentation for time series. The reviewers found the paper to be well-written and well-motivated, and the approach to be novel and interesting. However, the reviewers also found the experimental results to be insufficiently convincing. In particular, some of the analyses and ablations could be more complete, in order to truly investigate whether the conclusions are correct. (ROUGE-1: 0.33010, ROUGE-2: 0.06863, ROUGE-L: 0.33010)
Example 5, https://github.com/oaimli/PeerSum/blob/main/examples/e5.pdf	
Ground truth meta-review	This works presents a new and interesting notion of margin for deep neural networks (that incorporates representation at all layers). It then develops generalization bounds based on the introduced margin. The reviewers pointed some concerns, including some notation issues, complexity in case of residual networks, removal of exponential dependence on depth, and dependence on a hard to compute quantity - \kapp^{adv}. Some of these concerns were addressed by the authors. At the end, most of the reviewers find the notion of all-layer margin introduced in this paper a very novel and promising idea for characterizing generalization in deep networks. Agreeing with reviewers, I recommend accept. However, I request the authors to accommodate remaining comments /concerns raised by R1 in the final version of your paper. In particular, in your response to R1 you mentioned for one case you saw improvement even with dropout, but that is not mentioned in the revision; Please include related details in the draft.
PRIMERA	The paper presents a bound on the generalization error of a deep network in terms of margin at each layer of the network. The starting premise is that extending the existing margin generalization bounds to deep networks worsen exponentially with the depth. Recent work which removed that exponential dependency is claimed to require a more involved proof and complicated dependence on input. The paper provides a new bound that is simpler and tighter. A second contribution is to extend their bounds to robust classifier. Since their bounds depend on instance-specific margins, the extension to the robust case is straightforward. Finally, they present a new algorithm motivated by their bounds, that maximized margin on all layers. They show that the resultant network has much lower error than standard training. The paper is well-presented and in spite of being theoretical is very nicely developed so that the main contributions come out clearly to non-specialists too. A few minor comments: The inner min in Equation 2.2 seems to be a typo. In Theorem 2.1, there is typo around the definition of ξ. Below thoerem 2. 1, the phrase "depend on the q-th moment" has 'q' undefined. Typo "is has a" in Theorem 3.1 is the the most insightful result, though simple to prove in hindsight. It would greatly help the readers if simple figures are used to explain this insightful result in the final manuscript. (ROUGE-1: 0.36181, ROUGE-2: 0.07576, ROUGE-L: 0.35176)

LED	This paper presents a new way to measure the margin in the context of deep networks that removes the exponential dependency of depth in the corresponding generalization bounds. The reviewers agree that this is a valuable contribution both for theory and practices. (ROUGE-1: 0.32836, ROUGE-2: 0.10050, ROUGE-L: 0.28856)
PegasusX	The paper presents a new generalization bound for deep neural networks and develops a regularizer which optimize quantities related to the bound and improve generalization error on competitive baselines. The paper is well-written and the main contributions come out clearly to non-specialists. However, the paper still requires rewriting to properly emphasize that it does not avoid exponential dependence on depth, but rather avoids an explicit dependence. The authors also need to discuss the computability of kappadv. (ROUGE-1: 0.37975, ROUGE-2: 0.08511, ROUGE-L: 0.32911)
RAMMER	The paper shows that the all-layer margin has a clear and direct relationship with generalization for deep models. The paper also shows that this bound can be adapted to adversarial robustness of the deep model. The paper is well-written and well-motivated. The reviewers found the paper to be a valuable contribution both for theory and practices. (ROUGE-1: 0.29358, ROUGE-2: 0.04630, ROUGE-L: 0.27523)

Table 11: Generated meta-reviews by fully-supervised PRIMERA, LED, PegasusX and RAMMER for random samples, and ROUGE scores measuring the quality of generated meta-reviews in comparison to the ground truth one.

4.2 Hierarchical Sentiment Consolidation

Based on results from Chapter 4.1, we found that it is challenging for our baseline and other benchmark models to recognize and resolve conflicts among source documents when generating the meta-review. It shows that the experimented models still fail to truly understand these cross-document relationships. LLMs such as LLaMA2 (Touvron et al., 2023) and GPT-4 (OpenAI, 2023) have since emerged and they exhibit strong summarisation capability (Zhang et al., 2024b). However, it is unclear if these models could truly consolidate information from multiple source documents to generate meta-reviews. These models may generate summaries with inaccurate overall sentiments because they may not be able to aggregate opinionated information.

To tackle this, we hypothesize that the human meta-reviewers follow a three-layer framework of sentiment consolidation, including the input layer, the consolidation layer, and the generation layer. We ground the scientific meta-reviewing process on a set of review aspects, including *Advancement*, *Soundness*, *Novelty*, *Overall*, *Clarity*, and *Compliance*. The framework first identifies and extracts sentiments from the input documents and then consolidate sentiments for each review aspect in the consolidation layer. The generation layer generates the final meta-review based on the aggregated opinions from the consolidation layer.

To validate our hypothesis, we prompt LLMs with the hypothesised sentiment consolidation logic, decomposing the meta-review generation process into multiple steps. We run our experiments on proprietary and open-source LLMs, and conduct comprehensive comparison across different prompting approaches including prompting with chain-of-thought instructions (Wei et al., 2022) and following a pipeline. We collected human annotations to extract sentiments from source documents, and found high agreement between the annotators. We then attempt to automate the annotation (for extracting sentiments) using GPT-4, and find that GPT-4 has a moderate agreement with human annotators. To evaluate the quality of generated meta-reviews, we conduct reference-based and reference-free automatic evaluation, and reference-free human evaluation. As current MDS evaluation metrics cannot capture the sentiment nuances, we propose two sentiment-aware evaluation metrics: (1) a reference-based metric that measures sentiment similarity between the generated meta-review and the human-written reference based on review aspects; and (2) a reference-free one that measures sentiment fusion for each review aspect based on our sentiment extraction with GPT-4. Specifically The reference-based metric is based on sentiment similarity between the machine-generated summary and the ground truth summary. The sentiment similarity is calculated as the cosine similarity of the two sentiment vectors for the machine-generated summary and the ground truth summary. The sentiment vector is constructed based on the frequency of sentiments in all the review aspects. The reference-free metric is based on the accuracy of sentiment prediction by GPT-4 from the sentiments in the input reviews compared with the sentiments in the machine-generated summary.

A Sentiment Consolidation Framework for Meta-Review Generation

Miao Li[1] **Jey Han Lau**[1] **Eduard Hovy**[1,2]
[1]School of Computing and Information Systems, The University of Melbourne
[2]Language Technologies Institute, Carnegie Mellon University
miao4@student.unimelb.edu.au,
{laujh, eduard.hovy}@unimelb.edu.au

Abstract

Modern natural language generation systems with Large Language Models (LLMs) exhibit the capability to generate a plausible summary of multiple documents; however, it is uncertain if they truly possess the capability of information consolidation to generate summaries, especially on documents with opinionated information. We focus on meta-review generation, a form of sentiment summarisation for the scientific domain. To make scientific sentiment summarization more grounded, we hypothesize that human meta-reviewers follow a three-layer framework of sentiment consolidation to write meta-reviews. Based on the framework, we propose novel prompting methods for LLMs to generate meta-reviews and evaluation metrics to assess the quality of generated meta-reviews. Our framework is validated empirically as we find that prompting LLMs based on the framework — compared with prompting them with simple instructions — generates better meta-reviews.[1]

1 Introduction

Notable strides have been made in abstractive text summarization (El-Kassas et al., 2021) with the advancement of Large Language Models (LLMs) (Zhao et al., 2023) over recent years. With even a simple instruction such as "*tl;dr*" or "*please write a summary*", these models can generate plausible summaries which are found more preferred over those written by humans (Pu et al., 2023). However, it is uncertain if these models truly possess the ability of information consolidation, especially when summarizing documents that are composed of opinionated information. The models may take shortcuts to generate texts instead of correctly understanding and aggregating information from the source documents (Gehrmann et al., 2023) and they may generate abstractive summaries with incorrect overall sentiment.

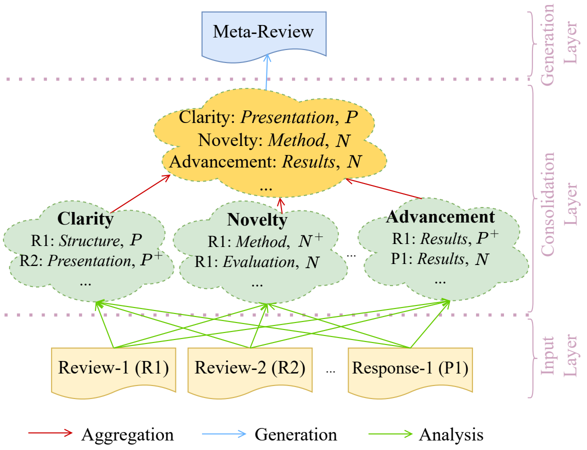


Figure 1: The three-layer framework of the underlying information consolidation logic in meta-reviewing (P: Positive, P^+: Strongly positive, N: Negative, N^+: Strongly negative).

Automated sentiment summarization holds significant importance (Kim et al., 2011) and there have been sentiment summarization datasets; however, most of them are in the product review domain. These datasets are less interesting for investigating information consolidation as (1) the summaries are extractive, composed of a simple combination of extracted snippets (Amplayo et al., 2021), and (2) the summary of product reviews is about extracting the majority sentiment (which is a simple consolidation function). To address this, in this paper, we propose the task of scientific sentiment summarization, taking the meta-reviews in scientific peer review as summaries.[2] The investigation of meta-review generation (Li et al., 2023a) presents an exciting opportunity for exploring the intricate process of multi-document information consolidation that involves complex judgement.

[1] The code and annotated data are accessible at https://github.com/oaimli/MetaReviewingLogic.

[2] The representative peer review platform which is publicly available is www.openreview.com.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 10158–10177
August 11-16, 2024 ©2024 Association for Computational Linguistics

This is because (1) meta-reviewers are supposed to understand not only all the reviews from different reviewers but also the multi-turn discussions between the reviewers and the author and write their comments to support the acceptance decision of the manuscript, (2) the logic of arguments (from reviewers and authors) has to be taken into account to arrive at the final sentiment in the meta-reviews and it is not a matter of majority voting and (3) meta-reviews have to recognize and resolve conflicts and consensus among reviewers.

In this paper, we hypothesize that human meta-reviewers follow a three-layer sentiment consolidation framework as shown in Figure 1 to write meta-reviews based on reviews and multi-turn discussions in the peer review process. Human and automatic annotation is then conducted to extract sentiments and expressions on various review facets (e.g., novelty and soundness) from corresponding source documents (i.e., reviews and discussions) and these judgements play a critical role in generating the meta-reviews. We also propose two evaluation metrics which focus on assessing sentiments in generated meta-reviews, and experiments empirically validate our proposed three-layer framework when they are integrated as prompts for LLMs to generate meta-reviews.

Contributions of our paper:

- We hypothesize that human meta-reviewers follow a three-layer sentiment consolidation framework when writing meta-reviews;
- We collect human annotations on meta-reviews and corresponding source documents based on the consolidation framework;
- We propose two automatic metrics (reference-free and reference-based) to evaluate the sentiment in the generated meta-reviews.
- Experiments validate the empirical effectiveness of the framework when we incorporate it as prompts for LLMs to generate meta-reviews.

2 Related Work

In this section, we discuss large-scale information consolidation in abstractive summarization, and automated sentiment summarization.

2.1 Large-Scale Information Consolidation

Natural language generation systems are expected to not only have high-quality generations but also have the ability to comprehend the input information, especially for conditional text generation such as multi-document summarization which has to integrate and aggregate information from different source documents (Gehrmann et al., 2023). Most work in the text summarization community only attempts to improve the generation quality of text summarization, such as relevance and faithfulness, without considering the intricate generation process (Phang et al., 2022; El-Kassas et al., 2021; Xiao et al., 2022). For example, Li et al. (2023b) use heterogeneous graphs to represent source documents and borrow the idea of graph compression to train the summarization model to get improvement of the generated summaries. However, it is uncertain if these models truly possess the ability to consolidate information from different source documents.

2.2 Automated Sentiment Summarization

Sentiment summarization aims to summarise the overall sentiment given a set of documents (Hossain et al., 2023). However, most datasets for sentiment summarization are in the product review domain (Amplayo et al., 2021), and scientific sentiment summarization is under-explored. Meta-review generation, which is a typical scenario of scientific sentiment summarization, is to automatically generate meta-reviews based on reviews and the multi-turn discussions between reviewers and the author of the corresponding manuscript (Li et al., 2023a). It is mostly modelled as an end-to-end task (Bhatia et al., 2020; Wu et al., 2022; Shen et al., 2022; Chan et al., 2020). Although Li et al. (2023a) considered the conversational structure of reviews and discussions, their models do not explain how human meta-reviewers write the meta-reviews. Different from investigating checklist-guided iterative introspection for meta-review generation with prompting (Zeng et al., 2024), our work is based on a three-layer sentiment consolidation framework and focuses on various review facets, and we explicitly investigate the sentiment fusion process which is arguably an important aspect of meta-review generation.

3 Sentiment Consolidation Over Multiple Opinionated Documents

In the following section, we introduce the task of scientific sentiment summarization and our three-layer sentiment consolidation framework in meta-review generation, conduct sentiment and expres-

Component	Definition
Content Expression	What the sentiment is talking about
Sentiment Expression	The value of the sentiment
Review Facet	The specific review facet that the judgement belongs to
Sentiment Level	The polarity and strength of the sentiment
Convincingness Level	How well the sentiment is justified in the document

Table 1: Definitions of components in a judgement.

	Min	Max	Average
#Documents/Sample	5	30	12.4
#Words/Sample	1,541	11,901	4,260.9
#Words/Source document	10	1,562	360.5
#Words/Meta-review	16	648	150.9

Table 2: Statistics of the human annotated data.

sion extraction, and analyze the fusion process of scientific sentiments.

3.1 Hierarchical Sentiment Consolidation

The task is meta-review generation. We use the PeerSum[3] dataset where the input is reviews and discussions and the target output is the corresponding human-written meta-review. We should clarify that even though the task is to generate meta-reviews, our focus here is to get the overall sentiment in the meta-reviews to be correct. Our method and evaluation reflect this focus.

Reading the reviewer guidelines from popular academic presses such as ACM and IEEE[4], we find they are mostly about *judgements* on the quality and merit of the manuscript. These judgements are generally based on six review facets of criteria: *Novelty*, *Soundness*, *Clarity*, *Advancement*, *Compliance* and *Overall quality*. The meta-reviewers must form their final opinion based on these judgements from the reviewers and authors. Looking at the meta-reviewer guidelines for ICLR[5] and NeurIPS[6], it recommends the meta-reviewer to understand and aggregate information from the whole peer-reviewing process. That is, a human meta-reviewer should first identify judgements from reviews and discussions, and then consolidate these opinions from different review facets to write their meta-review.

To conceptualize this, we propose a three-layer framework, as shown in Figure 1. The three layers include the input layer, the consolidation layer, and the generation layer. The input layer is the input documents of different types: official reviews and multi-turn discussions. The consolidation layer represents how meta-reviewers process the documents: they first identify and extract judgements from different documents, reorganize the judgements based on review facets, and then consolidate the opinions to form the final opinions of each review facet. In the generation layer, the meta-reviewer writes the meta-review to express the final opinions that they have developed from the previous layer.

3.2 Judgement Identification and Extraction

Judgements lay the foundation of our proposed framework and the whole peer review process. A judgement here expresses sentiment on a review facet and it contains several components: Content Expression, Sentiment Expression, Review Facet, Sentiment Level, and Convincingness Level (definitions are shown in Table 1, and an example is given in Appendix Figure 5). To automate judgement identification and extraction, we first conduct human annotation, and then leverage in-context learning of LLMs to perform more (automatic) annotation.

In human annotation, there are three types of documents including meta-reviews, official reviews, and discussions (the same definition used in Li et al. (2023a)) to be annotated. We recruit two annotators[7] to do this annotation (annotation instructions and design are detailed in Appendix B). 30 samples (i.e., one sample = one meta-review and its corresponding reviews and discussions) are annotated[8],

[3] https://github.com/oaimli/PeerSum

[4] The complete table of official guidelines that we consider is in Appendix A.

[5] https://iclr.cc/Conferences/2024/SACguide

[6] https://nips.cc/Conferences/2020/PaperInformation/AC-SACGuidelines

[7] The two annotators are senior PhD students who are familiar with the peer-review process.

[8] Annotating one sample takes about one hour on average and it costs about 60 hours and 2,100 US dollars in total.

and in total, we have 1,812 and 1,744 judgements from the two annotators. The statistics of these 30 samples are presented in Table 2. We present the agreement of the two annotators in Figure 2.[9] Generally, we see a moderate to high agreement, suggesting that the annotation task is robust and reproducible.

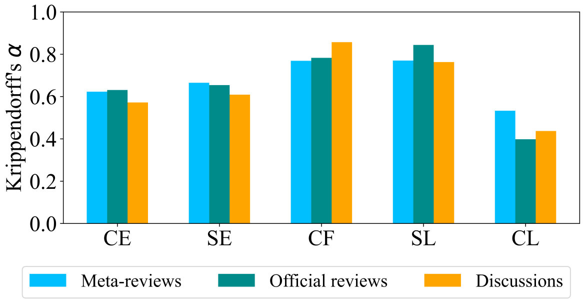


Figure 2: Inter-annotator agreement on meta-reviews, official reviews and discussions in terms of Krippendorff's α for different judgement components including Content Expression (CE), Sentiment Expression (SE), Review Facet (RF), Sentiment Level (SL), and Convincingness Level (CL).

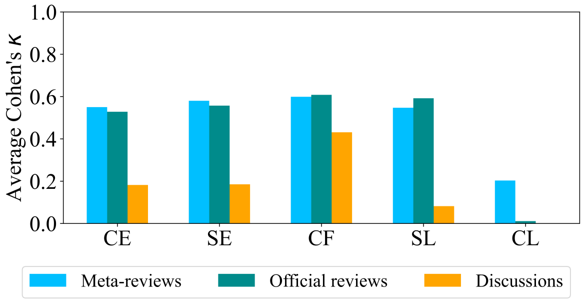


Figure 3: The averaged GPT-4's agreement with two human annotators on meta-reviews, official reviews and discussions in terms of Krippendorff's α for different judgement components including Content Expression (CE), Sentiment Expression (SE), Review Facet (RF), Sentiment Level (SL), and Convincingness Level (CL).

To get more annotated judgements for further experiments and analysis and investigate whether LLMs can be prompted to identify and extract judgements, we split the annotation task into two sub-tasks, extracting content and sentiment expressions and predicting other components of judgements, and use GPT-4 (OpenAI, 2023) with in-context learning (see full prompts in Appendix D and E respectively for the two sub-tasks).[10] We present the average agreement of GPT-4 with the two human annotators in Figure 3.[11] We can see GPT-4 has a moderate agreement with human annotators for meta-reviews and official reviews, but a low agreement for discussions. We suspect this may be because the discussions often contain rebuttals which have a different language to reviews or meta-reviews and extracting judgements from them may be more difficult. Interestingly, we also see that GPT-4 has a poor agreement in terms of convincingness (Figure 3), although the human inter-annotator agreement isn't strong in the first place (Figure 2). These observations suggest convincingness is perhaps a subjective assessment.

Facets	**%Judgements**	**%Documents**
Advancement	0.2545	0.8000
Soundness	0.2786	0.7833
Novelty	0.1817	0.6833
Overall	0.1414	0.5833
Clarity	0.1264	0.4500
Compliance	0.0174	0.0667

Table 3: Frequency of different review facets in meta-review judgements and meta-review documents.

3.3 Sentiment Fusion for Consolidation

With all the annotated judgements extracted by humans and GPT-4, we next dive more into the process of sentiment aggregation. Among all the review facets, we find that *Soundness* and *Advancement* are the two most important review facets when the meta-reviewers write their meta-reviews, while *Compliance* is rarely an issue in meta-reviews (shown in Table 3). This is consistent with our understanding of the peer-reviewing process.

More importantly, we find that human meta-reviewers do not always follow the majority review sentiment. We find that in PeerSum there are 23.7% samples where the meta-reviewer's acceptance decision is not consistent with the prediction based on majority voting by review ratings (a sample is defined as consistent when the number of reviews whose rating ≥ 5 is larger than the number of reviews whose rating < 5 and the final decision is *Accept*). We present an example in Table 4 where the meta-review does not follow the majority view on *Novelty* from the reviews.

[9] Calculation details and more results in terms of both Cohen's κ and Krippendorff's α are in Table 11, Table 12 and Table 13 in Appendix C.

[10] The version of GPT-4 we use is gpt-4-0613.

[11] More agreement results are in Table 14, Table 15 and Table 16 in Appendix C.

Human-written meta-review sentiment sentence
"Although each module in the proposed approach is **not novel**, it seems that the way they are used to address the specific problem of explainability and especially in text games is **novel** and sound."
All corresponding sentiment texts on Novelty in source reviews and discussions
"The generation of temporally extended explanations consists of a cascade of different components, **either straightfoward statistics or prior work**." "The novelty is **a bit low**." "overall novelty is **limitted**" "We contend that all steps are **individually novel as well as their combination**." "we are **the first** to use knowledge graph attention-based attribution to explain actions in such grounded environments"

Table 4: The example of a meta-review sentiment on *Novelty* which is not following majority voting of sentiments in source documents. The **green** and **red** texts indicate positive and negative sentiments, respectively.

Review Facets	Judgements	Full Texts
Advancement	0.677	**0.697**
Soundness	**0.684**	0.667
Novelty	**0.700**	0.650
Overall	**0.643**	0.631
Clarity	**0.712**	0.645
Compliance	0.555	**0.593**

Table 5: Accuracy of GPT-4 in predicting the sentiment levels in meta-reviews for each facet, using either only the annotated judgements ("Judgements") or the full text ("Full Texts") from the source documents.

To understand how well the judgements from the source documents (i.e., reviews and discussions) predict the overall sentiments in the meta-reviews for each review facet, we next formulate a text classification task where the output is the sentiment level of a content expression for a review facet in the meta-review, and the input is either: (1) the annotated judgements for the facet from the source documents; or (2) the full text of the source documents. We (zero-shot) prompt GPT-4 (full prompt detailed in Appendix F) with either input to predict 100 randomly sampled human-annotated instances and present the results in Table 5. Using judgements only as input, we see that it works better in 4 out of 6 facets — this preliminary result suggests our framework of extracting these judgements as an intermediate step may help generate better meta-reviews.

4 Sentiment-Aware Evaluation on Information Consolidation

In this section, we focus more on how to evaluate the sentiments of the generated summaries or meta-reviews in meta-review generation based on our proposed framework. We propose FacetEval and FusionEval which are reference-based and reference-free metrics, respectively.

4.1 Measuring Sentiment Similarity to Human-Written Meta-Review

To assess the quality of generated meta-reviews, we propose a reference-based evaluation metric, FacetEval, measuring the sentiment consistency c between the generated meta-review and the corresponding human-written meta-review in all review facets. Different from the generic evaluation metrics for abstractive summarization or text generation which mostly adopt surface-form matching, we focus more on review facets and their corresponding sentiment levels.

Specifically, we use the distribution of sentiments in all review facets to represent the meta-review and use the cosine similarity of the two vectors as the final score s.

$$s = \cos\left(\boldsymbol{m}_h, \boldsymbol{m}_g\right) \tag{1}$$

$$\boldsymbol{m} = \|_f [P_f^+, P_f, N_f^+, N_f, O_f] \tag{2}$$

where $\|$ denotes concatenation of representations for different facets, $\boldsymbol{m}_h$ and $\boldsymbol{m}_g$ are representations of the human-written and model-generated meta-reviews respectively. The representation $\boldsymbol{m}$ of the meta-review is the concatenation of vector representations of all review facets. Each facet of the document is represented by the frequency of different sentiment levels on the facet. The facet f is represented by a five-dimension vector $[P_f^+, P_f, N_f^+, N_f, O_f]$ where P_f^+ denotes the frequency of *Strongly positive* for the facet f, P_f the frequency of *Positive*, N_f^+ the frequency of *Strongly negative*, N_f the frequency of *Negative*, and O_f whether this facet is involved in the document. All the sentiments are obtained with GPT-4 following in-context learning in Section 3.2.

Following the similarity of meta-reviews, we could also calculate sentiment consistency among official reviews. Specifically, for every two official reviews i and j, the consistency in the facet f is the cosine similarity between two vector represen-

Review Facet	w/ Conflicts	w/o Conflicts
Advancement	0.463 (0.135)	0.551 (0.137)
Soundness	0.526 (0.158)	0.501 (0.110)
Novelty	0.300 (0.159)	0.357 (0.168)
Overall	0.433 (0.147)	0.597 (0.172)
Clarity	0.317 (0.133)	0.337 (0.145)
Compliance	0.827 (0.071)	0.771 (0.118)

Table 6: Sentiment consistency among different official reviews. (Variances are in the brackets.)

tations of documents.

$$c_{ij}^{f} = \cos\left(\boldsymbol{d}_i, \boldsymbol{d}_j\right) \quad (3)$$

where $\boldsymbol{d}^f = [P_f^+, P_f, N_f^+, N_f, O_f]$. Results shown in Table 6 suggest that different reviews are consistent in the sentiment to *Compliance* while there is much lower consistency in *Clarity* and *Novelty*. Moreover, we find that conflict reviews[12] would prefer showing conflicts in *Advancement*, *Novelty*, *Clarity* and *Overall*. This is also consistent with our typical understanding of peer reviews and occasional conflicts among them.

4.2 Measuring Sentiment Fusion for Individual Facets

Sentiments in the generated meta-reviews should be in line with the aggregate sentiment from the individual source documents including reviews and discussions. Seeing GPT-4 can predict the overall sentiment using judgements from source documents (Section 3.3), we introduce a reference-free evaluation metric, FusionEval, which assesses the consistency between the sentiments in the generated meta-review and that predicted by GPT-4 (with zero-shot prompting) from the source documents. Higher consistency implies the overall sentiment in generated meta-reviews are representative of the sentiments in the reviews and discussions (source documents).

Specifically, we first extract judgements from the generated meta-review following Section 3.2, and these judgements consist of *Content Expressions*, E and *Sentiment Levels*, L, and the corresponding *Review Facets*, F. Next, for each expression, $e \in E$, we predict the *Sentiment Level*, l', using GPT-4 (zero-shot) based on all judgements for the same *Review Facet* in the source documents following Section 3.3, and we get predicted *Sentiment Levels* for all judgements, L'. Lastly, FusionEval computes an accuracy score by evaluating L' against L. FusionEval only considers the precision instead of the recall for meta-review sentiments as it is reference-free and we have no information about the count of judgements that should be synthesized.

5 Enhancing LLMs with Explicit Information Consolidation

In this section, we propose two prompting methods to integrate the sentiment consolidation framework to generate meta-reviews. We compare the two methods with other prompting strategies including naive prompting and prompting with LLM-generated logic. We also run experiments on open-source models besides OpenAI ones to investigate the influence of different prompting methods on different models. The experiments are based on automatic and human annotation on 500 samples from PeerSum.[13]

5.1 Prompting LLMs with Sentiment Consolidation Logic

Following the process in Figure 1 we propose decomposing the meta-review generation process in the following steps: (1) Extracting content and sentiment expressions of judgements from source documents; (2) Predicting *Review Facets*, *Sentiment Levels*, and *Convincingness Levels*; (3) Clustering extracted judgements for different review facets; (4) generate a "mini summary" for judgements on the same review facet; and (5) Generating the final meta-review based on the mini summaries for all review facets.

We explore two methods to integrate this process for prompting an LLM. (1) Prompt-Ours: we describe the five steps in a single prompt and ask GPT-4 to generate the final meta-reviews (full prompt in Appendix G.1); (2) Pipeline-Ours: we create one prompt for each of the five steps, where the input for the intermediate step is the output from the previous step (full prompts in Appendix G.2).

We experiment with four open-source and close-source LLMs: GPT-4, GPT-3.5, LLaMA2-70B and LLaMA2-7B.[14]

[12] The same as in PeerSum (Li et al., 2023a), if any two reviews have ratings where the gap is larger than 4 they are conflict reviews.

[13] To avoid data contamination, we only use samples which were produced in and after 2022.

[14] Precise model names for them are: gpt-4-0613, gpt-3.5-turbo-1106, LLaMA2-70B-Chat, LLaMA2-7B-Chat. Note

LLM	Evaluation Metric	Prompt-Naive	Prompt-LLM	Prompt-Ours	Pipeline-Ours
GPT-4	FusionEval	50.14	48.90	53.62	**57.43**
	FacetEval	35.42	40.54	41.98	**42.36**
	ROUGE-1	27.16	27.49	**28.02**	24.91
	ROUGE-2	**6.63**	6.03	6.57	4.57
	ROUGE-L	24.78	24.75	**25.51**	22.70
GPT-3.5	FusionEval	48.35	49.66	51.40	**55.96**
	FacetEval	38.44	36.83	**39.88**	39.50
	ROUGE-1	28.22	25.04	**29.56**	28.92
	ROUGE-2	06.63	05.79	**6.95**	5.52
	ROUGE-L	25.36	22.77	**26.69**	16.13
LLaMA2-7B	FusionEval	46.85	46.83	50.18	**52.68**
	FacetEval	35.89	32.49	38.07	**38.35**
	ROUGE-1	25.94	23.88	**27.00**	19.39
	ROUGE-2	6.04	4.50	**6.86**	4.12
	ROUGE-L	23.57	21.59	**24.59**	17.37
LLaMA2-70B	FusionEval	47.35	48.53	50.24	**52.80**
	FacetEval	35.90	36.40	36.64	**36.82**
	ROUGE-1	26.61	16.60	**26.98**	26.41
	ROUGE-2	**6.56**	3.13	5.58	4.48
	ROUGE-L	**24.62**	14.63	24.20	23.71

Table 7: Performances of different LLMs with different prompting methods. For all metrics, a larger value denotes better performance. The bold and underlined values are the best and second in each row, respectively ($\times 0.01$)

Competition Groups	Preferred	IAA
Prompt-Naive LLaMA2-70B	46.67%	0.64
Prompt-Ours LLaMA2-70B	53.33%	
Prompt-Ours GPT-4	73.33%	0.74
Human-Written	26.67%	

Table 8: Two groups of human evaluation results based on human preferences: (1) comparing generated meta-reviews by Prompt-Naive and Prompt-Ours, and (2) comparing human-written meta-reviews and those generated by Prompt-Ours. IAA denotes inter-annotator agreement calculated with nominal Krippendorff's α.

5.2 Baselines

As baselines, we include two more methods: (1) Prompt-Naive: which prompts an LLM with a simple instruction to generate the meta-review (full prompt in Appendix G.3); and (2) Prompt-LLM: where we ask an LLM to self-generate the detailed steps for meta-review generation and we include these steps in the final prompt for meta-review generation (full prompt in Appendix G.4).

5.3 Reference-Based and Reference-Free Automatic Evaluation

For automatic evaluation, we adopt ROUGE (Lin and Hovy, 2003)[15], FacetEval (Section 4.1) and FusionEval (Section 4.2).[16]

We present the results in Table 7. Most LLMs perform better with either of our prompting methods (Prompt-Ours and Pipeline-Ours) than the baselines (Prompt-Naive and Prompt-LLM). Comparing between Prompt-Ours and Pipeline-Ours, the former seems to do best for ROGUE while the latter for FusionEval/FacetEval. As FusionEval/-FacetEval assesses the accuracy of the overall sentiment specifically, Pipeline-Ours is marginally better when it comes to getting the overall sentiment correct. Comparing different LLMs, GPT-4 seems to work best, although that is mostly true for FusionEval/FacetEval.

that for Pipeline-Ours, we always use GPT-4 for the first two steps, as we find that the other LLMs perform poorly for these tasks.

[15] We use the implementation of the algorithm in https://pypi.org/project/rouge-score/

[16] We do not consider other metrics such as BERTScore (Zhang et al., 2020), UniEval (Zhong et al., 2022), or G-Eval (Liu et al., 2023) as they have only been validated for summarization in news or a more general domain.

Human-Written Meta-Review
This paper suggests a method for defending against adversarial examples and out-of-distribution samples via projection onto the data manifold. The paper suggests a new method for detecting when hidden layers are off of the manifold, and uses auto encoders to map them back onto the manifold. The paper is well-written and the method is novel and interesting. However, most of the reviewers agree that the original robustness evaluations were not sufficient due to restricting the evaluation to using FGSM baseline and comparison with thermometer encoding (which both are known to not be fully effective baselines). After rebuttal, Reviewer 4 points out that the method offers very little robustness over adversarial training alone, even though it is combined with adversarial training, which suggests that the method itself provides very little robustness.
Meta-Review Generated by GPT-4 with Prompt-Naive
The reviewers have raised several concerns about the empirical evaluation and effectiveness of the proposed fortified network model. They note that the benefits of the fortified layers seem to be marginal and the experimental evaluation doesn't sufficiently demonstrate that the increased robustness is due to the fortified layers rather than the adversarial training. Furthermore, most of the reviewers would like to see benchmarks against state-of-the-art defenses such as PGD. Among other issues, reviewers have questioned the use of an autoencoder and the decision to use a denoising module in the hidden layers. Simply put, the reviewers would like a stronger reasoning for the methodological choices. There are also requests for clarification on the denoising terms and a comparison with a simple regularization loss. Moreover, a reviewer also raised a question on the effectiveness of the defense method against other types of attacks that might potentially compromise the proposed network. Most reviewers felt that the paper does not provide enough evidence that defends the proposed fortified network against adversarial attacks. Finally, concerns about the presentation of results and tables, and grammar errors have also been pointed out. In conclusion, while the reviewers acknowledge the relevance and importance of the research problem, the manuscript requires significant amendments to convince the experts about the effectiveness of the proposed method in tackling the problem. The authors must address these concerns and improve the manuscript to make a compelling case for their proposed approach.
Meta-Review Generated by GPT-4 with Prompt-Ours
The paper at hand proposes a fortified network model as a defense against adversarial attacks, based on the concept of denoising autoencoders. A key novelty in the presented approach is the placement of the denoising feature within the hidden layers, rather than just the input layer.The clarity and structure of the document have been appreciated by a number of reviewers, particularly the experimental data presentation and robust bibliography. There are, however, a few areas noted for improvement - predominantly around grammatical mistakes and confusing sentence structure. A significant source of debate among reviewers centers on the empirically evaluated efficacy of the fortified model. The authors propose that their model improves robustness by up to 5% over the baseline, but several reviewers argue that this improvement is marginal and does not sufficiently justify the implementation of an additional objective. This argument is most prominent when looking at the study's use of non-standard models and a PGD baseline that performs below the state-of-the-art, leading to questions about the accuracy of the baseline evaluation. Further, accuracy results against an Linf adversary of eps>0.1 on CIFAR-10 raises questions around the experiment's setup given known limitations relating to image perturbations. Suggestions were made to explore a nearest neighbor attack from the test set, multiple gradient queries per PGD step, and report standard non-robust accuracies to ensure robustness was not an artefact of model expressivity.Lastly, some apprehension was expressed about the motivation behind and efficacy of using a denoising autoencoder, with calls for comparison to a simpler baseline to justify the approach. In conclusion, while the fortified network model presents an intriguing approach to defending against adversarial attacks, the reviewers do not overwhelmingly agree that the empirical results sufficiently demonstrate advancement over existing methods. The majority find the defensive gain too marginal given the additional complexity and question the setup of the empirical evaluation. Further clarity in method and expanded empirical evidence would facilitating a stronger case for the proposed model.

Table 9: Human-written meta-review and the corresponding meta-reviews generated by GPT-4 with the naive prompt and the prompt based on the sentiment consolidation. (The green spans are positive sentiment values, red spans are negative sentiment values, while blue spans are the content expressions.)

5.4 Reference-Free Human Evaluation

To further validate the effectiveness of our prompting methods, we conduct human evaluations to assess the quality of meta-reviews generated by different prompting methods or written by human meta-reviewers. We recruited three volunteer annotators who are senior PhD students familiar with artificial intelligence research and the peer review process. They are asked to select their preferred meta-reviews based on their own understanding of high-quality meta-reviews without knowing the source.[17]

Prompt-Naive vs Prompt-Ours We randomly select 30 samples and the annotators are asked to compare the generated meta-reviews by Prompt-Naive and Prompt-Ours (using LLaMA2-70B) and select which one is better. Table 8 shows that the meta-reviews generated by Prompt-Ours are selected more by the annotators.

Prompt-Ours vs Human-Written We repeat the same experiments, but this time comparing meta-reviews generated by Prompt-Ours (GPT-4) vs. written by humans. Looking at Table 8, interestingly Prompt-Ours are much more preferred by the annotators. We suspect this may be because the generated meta-reviews tend to be more consistent in terms of the amount of detail it writes for each review facet, where else there is more variance for the human-written meta-reviews.

[17] We use majority voting to get the final human preference.

5.5 Case Study on Generated Meta-Reviews

To dive deeper into what difference the integration of sentiment consolidation framework makes, we also conduct a case study on generated meta-reviews with different prompting methods. We find that generated meta-reviews all seem plausible and machine-generated meta-reviews are much longer than human-written ones. In machine-generated meta-reviews, there are more details which are sometimes unnecessary or redundant. As shown in the example in Table 9, details such as "PGD" or "CIFA-10" are not essential to form the meta-review.

Our proposed Prompt-Ours tend to have a more balanced judgements. For example, in Table 9, Prompt-Naive does not talk about the positive aspects for *Clarity* and only highlights some issues, but Prompt-Ours comments on both the strengths and weaknesses for *Clarity*. This is consistent with the finding in Table 7 that Prompt-Naive gets worse sentiments than Prompt-Ours.

6 Conclusions and Future Work

In this paper, we explore sentiment-focused multi-document information consolidation within the task of scientific sentiment summarization. We introduce a three-layer framework of sentiment consolidation to focus on generating meta-reviews and it considers the sentiments for each review facet in the reviews and discussions. We also propose automatic evaluation metrics that assess the overall sentiments in the generated meta-reviews. Experiments on meta-review generation show that prompting LLMs by following the processes in the three-layer framework results in better meta-reviews, providing an empirical validation of our framework for describing the meta-review writing process. As the sentiment consolidation also exist in other domains where human reviews or comments exist such as politics and advertisement, we will explore adapting our proposed sentiment consolidation framework into other domains in the future.

Limitations

Although integration of the sentiment consolidation framework could improve the generation results, there are still some limitations of this work.

- As in other areas peer review data is not publicly available, we use the data only from some artificial intelligence conferences, and this may make the models biased. We hope that more data from diverse areas could be included.
- Experiments are only in English texts rather than other languages.
- We only inject the information consolidation logic into prompting based models instead of fine-tuning based models. We will investigate leveraging the information consolidation framework to improve fine-tuned models in the future.
- Although GPT-4 can predict meta-review sentiments based on source judgements to some extent, we have to understand more about how these models achieve this and what makes them fail in error cases.
- Meta-review generation is not only about sentiment prediction, future work has to consider more information such as argumentation in source reviews and justification in meta-reviews.

Ethics Statement

While our experiments demonstrate that the models exhibit potential in generating satisfactory meta-reviews to a certain degree, we strongly advise against solely relying on the generated results without manual verification and review, as instances of hallucinations exist in the generations. It is important to emphasize that we do not advocate for replacing human meta-reviewers with LLMs. However, it is noteworthy that these models have the capacity to enhance the meta-reviewing process, rendering it more efficient and effective.

Acknowledgements

We would thank all the anonymous reviewers for their great comments to improve the work. It is impossible to have this work without the volunteer contributions of the annotators on human evaluations.

References

Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. 2021. Unsupervised opinion summarization with content planning. In *AAAI*, pages 12489–12497.

Chaitanya Bhatia, Tribikram Pradhan, and Sukomal Pal. 2020. Metagen: An academic meta-review generation system. In *SIGIR*, pages 1653–1656.

Hou Pong Chan, Wang Chen, and Irwin King. 2020. A unified dual-view model for review summarization and sentiment classification with inconsistency loss. In *SIGIR 2020*, pages 1191–1200.

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

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2023. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *JAIR*, 77:103–166.

Md. Murad Hossain, Luca Anselma, and Alessandro Mazzei. 2023. Exploring sentiments in summarization: Sentitextrank, an emotional variant of textrank. In *Proceedings of the 9th Italian Conference on Computational Linguistics*, volume 3596.

Hyun Duk Kim, Kavita Ganesan, Parikshit Sondhi, and ChengXiang Zhai. 2011. Comprehensive review of opinion summarization.

Miao Li, Eduard Hovy, and Jey Han Lau. 2023a. Summarizing multiple documents with conversational structure for meta-review generation. In *Findings of EMNLP*.

Miao Li, Jianzhong Qi, and Jey Han Lau. 2023b. Compressed heterogeneous graph for abstractive multi-document summarization. In *AAAI*.

Chin-Yew Lin and Eduard H. Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *HLT-NAACL*, pages 71–78.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using GPT-4 with better human alignment. *CoRR*, abs/2303.16634.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Jason Phang, Yao Zhao, and Peter J. Liu. 2022. Investigating efficiently extending transformers for long input summarization. *CoRR*, abs/2208.04347.

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead. *CoRR*, abs/2309.09558.

Chenhui Shen, Liying Cheng, Ran Zhou, Lidong Bing, Yang You, and Luo Si. 2022. Mred: A meta-review dataset for structure-controllable text generation. In *Findings of ACL*, pages 2521–2535.

Po-Cheng Wu, An-Zi Yen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2022. Incorporating peer reviews and rebuttal counter-arguments for meta-review generation. In *CIKM*, pages 2189–2198.

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2022. PRIMERA: pyramid-based masked sentence pre-training for multi-document summarization. In *ACL*, pages 5245–5263.

Qi Zeng, Mankeerat Sidhu, Hou Pong Chan, Lu Wang, and Heng Ji. 2024. Scientific opinion summarization: Meta-review generation with checklist-guided iterative introspection. *CoRR*, abs/2305.14647.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *ICLR*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *CoRR*, abs/2303.18223.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *EMNLP*, pages 2023–2038.

A Review Criteria in Different Reviewer Guidelines

Academic Press	Review guidelines
ACM	https://dl.acm.org/journal/dgov/reviewer-guidelines
ACL Rolling Review	https://aclrollingreview.org/reviewertutorial
IEEE	https://conferences.ieeeauthorcenter.ieee.org/understand-peer-review/
Springer	https://www.springer.com/gp/authors-editors/authorandreviewertutorials/howtopeerreview/evaluating-manuscripts/10286398
NeurIPS	https://neurips.cc/Conferences/2021/Reviewer-Guidelines
ICLR	https://iclr.cc/Conferences/2023/ReviewerGuide#Reviewinginstructions
ACL	https://2023.aclweb.org/blog/review-acl23/
Cambridge University Press	https://www.cambridge.org/core/services/aop-file-manager/file/5a1eb62e67f405260662a0df/Refreshed-Guide-Peer-Review-Journal.pdf

Table 10: Review guidelines from different academic presses.

B Annotation Instructions for Human Annotation

The screenshots of the two-page annotation instruction for human annotation are shown in Figure 4 and Figure 5 in the last two pages of the Appendix.

C Inter-Annotator Agreement Among Human Annotators and GPT-4

We describe how we calculate inter-annotator agreement among human annotators and GPT-4 here. For Content Expression and Sentiment Expression, as they are highlighted text spans we calculate the character-level agreement with Krippendorf's α and Cohen's κ. Specifically, for each document, two annotators may highlight different text spans for Content Expression and Sentiment Expression. We construct two vectors of the same length as the characters to represent the highlighting behaviours of any two annotators. This agreement shows whether annotators identify sentiments from similar text spans.

For *Review Facet*, *Sentiment Level*, and *Convincingness Level*, we calculate Krippendorf's α and Cohen's κ in a common way. We first identify whether two annotators recognize sentiment from the same text span with a ROUGE threshold (the summation of ROUGE-1, ROUGE-2 and ROUGE-L between highlighted text spans for sentiment is larger than 2.0), and calculate agreement on the predicted values.

Inter-annotator agreement between two human annotators for human annotation in Section 3.2 are present in Table 11, Table 12, and Table 13. Averaged agreement of GPT-4 with the two human annotators are present in Table 14, Table 15, and Table 16.

D Prompt to Get Content and Sentiment Expressions with GPT-4

Annotation	Cohen's κ	Krippendorf's α
Content Expression	0.623	0.623
Sentiment Expression	0.666	0.665
Review Facet	0.769	0.769
Sentiment Level	0.770	0.770
Convincingness Level	0.534	0.533

Table 11: Human annotator agreement on annotating meta-reviews.

Annotation	**Cohen's κ**	**Krippendorff's α**
Content Expression	0.631	0.631
Sentiment Expression	0.654	0.654
Review Facet	0.783	0.783
Sentiment Level	0.844	0.844
Convincingness Level	0.405	0.398

Table 12: Human annotator agreement on annotating official reviews.

Annotation	**Cohen's κ**	**Krippendorff's α**
Content Expression	0.572	0.572
Sentiment Expression	0.609	0.609
Review Facets	0.857	0.857
Sentiment Levels	0.764	0.763
Convincingness Levels	0.455	0.437

Table 13: Human annotator agreement on annotating discussions.

Annotation	*A*	*B*	**Avg**
Content Expression	0.558	0.542	0.550
Sentiment Expression	0.565	0.594	0.580
Review Facets	0.588	0.610	0.599
Sentiment Levels	0.552	0.541	0.547
Convincingness Levels	0.213	0.192	0.203

Table 14: GPT-4 agreement in terms of Cohen's κ with human annotators *A* and *B* on annotating meta-reviews.

Annotation	*A*	*B*	**Avg**
Content Expression	0.522	0.534	0.528
Sentiment Expression	0.544	0.569	0.557
Review Facets	0.579	0.637	0.608
Sentiment Levels	0.594	0.589	0.592
Convincingness Levels	0.008	0.013	0.011

Table 15: GPT-4 agreement in terms of Cohen's κ with human annotators *A* and *B* on annotating official reviews.

Annotation	*A*	*B*	**Avg**
Content Expression	0.176	0.187	0.182
Sentiment Expression	0.182	0.188	0.185
Review Facets	0.480	0.381	0.431
Sentiment Levels	0.123	0.046	0.082
Convincingness Levels	0.0	0.0	0.0

Table 16: GPT-4 agreement in terms of Cohen's κ with human annotators *A* and *B* on annotating discussions.

```
Please read the document:

{{source_document}}

This task requires you to analyze the above document which is used to express opinions on the quality of a scientific manuscript. You are good at understanding the sentiment information with judgements in the document.
Please first identify the sentence with judgements only on the quality of scientific manuscripts based on the review facets for scientific peer-review: novelty, soundness, clarity, advancement, compliance and overall quality within the given document.
Once you have found a sentence that provides judgement in one or more of these areas, you then need to extract the specific expression of sentiment and the content it refers to.

The process can be broken into two steps:
1) Identify a judgement sentence that focuses on the quality of the manuscript based on the given criteria.

2) From the identified judgement sentence, extract two pieces of information: the sentiment expression and the content expression. The sentiment expression is the specific term or phrase that conveys the sentiment or opinion. The content expression pertains to the content that this sentiment is referring to.

Please provide the data in the following format:
{"judgement_sentence": "sentence", "content_expression": "content", "sentiment_expression": "sentiment"}

Here are a few examples for your reference:
{"judgement_sentence": "The writing of the paper is not well-written.", "content_expression": "The writing of the paper", "sentiment_expression": "not well-written"}
{"judgement_sentence": "Experimental results are not sufficiently substantiated.", "content_expression": "Experimental results", "sentiment_expression": "not sufficiently substantiated"}
{"judgement_sentence": "This paper presents two novel approaches to provide explanations for the similarity between two samples based on 1) the importance measure of individual features and 2) some of the other pairs of examples used as analogies.", "content_expression": "approaches", "sentiment_expression": "novel"}

The predicted judgments (following the same jsonline format of the above example):
```

E Prompt to Get Judgement Component Predictions with GPT-4

```
Please first read the document below:
```

```

{{source_document}}


Please predict the facet that the given judgements are talking about. You can refer to the context in the above source document.

Possible facets:

Novelty: How original the idea (e.g., tasks, datasets, or methods) is, and how clear where the problems and methods sit with respect to existing literature (i.e., meaningful comparison).

Soundness: (1) Empirical: how well experiments are designed and executed to support the claims, whether methods used are appropriate, and how correctly the data and results are reported, analysed, and interpreted. (2) Theoretical: whether arguments or claims in the manuscript are well supported by theoretical analysis, i.e., completeness and the methodology (e.g., mathematical approach) and the analysis is correct.

Clarity: The readability of the writing (e.g., structure and language), reproducibility of details, and how accurately what the research question is, what was done and what was the conclusion are presented.

Advancement: Importance of the manuscript to discipline, significance of the contributions of the manuscript, and its potential impact to the field.

Compliance: Whether the manuscript fits the venue, and all ethical and publication requirements are met.

Overall: Overall quality of the manuscript, not for specific facets.


You are also good at understanding sentiment information in the judgements.

Please predict the original expresser of the sentiment in the judgement sentence. You can refer to the context in the source document.

Possible sentiment expressers:

- Self: the sentiment is from the speaker
- Others: the sentiment is quoted from others


Please predict how well the sentiment in the judgement sentence is justified in the document in your understanding. You can refer to
```

```
the context in the source document.

Possible sentiment convincingness:

- Not applicable: the sentiment is explicitly excerpted from others.
- Not at all: not convincing at all or when there is no justification. How well the sentiment is justified in the document in your understanding
- Slightly Convincing: there is some evidence or logical reasoning, but it might not be comprehensive.
- Highly Convincing: leaving little room for doubt.


Please predict the polarity and strength of the sentiment in the judgement sentence. You can refer to the context in the source document.

Possible sentiments polarities:

- Strong negative: very negative
- Negative: minor negative
- Positive: minor positive
- Strong positive: very positive


Judgements:
{{judgement_expressions}}

Your predictions for the above judgements (following the same jsonlines format, return the same number of lines, and keep the same content and sentiment expressions):
```

F Prompts to Predict Meta-Review Sentiment Levels

F.1 Prediction with Judgements of Source documents

The judgements are extracted from source documents, and they are in the same review facet to the target meta-review judgement.

```
You will be given source judgements from reviewers for a scientific manuscript. Your task is to implicitly write a meta-review for these judgements and predict the sentiment level based on these judgements.

Source Judgements:

{{source_judgements}}

Candidate Sentiment Levels:

- Strong negative
- Negative
- Positive
- Strong positive

Content Expression:

{{content_expression}}

Predict the sentiment level of the given content expression based on the above judgements. You must follow the following format.
{"Content Expression": the above content expression, "Sentiment Level": your predicted sentiment level}
```

F.2 Prediction with Full Texts of Source documents

The source texts are the concatenation of the source documents.

```
You will be given multiple review documents for a scientific manuscript. Your task is to implicitly write a meta-review and predict the sentiment level based on these documents.

Source Documents:

{{source_texts}}

Candidate Sentiment Levels:

- Strong negative
- Negative
- Positive
- Strong positive

Content Expression:

{{content_expression}}

Predict the sentiment level of the given content expression based on related information in the above documents. You must follow the following format.
{"Content Expression": the above content expression, "Sentiment Level": your predicted sentiment level}
```

G Prompts for Meta-Review Generation with Integration of Information Consolidation Logic

G.1 Prompt with Descriptive Consolidation Logic

```
    Your task is to write a meta-review based on the following reviews and discussions for a scientific manuscript.

{{input_documents}}

Following the underlying steps below will get you better generated meta-reviews.
```

```

1. Extracting content and sentiment expressions of judgements in all above review and discussion documents;

2. Predicting Review Facets, Sentiment Levels, and Convincingness Levels;
Candidate review facets: Novelty, Soundness, Clarity, Advancement, Compliance, and Overall quality
Candidate sentiment levels: Strong negative, Negative, Positive and Strong positive
Candidate convincingness levels: Not at all, Slightly Convincing, Highly Convincing

3. Reorganize extracted judgements in different clusters for different review facets;

4. Generate a small summary for judgements on the same review facet with comparison and aggregation;

5. Aggregate judgements in different review facets and write a meta-review based on the aggregation.


You may follow these steps implicitly and only need to output the final meta-review. The final meta-review:
```

G.2 Prompts Used in the Pipeline Generation

Prompts for the first two steps, getting content and sentiment expressions and predicting other judgement components, are the same as prompts in Appendix D and Appendix E, respectively.

For the step of generating sub-summaries for individual facets, the prompt is as follows.

```
{{input_judgements}}

Write a summary of the above judgements on {{criteria_facet}} of a manuscript.
```

For the step of generating final meta-reviews based on sub-summaries of individual facets, the prompt is as follows.

```
{{input_sub_summaries}}

Write a meta-review to summarize the above sub-summaries of reviews and discussions in different review facets for a manuscript.
```

G.3 Prompts from Prompt-Naive

For Prompt-Naive in our experiments, the prompt we use is as follows.

```
{{input_documents}}

Write a meta-review based on the above reviews and discussions for a manuscript.
```

G.4 Prompts from Prompt-LLM

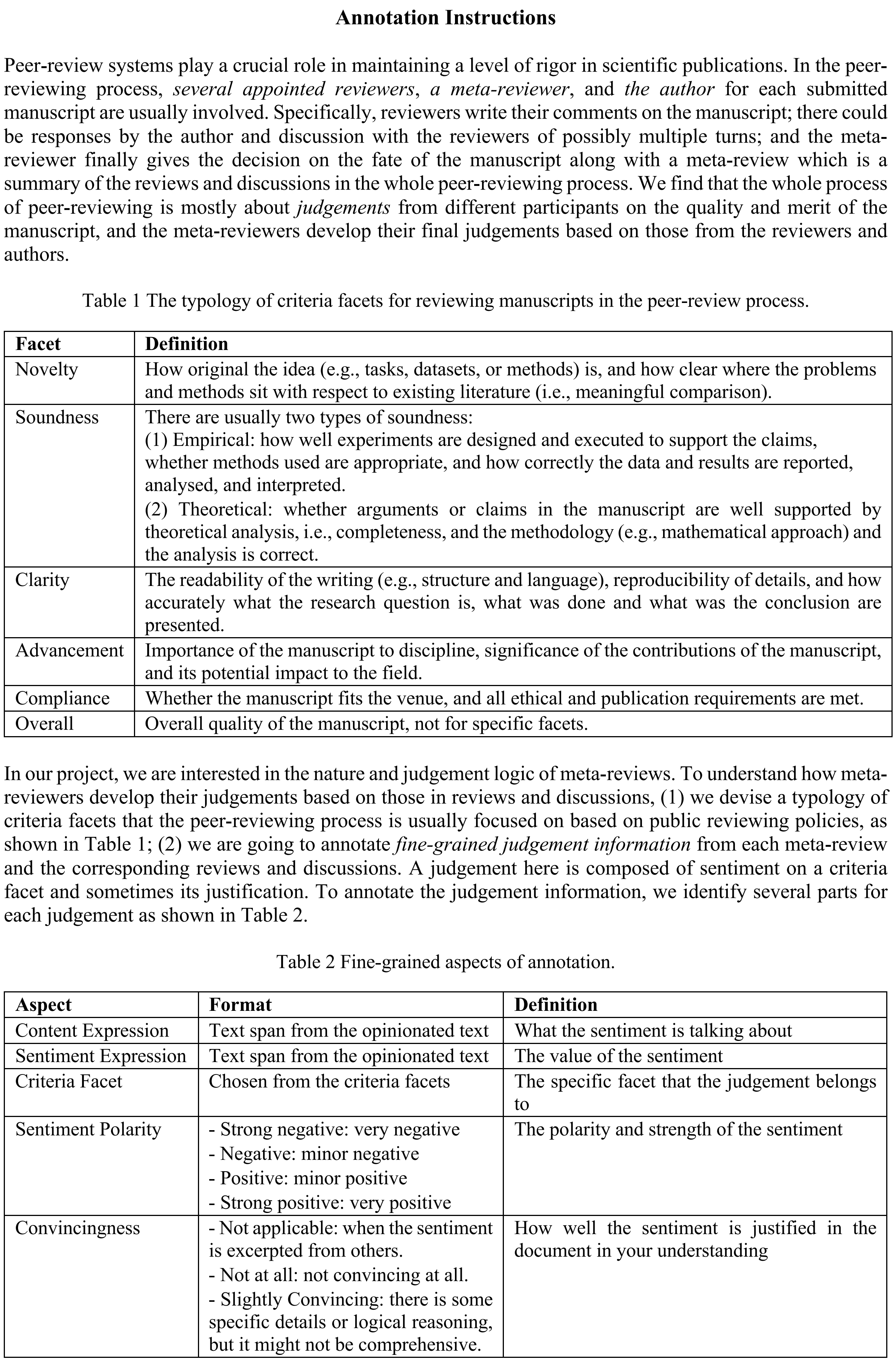
For Prompt-LLM, we have to generate first the steps with GPT-4 and then the meta-review based on the generated steps.

The prompt to generate the steps:

```
{{input_documents}}

What are the steps to write a meta-review specifically for the above reviews and discussions of a manuscript.
```

The prompt to generate the meta-review:

```
{{input_documents}}

Follow the following steps and write a meta-review based on the above reviews and discussions for a manuscript.

{{generated_steps}}
```



Annotation Instructions

Peer-review systems play a crucial role in maintaining a level of rigor in scientific publications. In the peer-reviewing process, *several appointed reviewers*, *a meta-reviewer*, and *the author* for each submitted manuscript are usually involved. Specifically, reviewers write their comments on the manuscript; there could be responses by the author and discussion with the reviewers of possibly multiple turns; and the meta-reviewer finally gives the decision on the fate of the manuscript along with a meta-review which is a summary of the reviews and discussions in the whole peer-reviewing process. We find that the whole process of peer-reviewing is mostly about *judgements* from different participants on the quality and merit of the manuscript, and the meta-reviewers develop their final judgements based on those from the reviewers and authors.

Table 1 The typology of criteria facets for reviewing manuscripts in the peer-review process.

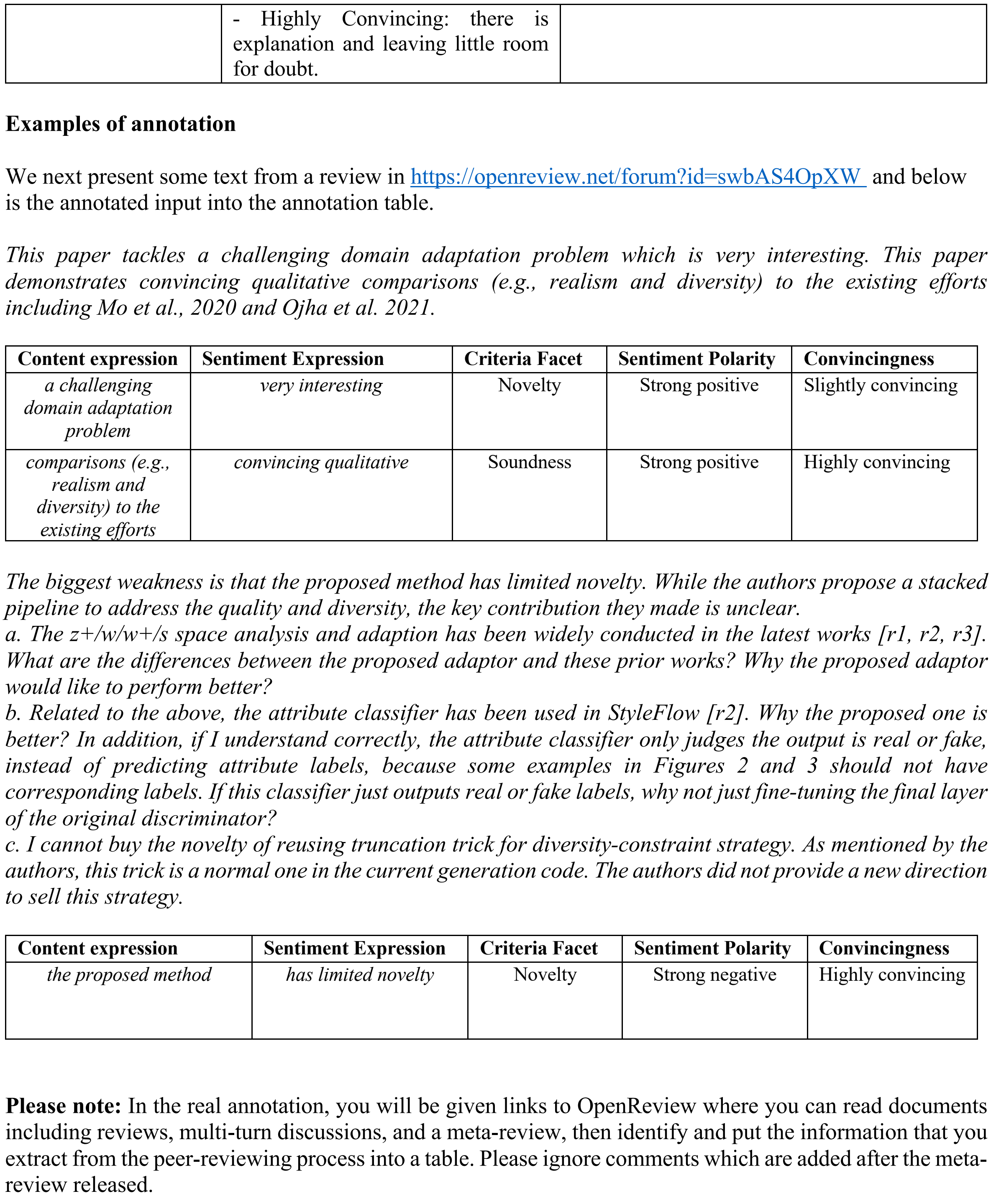
Facet	**Definition**
Novelty	How original the idea (e.g., tasks, datasets, or methods) is, and how clear where the problems and methods sit with respect to existing literature (i.e., meaningful comparison).
Soundness	There are usually two types of soundness: (1) Empirical: how well experiments are designed and executed to support the claims, whether methods used are appropriate, and how correctly the data and results are reported, analysed, and interpreted. (2) Theoretical: whether arguments or claims in the manuscript are well supported by theoretical analysis, i.e., completeness, and the methodology (e.g., mathematical approach) and the analysis is correct.
Clarity	The readability of the writing (e.g., structure and language), reproducibility of details, and how accurately what the research question is, what was done and what was the conclusion are presented.
Advancement	Importance of the manuscript to discipline, significance of the contributions of the manuscript, and its potential impact to the field.
Compliance	Whether the manuscript fits the venue, and all ethical and publication requirements are met.
Overall	Overall quality of the manuscript, not for specific facets.

In our project, we are interested in the nature and judgement logic of meta-reviews. To understand how meta-reviewers develop their judgements based on those in reviews and discussions, (1) we devise a typology of criteria facets that the peer-reviewing process is usually focused on based on public reviewing policies, as shown in Table 1; (2) we are going to annotate *fine-grained judgement information* from each meta-review and the corresponding reviews and discussions. A judgement here is composed of sentiment on a criteria facet and sometimes its justification. To annotate the judgement information, we identify several parts for each judgement as shown in Table 2.

Table 2 Fine-grained aspects of annotation.

Aspect	**Format**	**Definition**
Content Expression	Text span from the opinionated text	What the sentiment is talking about
Sentiment Expression	Text span from the opinionated text	The value of the sentiment
Criteria Facet	Chosen from the criteria facets	The specific facet that the judgement belongs to
Sentiment Polarity	- Strong negative: very negative - Negative: minor negative - Positive: minor positive - Strong positive: very positive	The polarity and strength of the sentiment
Convincingness	- Not applicable: when the sentiment is excerpted from others. - Not at all: not convincing at all. - Slightly Convincing: there is some specific details or logical reasoning, but it might not be comprehensive.	How well the sentiment is justified in the document in your understanding

Figure 4: The first page of the annotation instruction for human judgement annotation.



	- Highly Convincing: there is explanation and leaving little room for doubt.	

Examples of annotation

We next present some text from a review in https://openreview.net/forum?id=swbAS4OpXW and below is the annotated input into the annotation table.

This paper tackles a challenging domain adaptation problem which is very interesting. This paper demonstrates convincing qualitative comparisons (e.g., realism and diversity) to the existing efforts including Mo et al., 2020 and Ojha et al. 2021.

Content expression	Sentiment Expression	Criteria Facet	Sentiment Polarity	Convincingness
a challenging domain adaptation problem	*very interesting*	Novelty	Strong positive	Slightly convincing
comparisons (e.g., realism and diversity) to the existing efforts	*convincing qualitative*	Soundness	Strong positive	Highly convincing

The biggest weakness is that the proposed method has limited novelty. While the authors propose a stacked pipeline to address the quality and diversity, the key contribution they made is unclear.
a. The z+/w/w+/s space analysis and adaption has been widely conducted in the latest works [r1, r2, r3]. What are the differences between the proposed adaptor and these prior works? Why the proposed adaptor would like to perform better?
b. Related to the above, the attribute classifier has been used in StyleFlow [r2]. Why the proposed one is better? In addition, if I understand correctly, the attribute classifier only judges the output is real or fake, instead of predicting attribute labels, because some examples in Figures 2 and 3 should not have corresponding labels. If this classifier just outputs real or fake labels, why not just fine-tuning the final layer of the original discriminator?
c. I cannot buy the novelty of reusing truncation trick for diversity-constraint strategy. As mentioned by the authors, this trick is a normal one in the current generation code. The authors did not provide a new direction to sell this strategy.

Content expression	Sentiment Expression	Criteria Facet	Sentiment Polarity	Convincingness
the proposed method	*has limited novelty*	Novelty	Strong negative	Highly convincing

Please note: In the real annotation, you will be given links to OpenReview where you can read documents including reviews, multi-turn discussions, and a meta-review, then identify and put the information that you extract from the peer-reviewing process into a table. Please ignore comments which are added after the meta-review released.

Figure 5: The first page of the annotation instruction for human judgement annotation.

4.3 Reflections

In the chapter, we present two studies on scientific opinion summarization. For the first study, we developed a new meta-review generation dataset which has rich inter-document relationships with explicit hierarchical conversational structures, cross-references and conflicting opinions. This data facilitates the research of MDS over scientific reviews. We also present a baseline that implements the relationship-aware sparse attention and multi-task training to incorporate the conversational structure into the meta-review generation process. Our experiments results show that the simple baseline performs significantly better than other benchmark models. Manual analysis on the generated meta-reviews reveals that most models cannot recognize and resolve conflicts in the reviews, suggesting a promising avenue for future research. In the second study, we present a three-layer summarisation framework that better captures the sentiments in the reviews, our experimental results show that prompting based on our framework generate better meta-reviews than other strong prompting baselines.

In these two studies, our proposed approaches use different backbone foundation models. The first study explores encoder-decoder pre-trained language models, while the second study uses decoder-only LLMs. The reason for these differences is because when we started the first study LLMs has only emerged and encoder-decoder pre-trained models are the state-of-the-art for conditional text generation tasks such as text summarization (Beltagy et al., 2020; Lewis et al., 2020; Zhang et al., 2020a). The relationship-aware sparse attention is applied to these encoder-decoder models. Our second study, on the other hand, uses LLMs because at that time LLMs is the new state-of-the-art for generation and has strong instruction-following capability. We can therefore incorporate the three-layer framework into the generation process through prompting. This prompting approach has its own advantages, e.g., the meta-review generation process gets more transparent and controllable and we no longer require large-scale training data.

There are still some limitations in our studies. We frame scientific meta-review generation as a MDS problem, but the real-world meta-reviews may not be based on the source documents but by drawing from the meta-reviewer's 'wisdom', such as their own experience or opinions. This makes evaluation of the quality of generated meta-reviews even harder as there is no ground truth for the information which comes from meta-reviewer's 'wisdom'.

In modelling of source documents, we incorporate the explicit conversational structure by modifying the model's attention mechanism. However, this would not be possible for closed-source LLMs. It would be interesting to incorporate the conversational structure through prompting. Moreover, beyond explicit cross-document relationships we should also consider modelling implicit relationships, such as argument and discourse links. This may in turn help to generate higher-quality meta-reviews.

Although prompting based on our three-layer framework can generate better meta-reviews than other simple prompting approaches. We still do not understand how it achieves this and what makes them fail in error cases. To solve this, we have to have rigorous experiments to evaluate them with customized experimentation. We could get LLMs to synthesize test cases for specific evaluation in a large scale, such as edge cases which require the model to consider justifications to balance opinions from different source documents. Lastly, our proposed method specifically targets the scientific domain in this chapter, it would be interesting to see whether such framework can be transferred to other domains, such as product and business reviews. This is what we are going to study in the next chapter.

Chapter 5

Domain-General Opinion Summarization

Ideally, opinion summarization systems should be grounded and transparent and should work across different domains. This means that the generated meta-reviews should accurately reflect the opinions in the source documents and the models should offer some evidence to justify their outputs. In the previous chapter, our work focuses only on scientific reviews. Although our three-layer meta-reviewing framework gets better meta-reviews and transparency, it does not work in other domains. Most existing work in opinion summarization is based on end-to-end modelling. They lack transparency because of the black-box nature of the models. These studies also tend to exclusively focus on product reviews or business reviews. Although some of them have transparent pipelines (Hosking et al., 2024), their approach cannot be adapted to summarizing scientific reviews. This is because sentiment consolidation in the product or business domains is based on majority voting (i.e., the overall opinion corresponds to the majority view) (Angelidis et al., 2021; Hosking et al., 2024) but in Chapter 4.2 we found that the majority assumption fails in almost a quarter of the instances when summarizing scientific reviews.

Therefore, this chapter addresses the last research question of the thesis: *how to build grounded and transparent opinion summarisation systems that work across different domains*. The content of this chapter is based on the following publication.

- **Miao Li**, Jey Han Lau, Eduard Hovy, and Mirella Lapata. 2025. Decomposed Opinion Summarization with Verified Aspect-Aware Modules. In *Findings of the Association*

for Computational Linguistics: ACL 2025, pages 24805–24841, Vienna. Association for Computational Linguistics.

5.1 Aspect-Aware Decomposition Across Domains

To make the process of opinion summarization more transparent and grounded, we decompose the generation process by review aspects with a modular approach. In Chapter 4.2, we propose a three-layer meta-reviewing framework based on review aspects for the scientific domain, and it works well in generating the meta-reviews with better predicted sentiments. However, it is designed only for the scientific domain. Moreover, the framework for scientific sentiment consolidation focuses only on sentiments instead of considering the justifications in the meta-review generation. To make opinion summarization more transparent, existing studies first extract information clusters in the format of sentence fragments (Bhaskar et al., 2023b; Hosking et al., 2023, 2024) or sentiment spans (Li et al., 2024b), and then generate summaries (e.g., using a language model) based on the clusters. However, they are limited in that they are based on assumptions that are only valid for specific domains (e.g., based on the popularity of opinions) and are not entirely transparent (e.g., the clusters or aggregation step cannot be easily verified).

Therefore, in this chapter we propose to decompose opinion summarization into simpler domain-agnostic sub-tasks: *Aspect Identification*, *Opinion Consolidation*, and *Meta-review Synthesis*. Intuitively, *Aspect Identification* first identifies text fragments in the input documents that are about a particular aspect (e.g., novelty); next *Opinion Consolidation* creates meta-reviews for each aspect, and finally *Meta-review Synthesis* generates a global meta-review for all aspects. The three modules are implemented by zero-shot prompting of LLMs. We take this approach because large-scale human-written meta-reviews and intermediate outputs for modules are usually not available, and they are difficult to annotate. Our implementation is also inspired by recent applications of chain-of-thought prompting (Wei et al., 2022) and its variants (Khot et al., 2023; Zhou et al., 2023) in Chapter 2.2, which address reasoning problems by decomposing complex tasks into a sequence of simpler sub-problems.

This decomposition is different from our previous work although both are guided by review aspects. Our previous work extracts sentiments based on the pre-defined format for scientific opinions and predicts the overall sentiments based on them. In contrast, in this work, we extract text fragments and generate aspect-focused meta-reviews based on extracted texts. The length of the text fragments is dynamic, and that makes the model work for other domains and consider more opinion information such as justification when generating the final meta-review.

We conduct evaluation experiments on not only generated meta-reviews but also intermediate outputs. To showcase the versatility of our approach for different domains, we conduct experiments on opinion summarization of shoe reviews, hotel reviews, and scientific article reviews. We implement our approach with different LLM backbones including open- and closed-source models and compare them with strong prompting and fine-tuning baselines. The prompting baselines include automatic decomposition which prompt LLMs with model-generated summarization steps, chunk-wise decomposition which recursively summarizes the reviews chunk-by-chunk with prompting, and naive aspect-aware prompting which does not perform task decomposition but is aspect aware (i.e., injecting aspect descriptions into the prompt). Our automatic evaluation also assesses aspect coverage and faithfulness of the generated meta-reviews.

Decomposed Opinion Summarization with Verified Aspect-Aware Modules

Miao Li[1,3] **Jey Han Lau**[1] **Eduard Hovy**[1,2] **Mirella Lapata**[3]
[1]School of Computing and Information Systems, The University of Melbourne
[2]Language Technologies Institute, Carnegie Mellon University
[3]School of Informatics, The University of Edinburgh
miao4@student.unimelb.edu.au,
{laujh, eduard.hovy}@unimelb.edu.au,
mlap@inf.ed.ac.uk

Abstract

Opinion summarization plays a key role in deriving meaningful insights from large-scale online reviews. To make the process more explainable and grounded, we propose a domain-agnostic modular approach guided by review aspects (e.g., cleanliness for hotel reviews) which separates the tasks of aspect identification, opinion consolidation, and meta-review synthesis to enable greater transparency and ease of inspection. We conduct extensive experiments across datasets representing scientific research, business, and product domains. Results show that our approach generates more grounded summaries compared to strong baseline models, as verified through automated and human evaluations. Additionally, our modular approach, which incorporates reasoning based on review aspects, produces more informative intermediate outputs than other knowledge-agnostic decomposition approaches. Lastly, we provide empirical results to show that these intermediate outputs can support humans in summarizing opinions from large volumes of reviews.[1]

1 Introduction

Reviews are omnipresent in the digital world, providing invaluable insights into products (Bražinskas et al., 2021), businesses (Angelidis et al., 2021), even scientific articles (Li et al., 2023b). Automatic opinion summarization aims to *aggregate* a large and diverse set of reviews about a particular *entity* (e.g., hotel) into a single easy-to-read *meta-review* (or summary). A good meta-review should accurately reflect the balance of opinions in the source reviews and speak to the entity's specific *aspects* (e.g., *Cleanliness*, *Service*, *Location*). A useful meta-review should also present some *evidence* justifying its content.

Opinion summarization has distinct characteristics that set it apart from other summarization tasks. Firstly, it cannot rely on reference summaries for training, as human-written meta-reviews are not generally available (e.g., across entities and domains) and can be difficult to crowdsource (e.g., for entities represented by thousands of reviews). Secondly, a meta-review needs to cover the most important aspects related to the entity of interest. And finally, given the subjective nature of the summarization task, systems should offer some evidence to justify their output.

Prior approaches to generating meta-reviews broadly fall into three categories. *Extractive* methods create summaries by selecting a few representative sentences from source reviews (Angelidis et al., 2021; Basu Roy Chowdhury et al., 2022; Li et al., 2023a). While these approaches are scalable and inherently attributable, the summaries tend to be overly detailed and lack coherence. *Abstractive* methods rely on neural language models to generate fluent and coherent meta-reviews with novel language (Fremann and Klementiev, 2019; Chu and Liu, 2019; Coavoux et al., 2019; Bražinskas et al., 2020; Amplayo et al., 2021a,b; Iso et al., 2021; Bražinskas et al., 2021; Cattan et al., 2023). In the era of large language models (LLMs), long-context language models could be directly used on opinion summarization with prompting (Touvron et al., 2023; OpenAI, 2023). However, these abstractive approaches are neither transparent nor controllable due to the black-box nature of end-to-end modeling.

Hybrid methods (Hosking et al., 2023; Bhaskar et al., 2023; Hosking et al., 2024; Li et al., 2024), the third category of summarisation approaches, could generate fluent and explainable summaries. They first extract information clusters in the format of sentence fragments (Hosking et al., 2023; Bhaskar et al., 2023; Hosking et al., 2024) or sentiment spans (Li et al., 2024), and then generate

[1] Code and data are released at: https://github.com/oaimli/ModularMetaReview

Findings of the Association for Computational Linguistics: ACL 2025, pages 24805–24841
July 27 - August 1, 2025

summaries (e.g., using a language model) based on the clusters. However, they are limited in that they are based on assumptions valid for specific domains (e.g., based on the popularity of opinions) and are not entirely transparent (e.g., either the clusters or aggregation step cannot be easily verified).

In this paper, we propose to decompose opinion summarization into simpler sub-tasks that are domain-agnostic and can be executed by prompting-based LLMs dedicated to these sub-tasks. Our approach is also inspired by recent applications of chain-of-thought prompting (Wei et al., 2022) and its variants (Khot et al., 2023; Zhou et al., 2023), which address reasoning problems by decomposing complex tasks into a sequence of simpler sub-problems, which are solved sequentially. Our decomposition consists of three high-level modules, namely *Aspect Identification*, *Opinion Consolidation*, and *Meta-Review Synthesis*. Intuitively, we first identify text fragments in the input reviews discussing aspects about the entity and domain in question; next we create meta-reviews for *each* aspect, and finally we generate a global meta-review for *all* aspects (see Figure 1). Our approach eschews problems relating to the scale of the input to some extent, since reviews can be processed in parallel to identify their aspects. It also avoids problems with clusters being diffuse or irrelevant since we leverage domain specific aspect definitions (as part of the prompt) to obtain interpretable clusters. Finally, our decomposition is controllable, and evidence-based, as the output of each module can be traced back to its input. Our contributions can be summarized as follows:

- We propose a domain-agnostic decomposition of opinion summarization into three verifiable modules that are instantiated with LLMs using zero-shot prompting.
- Extensive experiments on three datasets from different domains demonstrate that our aspect-informed approach produces more grounded meta-reviews than strong baselines in terms of automatic and human evaluation.
- Aspect-aware decomposition yields more useful reasoning chains compared to end-to-end prompting with automatic decomposition (Khot et al., 2023), and our experiments show that our generated intermediate reasoning steps are empirically helpful in assisting humans with summarizing reviews.

2 Related Work

Our work focuses on abstractive opinion summarization that aims to generate fluent and coherent summaries with novel language (Bražinskas et al., 2021; Li et al., 2023b). This task has been explored in different domains, such as summarizing reviews of products, businesses, and scientific articles (Chu and Liu, 2019; Bražinskas et al., 2021; Li et al., 2023b; Hosking et al., 2024). Previous abstractive methods lack transparency in their decision-making process due to their end-to-end nature (Bražinskas et al., 2021; Cattan et al., 2023; Touvron et al., 2023; OpenAI, 2023). Hybrid approaches implement pipelines with transparent intermediate outputs, however, most of them are aspect agnostic, focusing on how to organize or annotate the input for downstream processing. For example, Hosking et al. (2024) propose a method that represents sentences from reviews as paths through a learned discrete hierarchy, and then use LLMs to generate sentences based on frequent paths retrieved from this hierarchy. Their retrieval module relies on majority voting, which is less effective in domains where minority but well-argued opinions are valuable, such as in scientific reviews (Li et al., 2023b).

Inspired by a well-known decomposition of multi-document summarization into three modules, namely content selection, content consolidation (or fusion), and output generation (Barzilay and McKeown, 2005; Radev and McKeown, 1998; Lebanoff et al., 2020; Slobodkin et al., 2024; Krishna et al., 2021; Li et al., 2024), we apply a similar aspect-guided decomposition to the task of opinion summarization. A few approaches take aspects into account. For example, Angelidis et al. (2021) cluster opinions through a discrete latent variable model and extract sentences based on popular aspects or a particular aspect, while Li et al. (2023a) learn aspects by clustering of aspect-related words. Amplayo et al. (2021a) fine-tune pre-trained models on synthetic data enhanced with aspect annotations which can be used to control output summaries at inference time. Different from earlier work (Li et al., 2023a; Bhaskar et al., 2023; Hosking et al., 2023) which identifies aspects based on sentences, our approach identifies opinions in text fragments of variable lengths. We delegate the task of aspect identification to prompt engineering, demonstrating that LLMs can reliably extract aspects in the format of flexible text fragments given an input review and aspect definitions without additional

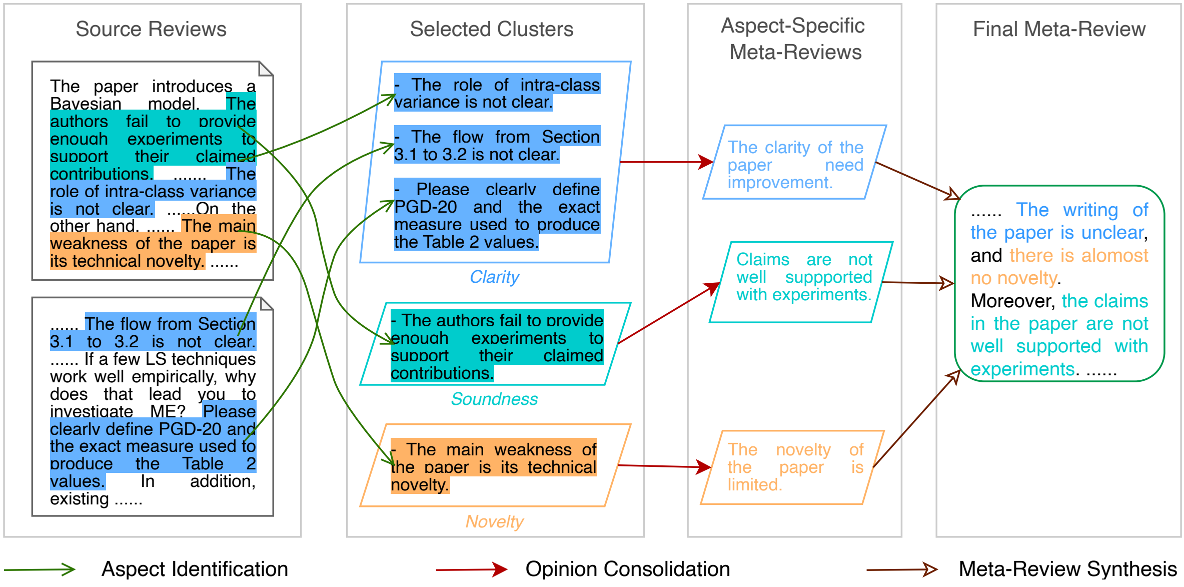

Source Reviews
The paper introduces a Bayesian model. The authors fail to provide enough experiments to support their claimed contributions. .... The role of intra-class variance is not clear. ......On the other hand, ... The main weakness of the paper is its technical novelty. ......
...... The flow from Section 3.1 to 3.2 is not clear. ...... If a few LS techniques work well empirically, why does that lead you to investigate ME? Please clearly define PGD-20 and the exact measure used to produce the Table 2 values. In addition, existing ......
Selected Clusters
- The role of intra-class variance is not clear.
- The flow from Section 3.1 to 3.2 is not clear.
- Please clearly define PGD-20 and the exact measure used to produce the Table 2 values.
Clarity
- The authors fail to provide enough experiments to support their claimed contributions.
Soundness
- The main weakness of the paper is its technical novelty.
Novelty
Aspect-Specific Meta-Reviews
The clarity of the paper need improvement.
Claims are not well suppported with experiments.
The novelty of the paper is limited.
Final Meta-Review
...... The writing of the paper is unclear, and there is alomost no novelty. Moreover, the claims in the paper are not well supported with experiments. ......
Aspect Identification
Opinion Consolidation
Meta-Review Synthesis

Figure 1: High-level overview of our decomposition for opinion summarization using an example from the scientific domain with three aspects (*Clarity*, *Soundness*, and *Novelty*). The modules *Aspect Identification*, *Opinion Consolidation*, and *Meta-Review Synthesis* are instantiated with prompt-based LLMs and operate in sequence. The output of *Aspect Identification* serves as input to *Opinion consolidation* and *Meta-Review synthesis* aggregates opinions found in aspect-specific meta-reviews. All prompts and inputs/outputs are in natural language.

training. Bhaskar et al. (2023) and Li et al. (2024) use similar modules to ours. However, intermediate results of their recursive prompting and aspect identification are not inspected or verified and there is limited transparency, and Li et al. (2024) focus exclusively aspects in scientific reviews (e.g., opinions about *Novelty* or *Soundness*).

Our work also relates to recent efforts aiming to improve the in-context learning performance of LLMs through intermediate reasoning chains (Wei et al., 2022; Yao et al., 2023; Khot et al., 2023). Previous approaches focus primarily on mathematical or symbolic reasoning, while intermediate reasoning for complex writing tasks such as opinion summarization remains under-explored (Li et al., 2024). Decomposed prompting (Khot et al., 2023) uses LLMs to predict both the task decomposition into modules and the modules themselves. However, it is unclear whether it is suited to complex tasks like opinion summarization.

3 Task Decomposition

Let $\mathcal{C}$ denote a corpus of reviews on entities $\{e_1, e_2, \dots\}$ from a domain d, for example, hotels or scientific articles. Reviews may discuss a number of relevant aspects $A_d = \{a_1, a_2, \dots\}$, like *Clarity* or *Soundness*, For each entity e_i, our task is to generate a meta-review $\widehat{y}_i$ by synthesizing opinions from a set of source reviews $R_i = \{r_1, r_2, \dots\}$ covering *all* attested aspects A_d. We decompose the task into three modules, namely *Aspect Identification*, *Opinion Consolidation*, and *Meta-Review Synthesis*. We present the inner workings of each module in Figure 1 with an example from the scientific domain. Due to the limited availability of training data, we implement our modules using an unsupervised approach, leveraging zero-shot prompting of LLMs and their instruction following and generation capabilities.[2]

Aspect Identification As not all content in the source reviews is relevant for generating meta-reviews, opinion summarization models must be able to isolate critical information in the input. The first module, *Aspect Identification*, selects text fragments of variable lengths from source reviews discussing any review aspect. Specifically, for reviewed entity e_i, our module identifies text fragments for aspect a_j from source reviews R_i. The module essentially partitions text fragments into aspect-specific clusters $C_{i,j} = \{f_1, f_2, \dots\}$, where fragments f_m can originate from any source review in R_i. For example, in Figure 1, the module identifies fragments in scientific reviews for the aspects *Clarity*, *Soundness*, and *Novelty*. We implement this module with zero-shot LLM prompting. Our prompt template is shown in Appendix A (Figure 3)

[2] It is worth noting that our prompts could be further improved, however, we leave prompt optimization to future work.

and can be modified for different aspects and domains. Our aspect identification is not based on sentence clustering as well-justified opinions may be composed of multiple sentences, and each text fragment could be identified for multiple aspects.

Opinion Consolidation As shown in Figure 1, the output of the first module consists of clusters of text fragments, each discussing a specific aspect. Depending on the domain, these clusters can have a lot of redundancy, often repeating the same opinion. Our second module, *Opinion Consolidation*, aggregates opinions into aspect-specific meta-reviews. We essentially adopt a divide-and-conquer strategy, since generating meta-reviews from aspect-specific clusters is significantly easier than producing an entire summary from reviews containing mixed aspects. Specifically, taking as input cluster $C_{i,j}$, the module generates meta-review $o_{i,j}$ for aspect a_j. As we do not have training data for these intermediate summaries, we also implement this module with zero-shot prompting.[3] Our template (shown in the Appendix, Figure 4) instructs LLMs to integrate opinions (i.e., text fragments) from a specific cluster. For example, in Figure 1, the three sentences in the *Clarity* cluster are aggregated into "*The clarity of the paper needs improvement*".

Meta-Review Synthesis After obtaining all aspect-specific summaries $O_i = \{o_{i,1}, o_{i,2}, \dots\}$, our last module generates the final meta-review $\widehat{y}_i$ for entity e_i; it combines the opinions mentioned in the individual summaries into a fluent and coherent overall summary. An example is given in Figure 1 where the meta-review focuses on the aspects of *Clarity*, *Soundness*, and *Novelty*. Again, this module leverages the generation capabilities of LLMs, and is instantiated via zero-shot prompting. Our template (given in the Appendix, Figure 5) asks the LLM to write a concise meta-review which summarizes the provided opinions and covers all mentioned aspects.

4 Experimental Setup

We showcase the versatility of our approach on different domains. We first describe the datasets used in our experiments, discuss implementation details and comparison baselines, and explain how we evaluate performance with automatic metrics.

Dataset	#Train/ Dev/Test	#Reviews	SourceL	MetaL	#Aspects
PeerSum	22,420/50/100	14.9	5,146	156.1	5
AmaSum	25,203/50/50	381.8	14,495	94.8	10
SPACE	0/25/25	100	14,439	75.7	6

Table 1: Statistics of our experimental datasets. #Train/Dev/Test refer to the number of training, development, and test instances, respectively; #Reviews is the average number of reviews per entity; SourceL refers to the total length of the source reviews (when concatenated) and MetaL to the average meta-review length; #Aspects is the number of aspects covered in each dataset. For AmaSum, the statistics are for the sports shoes subset.

Datasets We conducted experiments on three domains, product reviews for sports shoes, business reviews for hotels, and scientific reviews for research articles. For business reviews, we use SPACE, an opinion summarization dataset constructed by Angelidis et al. (2021). For product reviews, we use the sports shoes subset from AmaSum (Bražinskas et al., 2021). For scientific reviews, we use PeerSum (Li et al., 2023b) and also the human annotations of review aspects from Li et al. (2024).[4] Statistics for these datasets are shown in Table 1.

SPACE (Angelidis et al., 2021) consists of hotel reviews from TripAdvisor, with 100 reviews per entity, as well as reference meta-reviews of customer experiences created by annotators. The dataset provides six aspects for hotels, which we adopt in our experiments, namely *Building*, *Cleanliness*, *Food*, *Location*, *Rooms*, and *Service*. AmaSum contains meta-reviews for a variety of Amazon products, with reference summaries collated from professional review platforms. To cover more domains in our experiments with limited computing resources, we randomly choose the sports shoes subset curated from the RunRepeat platform[5] which covers the aspects: *Breathability*, *Durability*, *Weight*, *Cushioning*, *Stability*, *Flexibility*, *Traction*, *Size and Fit*, *Comfort*, and *Misc*.

PeerSum (Li et al., 2024) contains reviews for scientific articles and corresponding meta-reviews from OpenReview focusing on the aspects of *Novelty*, *Soundness*, *Clarity*, *Advancement*, and *Compliance*. Detailed definitions for all aspects in the datasets (SPACE, AmaSum, and PeerSum) are given in the Appendix B–D.

[3] Some aspects may not have corresponding text fragments in the source reviews, as they do not always cover every aspect.

[4] To make our experiments cost-effective, we randomly sampled 100 test instances from the PeerSum dataset.

[5] https://runrepeat.com/

Models	Coverage↑	G-Eval↑	AlignScore-R/M↑	Rouge↑
Sentiment CoT-GPT-4o (Li et al., 2024)	0.96	0.75	0.72/0.08	**23.47**
FT-Llama 8B (Touvron et al., 2023)	0.87	0.60	0.33/0.06	20.60
Aspect-aware decomposition-GPT-4o (ours)	0.95	0.76	0.68/0.06	20.78
Automatic decomposition-Llama 8B (Khot et al., 2023)	0.58	0.20	0.36/0.03	11.98
Chunk-wise decomposition-Llama 8B (Khot et al., 2023)	0.79	0.65	0.65/0.03	21.19
Naive aspect-aware prompting-Llama 8B (Radford et al., 2019)	0.72	0.62	0.70/0.06	16.93
Aspect-aware decomposition-Llama 8B (ours)	0.90	0.66	0.71/0.07	21.12
Automatic decomposition-Llama 70B (Khot et al., 2023)	0.59	0.31	0.51/0.03	12.0
Chunk-wise decomposition-Llama 70B (Khot et al., 2023)	0.84	0.72	0.65/0.06	21.80
Naive aspect-aware prompting-Llama 70B (Radford et al., 2019)	0.72	0.62	0.70/0.07	16.82
Aspect-aware decomposition-Llama 70B (ours)	**0.97**	**0.76**	**0.76**/**0.09**	22.58

Table 2: Results on scientific **reviews of research articles**. The first section of the table presents results for GPT-4o and state-of-the-art models. The second section has results for Llama-8B, and the third one for Llama 70B. Underlined scores denote best in section per metric while bold scores denote best overall. AlignScore-R calculates AlignScore against source reviews, while AlignScore-M is computed against reference meta-reviews.

Model Comparisons We implement our modular approach with different backbone LLMs, including closed- and open-source models. Since the modules need to have reasonable language generation and instruction following capabilities, we conduct experiments with gpt-4o-2024-05-13[6] from OpenAI, and Llama-3.1-70B-Instruct[7] and Llama-3.1-8B-Instruct[8] from Meta.[9] The prompts used in our experiments are provided in Appendix B–D.

We compare our approach with representative prompting and fine-tuning baselines (see more details in Appendix E). We implement two strong prompting approaches which do not take aspect information into account: *automatic decomposition* breaks down complex reasoning tasks into simpler ones (Khot et al., 2023) by automatically predicting the decomposition and the modules, while *chunk-wise decomposition* (Khot et al., 2023) recursively summarizes the input reviews chunk-by-chunk with prompting.[10] We also compare against the *naive aspect-aware prompting* which does not perform task decomposition but is aspect-aware (Radford et al., 2019). For fine-tuning, we conduct experiments on decoder-only LLMs. Due to computational limitations, we present fine-tuning results only with Llama-3.1-8B[11] on all three datasets. Moreover, we also include generations from strong baseline approaches on our datasets.

Automatic Evaluation Metrics We evaluate the quality of generated meta-reviews in terms of *aspect coverage* and *faithfulness* (against source reviews). Aspect coverage measures how well the generated meta-review for entity e_i captures the aspects discussed in the source reviews. Specifically, we compute the F_1 between the set of aspects present in the generated meta-review and those in the source reviews. We recognize aspects automatically by running our Aspect Identification module (see Section 3) on the system input and output. Opinion faithfulness measures how well opinions in generated meta-reviews are supported by the source reviews. Specifically, we use G-Eval (Liu et al., 2023), a prompting-based evaluation[12] metric, and AlignScore (Zha et al., 2023)[13], a fine-tuned evaluation metric based on information alignment between two arbitrary text pieces. We use the large version of the pre-trained backbone for AlignScore, and we set *nli_sp* as our evaluation mode. We also report Rouge F1 (Lin and Hovy, 2003), as a measure of overall summary quality.[14] To obtain fair conclusions, we make the models output three generations for each instance and present the average performance in the tables.

5 Results and Analysis

We perform experiments on datasets covering multiple domains, comparing meta-reviews generated by our approach with those from strong baselines and state-of-the-art approaches. We further eval-

[6] https://platform.openai.com/docs/models/gpt-4o

[7] https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct

[8] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[9] All models used in our experiments are instruction-tuned.

[10] The input is chunked based on document boundaries. For PeerSum each review is a chunk, while for AmaSum and SPACE chunks correspond to 20% of the source documents.

[11] https://huggingface.co/meta-llama/Llama-3.1-8B

[12] Our prompts are provided in Appendix F.

[13] https://github.com/yuh-zha/AlignScore/tree/main

[14] We use the average F1 of ROUGE-1, ROUGE-2, and ROUGE-L.

Models	Coverage↑	G-Eval↑	AlignScore-R/M↑	Rouge↑
HIRO-abs (Hosking et al., 2024)	0.54	0.35	0.78/0.13	14.90
FT-Llama 8B (Touvron et al., 2023)	0.45	0.12	0.43/0.16	9.90
Aspect-aware decomposition-GPT-4o (ours)	**0.86**	0.87	**0.79**/**0.17**	16.10
Automatic decomposition-Llama 8B (Khot et al., 2023)	0.39	0.11	0.47/0.13	9.23
Chunk-wise decomposition-Llama 8B (Khot et al., 2023)	0.58	0.80	0.66/0.08	**16.59**
Naive aspect-aware prompting-Llama 8B (Radford et al., 2019)	0.54	0.29	0.50/0.07	8.80
Aspect-aware decomposition-Llama 8B (ours)	0.77	0.78	0.69/0.09	16.44
Automatic decomposition-Llama 70B (Khot et al., 2023)	0.31	0.28	0.68/0.14	7.74
Chunk-wise decomposition-Llama 70B (Khot et al., 2023)	0.57	**0.88**	0.54/0.07	15.28
Naive aspect-aware prompting-Llama 70B (Radford et al., 2019)	0.49	0.48	0.60/0.09	7.35
Aspect-aware decomposition-Llama 70B (ours)	0.83	0.86	0.74/0.16	16.40

Table 3: Results on product **reviews of sports shoes**. The first section of the table presents results for GPT-4o and state-of-the-art models. The second section has results for Llama-8B, and the third one for Llama 70B. Underlined scores denote best in section per metric while bold scores denote best overall. AlignScore-R calculates AlignScore against source reviews, while AlignScore-M is computed against reference meta-reviews.

Models	Coverage↑	G-Eval↑	AlignScore-R/M↑	Rouge↑
HIRO-abs (Hosking et al., 2024)	0.87	0.62	**0.83**/**0.24**	**26.50**
TCG (Bhaskar et al., 2023)	0.98	0.66	0.66/0.11	22.98
Aspect-aware decomposition-GPT-4o (ours)	**1.00**	**0.90**	0.81/0.10	21.38
Automatic decomposition-Llama 8B (Khot et al., 2023)	0.65	0.07	0.55/0.15	13.80
Chunk-wise decomposition-Llama 8B (Khot et al., 2023)	0.94	0.80	0.65/0.14	22.9
Naive aspect-aware prompting-Llama 8B (Radford et al., 2019)	0.55	0.06	0.34/0.18	10.30
Aspect-aware decomposition-Llama 8B (ours)	0.97	0.81	0.70/0.10	22.05
Automatic decomposition-Llama 70B (Khot et al., 2023)	0.63	0.38	0.70/0.22	10.0
Chunk-wise decomposition-Llama 70B (Khot et al., 2023)	0.93	0.84	0.65/0.01	22.02
Naive aspect-aware prompting-Llama 70B (Radford et al., 2019)	0.37	0.34	0.44/0.22	5.00
Aspect-aware decomposition-Llama 70B (ours)	0.99	0.88	0.79/0.11	23.46

Table 4: Results on business **reviews of hotels**. The first section of the table presents results for GPT-4o and state-of-the-art models. The second section has results for Llama-8B, and the third one for Llama 70B. Underlined scores denote best in section per metric while bold scores denote best overall. AlignScore-R calculates AlignScore against source reviews, while AlignScore-M is computed against reference meta-reviews.

uate the intermediate outputs obtained from our modules against human annotations and conduct ablations to examine the extent to which individual modules contribute to the summarization task. Finally, in addition to automatic evaluation we conduct human evaluation based on pair-wise system comparisons and intermediate outputs.

Aspect-aware decomposition leads to better aspect coverage and opinion faithfulness. Our results using automatic evaluation metrics are summarized in Table 2 (scientific articles), Table 3 (shoes), and Table 4 (hotels).[15] Across domains we find that our modular approach with GPT-4o or Llama-3.1-70B delivers the highest coverage of review aspects. Our approach with GPT-4o is also better than comparison systems in terms of opinion faithfulness (see AlignScore). Our aspect-aware decomposition is consistently superior to more naive decompositions and prompting methods in terms of aspect coverage across domains and model backbones. We also observe that using Llama-70B as a backbone gives our approach a boost across metrics which is not surprising as larger models tend to have better generation and instruction-following capabilities. Interestingly, the fine-tuned model (FT-Llama 8B) trails behind our modular system when using a backbone LLM of the same scale (Aspect-aware decomposition-Llama 8B), both in terms of aspect coverage and opinion faithfulness. Overall, our results suggest that prompt decomposition is useful in opinion summarization and intermediate reasoning steps based on task and domain-specific knowledge lead to meta-reviews of higher quality.

Llama-70B performs well at identifying and summarizing aspects. In addition to evaluating the generated meta-reviews, we conduct evaluations on the intermediate outputs of our modules. We only report results on the scientific domain

[15] We run inference three times, with different random seeds and report average performance.

Models	Recall↑	Precision↑	F_1 ↑
GPT-4o	**0.82**	0.27	0.40
Llama-3.1-8B	0.80	0.25	0.38
Llama-3.1-70B	0.74	**0.34**	**0.46**

Table 5: Evaluation of text fragments extracted by *Aspect Identification* against human annotations.

Models	AlignScore-S↑	Rouge↑
GPT-4o	0.86	18.40
Llama-3.1-8B	0.82	**18.24**
Llama-3.1-70B	**0.87**	16.93

Table 6: Evaluation of aspect-specific meta-reviews, i.e., intermediate outputs of *Opinion Consolidation*.

reusing the ground truth annotations[16] provided in Li et al. (2024). For *Aspect Identification*, we calculate word-level Recall, Precision, and F_1 between model-extracted text fragments and human-annotated text fragments following Li et al. (2024). The scores shown in Table 5 denote how accurately our approach extracts opinionated text from source reviews. We find that Llama-3.1-70B is the best model for this module, even better than GPT-4o (in terms of F_1). Moreover, Figure 2 shows that Llama-3.1-70B also performs well on individual review aspects, especially frequent ones including *Novelty*, *Soundness* and *Clarity*. For *Opinion Consolidation*, Table 6 shows that Llama-3.1-70B performs better than other models at generating aspect-specific meta-reviews. Taken together, the evaluations on intermediate outputs explain Llama-3.1-70B's superior performance at the end task.

***Opinion Consolidation* is the most important module.** We further examine the contributions of individual modules to meta-review generation. Specifically, we perform two ablations: (1) remove *Aspect Identification* and directly generate aspect-specific meta-reviews based on original reviews and (2) remove *Opinion Consolidation* and directly generate final meta-reviews based on text fragments from *Aspect Identification*. We use Llama-3.1-70B as our backbone LLM because of its superior performance in previous experiments. As we have ground truth text fragments for scientific reviews (Li et al., 2024), we include another experiment in this domain where we replace the output of *Aspect Identification* with human-annotated text fragments. According to Table 7, both *Aspect Identification* and *Opinion Consolidation* are crucial to generating more faithful meta-reviews and with higher aspect coverage, however *Opinion Consolidation* appears to be the most critical as its removal decreases performance across domains (exception: coverage for research articles). We also observe that model-extracted text fragments are on par with human-selected ones but more helpful to generating faithful meta-reviews.

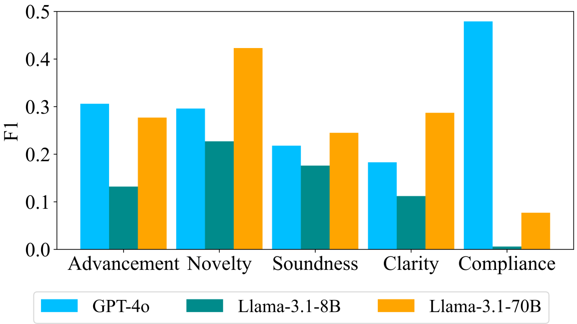


Figure 2: Evaluation of text fragments extracted for individual review aspects by *Aspect Identification*.

Domain	Modules	Coverage↑	AlignScore-S↑
Hotels	AI+OC+MS	**0.99**	0.80
	OC+MS	**0.99**	**0.83**
	AI+MS	0.55	0.62
Shoes	AI+OC+MS	**0.83**	**0.74**
	OC+MS	0.69	0.72
	AI+MS	0.61	0.69
Research Articles	AI+OC+MS	0.97	**0.79**
	OC+MS	**0.98**	0.78
	AI+MS	0.97	0.75
	$AI^{\dagger}$+OC+MS	0.97	0.69

Table 7: Ablations quantifying the contribution of different modules on three domains (hotels, shoes, research articles). AI: *Aspect Identification*, OC: *Opinion Consolidation*, MS: *Meta-Review Synthesis*, $AI^{\dagger}$: text fragments selected by humans. Results shown for *Aspect-aware decomposition-Llama 70B*.

Humans prefer meta-reviews generated by our modular system to gold-standard references. We conduct human evaluation to verify that our approach generates meta-reviews that reflect the review aspects of the input and are overall coherent and faithful. We recruited crowdworkers through Prolific[17], selected to be L1 English speakers from the US or UK, and compensated above the UK living wage at 12GBP/hr. We ask crowdworkers to read a set of source reviews followed by two generated meta-reviews and select which meta-review is best (allowing for ties) along two dimensions, as

[16] https://github.com/oaimli/MetaReviewingLogic

[17] https://www.prolific.com/

Model	Cover↑	Faith↑	Overall↑
Research Articles			
Sentiment CoT-GPT-4o	0%	0%	0%
Human-written reference	80%	80%	80%
Automatic decomposition-Llama 70B	90%	90%	90%
Chunk-wise decomposition-Llama 70B	70%	90%	90%
Naive aspect-aware prompting-Llama 70B	0%	0%	10%
Aspect-aware decomposition-GPT-4o	10%	50%	50%
Sports Shoes			
HIRO-abs	90%	90%	90%
Human-written reference	90%	90%	90%
Automatic decomposition-Llama 70B	100%	90%	100%
Chunk-wise decomposition-Llama 70B	80%	80%	70%
Naive aspect-aware prompting-Llama 70B	20%	20%	40%
Aspect-aware decomposition-GPT-4o	10%	20%	30%
Hotels			
HIRO-abs	80%	100%	100%
Human-written reference	30%	70%	100%
Automatic decomposition-Llama 70B	90%	100%	100%
Chunk-wise decomposition-Llama 70B	50%	60%	80%
Naive aspect-aware prompting-Llama 70B	100%	100%	100%
Aspect-aware decomposition-GPT-4o	0%	0%	10%

Table 8: Proportion of times (%) crowdworkers preferred our model (*Aspect-aware decomposition-Llama 70B*) against depicted systems. We highlight in red comparisons where our model is chosen as better more than 50% of the time (higher is better). For example, '90%' means that crowdworkers prefer our system on 9 out of 10 entities. We take a majority vote to determine a single system preference.

well as an overall preference:

- **Coverage** — Which meta-review covers more review aspects in the source reviews?
- **Faithfulness** — Which meta-review has a higher percentage of opinions supported by the source reviews?
- **Overall** — Which meta-review do you think is better overall?

We randomly select ten entities for each dataset (SPACE, AmaSum, and PeerSum) and construct six pairwise combinations between our approach (Aspect-aware decomposition with Llama-3.1-70B) and the systems shown in Table 8, including human-written reference meta-reviews. For AmaSum and SPACE, we only present crowdworkers with 20% of the reviews for each entity, to maintain a reasonable workload (reviews are sampled randomly). We elicit three annotations for each pairwise combination of system outputs, leading to a total of 1,260 ratings. Annotators have reasonable agreement, with average values of Krippendorff's α being 0.335 on shoes, 0.622 on hotels, and 0.463 on research articles. More details on experimental design and the full instructions are in Appendix G.

Present Reasoning Steps	Time↓	Preferred↑
No reasoning steps	10.9	20%
Automatic decomposition	10.3	20%
Aspect-aware decomposition (ours)	9.3	40%

Table 9: Average time (in minutes) humans take to write scientific meta-reviews and the proportion of times participants prefer meta-reviews when present with different intermediate reasoning steps (in exhausted pair-wise comparison).

Table 8 shows the proportion of times (%) crowdworkers prefer our approach against a comparison system. We find that human judgments are broadly consistent with automatic evaluation. Crowdworkers prefer our system to human references on two (shoes and research articles) out of three domains. We consistently win against automatic and chunk-wise decompositions (with Llama 70B), but lose against our own decompositions with GPT-4o.

Aspect-aware decomposition allows humans to create better summaries faster. We also evaluate the intermediate outputs produced by our modules. In particular, we examine whether the specific module decomposition adopted by our system is useful for real-world meta-review writing. We ask annotators to write meta-reviews for hotel reviews in three conditions: (1) they are not given any intermediate reasoning steps; (2) they are given reasoning steps produced by automatic knowledge-agnostic decomposition from *Automatic decomposition-Llama 70B*; and (3) they are provided with the intermediate outputs of our modules with *Aspect-aware decomposition-Llama 70B* as reasoning steps. We record the time it takes crowdworkers to finish the writing.

We randomly select ten entities and obtain three meta-reviews for each (according to the three conditions described above). We recruit five annotators, however, each annotator writes a meta-review for each entity once to avoid memorization. Based on the time reported in Table 9, we find that providing intermediate outputs of our aspect-aware decomposition accelerates participants' writing compared with the other two conditions, reducing the time of writing a meta-review by 14.7% (on average). More details about how we present different reasoning steps to annotators and annotation instructions are provided in Appendix H.

We also ask another set of annotators to assess the meta-reviews written above, by presenting pairwise comparisons (following the instructions of human annotation presented in the previous section). We find that participants prefer meta-reviews written based on the outputs of our modules twice as much compared to the other two settings (Krippendorff's α is 0.542).

6 Conclusion

We propose modular decomposition for opinion summarization based on review aspects. Our decomposition is evidence-based (the output of each module can be traced back to its input), enabling greater transparency and ease of inspection. Extensive experiments demonstrate that our modular framework outperforms state-of-the-art methods and other strong baselines in multiple domains. Human evaluations reveal that our approach not only produces higher-quality meta-reviews but also generates more useful intermediate outputs to assist humans in composing meta-reviews. While our work focuses on opinion summarization, the concept of aspect-aware decomposition holds promise for other complex language generation tasks.

Limitations

Despite promising results, our experimental findings are currently limited to English. Furthermore, the prompts used in our modular approach were not optimized in any way, suggesting a potential area for future improvement.

Our approach operates on a set of predefined aspects, whose definitions were sourced from previous work (Angelidis et al., 2021; Li et al., 2024). While reviewing platforms often feature similar criteria (e.g., `https://runrepeat.com/hoka-bondi-8` for the domain of sports shoes), our work prioritizes enhancing the grounding and transparency of opinion summarization. Therefore, the broader task of aspect engineering, including unsupervised methods for aspect discovery, falls outside the scope of this investigation.

Our meta-review synthesis module, aims to cover all aspects mentioned in the original reviews without considering their relative importance. We could easily filter the aspects based on the size of their corresponding text fragments. However, we found this approach is not universally applicable across all domains. In the scientific review domain, for example, *Advancement* opinions are more frequent than *Novelty* ones (25% vs. 14%) but both aspects are equally important (Li et al., 2024). We defer further investigation into aspect importance to future work.

Finally, our approach does not explicitly address the potential generation of biased or harmful content, even though our goal is to ensure that the generated meta-reviews remain grounded in the original content.

Ethics Statement

Our work primarily focuses on enhancing the capabilities of AI systems to assist humans, rather than aiming to replace them. As demonstrated in our experiments, the intermediate outputs generated by our approach can help humans produce higher-quality meta-reviews with greater efficiency.

Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable feedback, which greatly helped improve this work. We would also like to thank Tom Hosking and Alex Gurung for their help with the human annotation experiments. This work was partially supported by the Australian Research Council Discovery Grant awarded to Jey Han Lau (ID: DP240101006) and the UK Engineering and Physical Sciences Research Council (grant EP/W002876/1) awarded to Mirella Lapata.

References

Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. 2021a. Aspect-controllable opinion summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6578–6593, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. 2021b. Unsupervised opinion summarization with content planning. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 12489–12497. AAAI Press.

Stefanos Angelidis, Reinald Kim Amplayo, Yoshihiko Suhara, Xiaolan Wang, and Mirella Lapata. 2021. Extractive opinion summarization in quantized transformer spaces. *Transactions of the Association for Computational Linguistics*, 9:277–293.

Regina Barzilay and Kathleen R. McKeown. 2005. Sentence fusion for multidocument news summarization. *Computational Linguistics*, 31(3):297–328.

Somnath Basu Roy Chowdhury, Chao Zhao, and Snigdha Chaturvedi. 2022. Unsupervised extractive opinion summarization using sparse coding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1209–1225, Dublin, Ireland. Association for Computational Linguistics.

Adithya Bhaskar, Alexander R. Fabbri, and Greg Durrett. 2023. Prompted opinion summarization with GPT-3.5. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 9282–9300. Association for Computational Linguistics.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020. Unsupervised opinion summarization as copycat-review generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5151–5169, Online. Association for Computational Linguistics.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2021. Learning opinion summarizers by selecting informative reviews. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9424–9442, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Arie Cattan, Lilach Eden, Yoav Kantor, and Roy Bar-Haim. 2023. From key points to key point hierarchy: Structured and expressive opinion summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 912–928, Toronto, Canada. Association for Computational Linguistics.

Eric Chu and Peter J. Liu. 2019. Meansum: A neural model for unsupervised multi-document abstractive summarization. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 1223–1232. PMLR.

Maximin Coavoux, Hady Elsahar, and Matthias Gallé. 2019. Unsupervised aspect-based multi-document abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 42–47, Hong Kong, China. Association for Computational Linguistics.

Lea Frermann and Alexandre Klementiev. 2019. Inducing document structure for aspect-based summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6263–6273, Florence, Italy. Association for Computational Linguistics.

Tom Hosking, Hao Tang, and Mirella Lapata. 2023. Attributable and scalable opinion summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8488–8505, Toronto, Canada. Association for Computational Linguistics.

Tom Hosking, Hao Tang, and Mirella Lapata. 2024. Hierarchical indexing for retrieval-augmented opinion summarization. *Transactions of the Association for Computational Linguistics*, 12:1533–1555.

Hayate Iso, Xiaolan Wang, Yoshihiko Suhara, Stefanos Angelidis, and Wang-Chiew Tan. 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.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2023. Decomposed prompting: A modular approach for solving complex tasks. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Kundan Krishna, Sopan Khosla, Jeffrey Bigham, and Zachary C. Lipton. 2021. Generating SOAP notes from doctor-patient conversations using modular summarization techniques. 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 4958–4972, Online. Association for Computational Linguistics.

Logan Lebanoff, Franck Dernoncourt, Doo Soon Kim, Walter Chang, and Fei Liu. 2020. A cascade approach to neural abstractive summarization with content selection and fusion. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 529–535, Suzhou, China. Association for Computational Linguistics.

Haoyuan Li, Somnath Basu Roy Chowdhury, and Snigdha Chaturvedi. 2023a. Aspect-aware unsupervised extractive opinion summarization. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 12662–12678. Association for Computational Linguistics.

Miao Li, Eduard Hovy, and Jey Lau. 2023b. Summarizing multiple documents with conversational structure for meta-review generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7089–7112, Singapore. Association for Computational Linguistics.

Miao Li, Jey Han Lau, and Eduard Hovy. 2024. A sentiment consolidation framework for meta-review generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10158–10177, Bangkok, Thailand. Association for Computational Linguistics.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 150–157.

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

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Dragomir R. Radev and Kathleen R. McKeown. 1998. Generating natural language summaries from multiple on-line sources. *Computational Linguistics*, 24(3):469–500.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. In *OpenAI Blog*.

Aviv Slobodkin, Ori Shapira, Ran Levy, and Ido Dagan. 2024. Multi-review fusion-in-context. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3003–3021, Mexico City, Mexico. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V. Le, and Ed H. Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

A Prompts for Aspect-aware Decomposition

In this section we provide the prompt templates used to decompose opinion summarization into the modules of *Aspect Identification*, *Opinion Consolidation*, and *Meta-review Synthesis*. Domain-specific prompts are provided in Sections B–D.

Aspect Identification

You are good at understanding documents with {domain} review opinions.
Below is a {domain} review for an academic manuscript, please extract fragments that are related to {the-review-aspect} of the {the entity}.
Definition of {the review aspect}:{the definition of the review aspect}
Example input review:
{the example input review}
Example format of extracted fragments in different lines:
{the example output}
Target input review:
{input-document}
Final extracted fragments (follow the format above in different lines and if no resulted fragments just output "No related fragments"):

Figure 3: The few-shot prompt template for the *Aspect Identification* module; text fragments are extracted for each (domain) aspect. Please note that for research articles we use few-shot prompting to enable the model follow the output format while for sports shoes and hotels zero-shot prompting (with just removing the demonstration example) could get reasonable performances.

Opinion Consolidation

You are good at writing summaries for opinionated texts. You are given some opinionated text fragments, please write a concise summary for them.
Example input review fragments:
{the example text fragments}
Example summary of the input fragments:
{the example aspect-specific meta-review of the input fragments}
Target input fragments:
{input-fragments}
The final summary of these target input text fragments (just output the answer without any other content):

Figure 4: The few-shot prompt template for the *Opinion Consolidation* module; it outputs summaries for individual review aspects. Please note that for research articles we use few-shot prompting to get better performance while for sports shoes and hotels zero-shot prompting (with just removing the demonstration example) could get reasonable performances.

Meta-Review Synthesis

You are good at understanding documents with {domain} review opinions.
Below are comments on different review aspects for {the entity}, please write a concise and natural meta-review which summaries the provided comments and covers all mentioned review aspects.
Comments on different aspects:
{meta-reviews of individual review aspects}
The meta-review is (directly output the answer without any other content):

Figure 5: The prompt template for the *Meta-Review Synthesis* module based on aspect-specific meta-reviews from the *Opinion Consolidation* module. As zero-shot prompting gives us reasonable performances on all the three datasets, we used the same zero-shot prompt template for the module.

B Prompts for Scientific Reviews of Research Articles

Prompts for *Aspect Identification* are given in Tables 6–10 for the aspects *Advancement*, *Clarity*, *Compliance*, *Soundness*, and *Novelty*. The prompt for *Opinion Consolidation* is in Table 11 and all aspects share the same prompt for this module. The prompt for *Meta-Review Synthesis* is in Table 12.

Aspect Identification: Advancement

You are good at understanding documents with scientific review opinions.
Below is a scientific review for an academic manuscript, please extract text fragments that are related to Advancement of the research work.

Definition of Advancement:

Importance of the manuscript to discipline, significance of the contributions of the manuscript, and its potential impact to the field.

Example input review:

This paper theoretically studied one of the fundamental issue in CycleGAN (recently gained much attention for image-to-image translation). The authors analyze the space of exact and approximated solutions under automorphisms. Reviewers mostly agree with theoretical value of the paper. Some concerns on practical values are also raised, e.g., limited or no-surprising experimental results. In overall, I think this is a boarderline paper. But, I am a bit toward acceptance as the theoretical contribution is solid, and potentially beneficial to many future works on unpaired image-to-image translation.

Example output fragments in different lines:

Some concerns on practical values are also raised, e.g., limited or no-surprising experimental results.

Reviewers mostly agree with theoretical value of the paper.

But, I am a bit toward acceptance as the theoretical contribution is solid, and potentially beneficial to many future works on unpaired image-to-image translation.

Target input review:

{{input_document}}
Final extracted fragments (follow the format above in different lines and if no resulted fragments just output "No related fragments"):

Figure 6: The prompt of *Aspect Identification* for the aspect *Advancement*.

C Prompts for Business Reviews of Hotels

Prompts for *Aspect Identification* on hotels are shown in Tables 13–18 for the aspects *Building*, *Cleanliness*, *Food*, *Location*, *Rooms*, and *Service*. The prompt for *Opinon Consolidation* for any review aspect is in Table 19. The prompt for *Meta-Review Synthesis* is present in Table 20.

D Prompts for Product Reviews of Sports Shoes

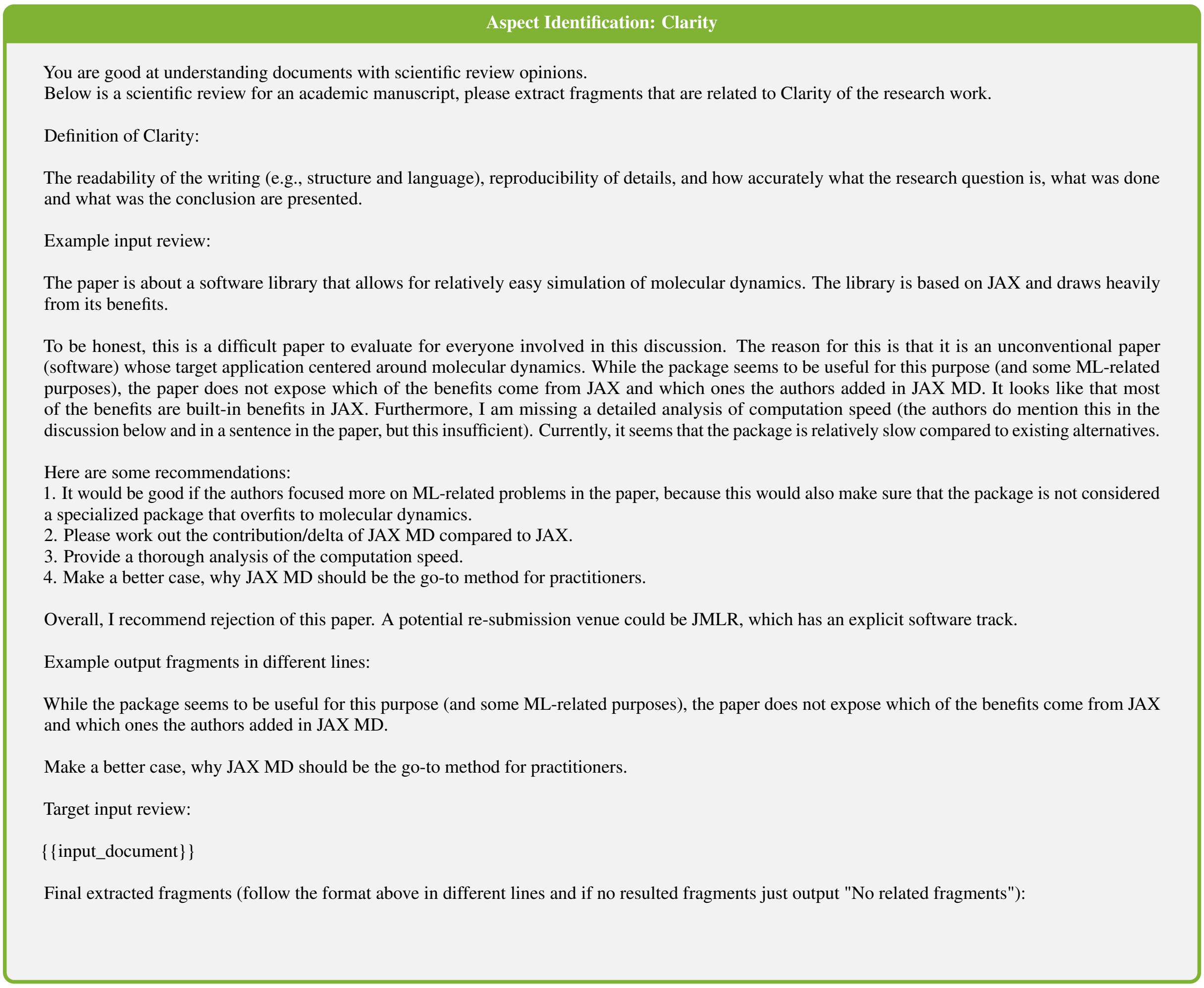
Prompts for *Aspect Identification* are given in Tables 21–30 for the aspects *Breathability*, *Comfort*, *Cushioning*, *Durability*, *Flexibility*, *Misc*, *Size and Fit*, *Stability*, *Traction*, and *Weight*. The prompt for *Opinion Consolidation* for any aspect is in Table 31. The prompt for *Meta-Review Synthesis* is in Table 32.

E Implementation Details of Comparison Models

In this section we provide implementation details for the various comparison models used in our experiments.

- For HIRO-abs (Hosking et al., 2024), we obtain generations for AmaSum and SPACE from https://github.com/tomhosking/hiro. There are three outputs for each entity and we use the first one as the generation of HIRO-abs.

- For TCG (Bhaskar et al., 2023), we made some adaptation to get fair comparison. TCG only generates aspect-oriented summaries instead of an overall global summary, which we have to aggregate to ensure a fair comparison with our approach. We obtain their released aspect-oriented summaries and use the open-source Llama 70B to generate an overall summary. We use the same version of



Aspect Identification: Clarity

You are good at understanding documents with scientific review opinions.
Below is a scientific review for an academic manuscript, please extract fragments that are related to Clarity of the research work.

Definition of Clarity:

The readability of the writing (e.g., structure and language), reproducibility of details, and how accurately what the research question is, what was done and what was the conclusion are presented.

Example input review:

The paper is about a software library that allows for relatively easy simulation of molecular dynamics. The library is based on JAX and draws heavily from its benefits.

To be honest, this is a difficult paper to evaluate for everyone involved in this discussion. The reason for this is that it is an unconventional paper (software) whose target application centered around molecular dynamics. While the package seems to be useful for this purpose (and some ML-related purposes), the paper does not expose which of the benefits come from JAX and which ones the authors added in JAX MD. It looks like that most of the benefits are built-in benefits in JAX. Furthermore, I am missing a detailed analysis of computation speed (the authors do mention this in the discussion below and in a sentence in the paper, but this insufficient). Currently, it seems that the package is relatively slow compared to existing alternatives.

Here are some recommendations:
1. It would be good if the authors focused more on ML-related problems in the paper, because this would also make sure that the package is not considered a specialized package that overfits to molecular dynamics.
2. Please work out the contribution/delta of JAX MD compared to JAX.
3. Provide a thorough analysis of the computation speed.
4. Make a better case, why JAX MD should be the go-to method for practitioners.

Overall, I recommend rejection of this paper. A potential re-submission venue could be JMLR, which has an explicit software track.

Example output fragments in different lines:

While the package seems to be useful for this purpose (and some ML-related purposes), the paper does not expose which of the benefits come from JAX and which ones the authors added in JAX MD.

Make a better case, why JAX MD should be the go-to method for practitioners.

Target input review:

{{input_document}}

Final extracted fragments (follow the format above in different lines and if no resulted fragments just output "No related fragments"):

Figure 7: The prompt of *Aspect Identification* for the aspect of *Clarity*.

Llama 70B as in our experiments since the GPT-3.5 model used in their implementation has been deprecated.

- For fine-tuning Llama-3.1-8B, we trained the model with Transformers from Huggingface on the three datasets for 5 epochs on four NVIDIA A100 80G GPUs, with `max-predict-length=512`, `bf16=True`, `batch-size=1`, `optim=adafactor`, `learning-rate=1e-6`, `warmup-rate=0.2`, `label-smoothing-factor=0.1`, `lr-scheduler-type=cosine`, `fsdp='full_shard auto_wrap offload'`.

- For *naive aspect-aware prompting*, we only incorporate aspect descriptions into the prompt. As an example, we show the prompt for scientific reviews in Figure 33.

- For *Automatic decomposition* (Khot et al., 2023), the prompting approach cannot be directly transferred to opinion summarization. Based on the idea of automatic decomposition, we implement automatic knowledge-agnostic decomposition on our experimental datasets. The idea is to first generate intermediate reasoning steps and then follow those steps in sequence to generate the final meta-review. We provide example prompts for scientific reviews in Figure 34 and 35.

- For *chunk-wise decomposition* (Khot et al., 2023), we first generate small meta-reviews for each chunk, and then combine all chunk-specific meta-reviews with another prompt to generate the global meta-review. Example prompts for scientific reviews are shown in Figures36 and 37.

Aspect Identification: Compliance

You are good at understanding documents with scientific review opinions.
Below is a scientific review for an academic manuscript, please extract fragments that are related to Compliance of the research work.

Definition of Compliance:

Whether the manuscript fits the venue, and all ethical and publication requirements are met.

Example input review:

"The paper proposes a method to identify and correct regions on the data manifold in which a trained classifier fails. The *identification* phase is based on clustering classification failure regions in a GAN latent space and the *correction* phase is based on fine-tuning the classifier with additional synthetic samples from the GAN. The proposed method is strongly based on Zhao et al 2018 (Generating Natural Adversarial Examples), a method to generate on-manifold black-box adversarial examples using a GAN. The authors of the current paper describe some differences of their identification step from Zhao et al (end of section 3.2.1), but in my opinion they are minor. The main contribution of the current paper over Zhao et al seems to be clustering the adversarial examples (using GMM) and using them to fine-tune the classifier. This, in my opinion, is potentially an interesting idea, however, the authors do not show sufficient evidence of its success. Specifically, the authors claim to "achieve near perfect failure scenario accuracy with minimal change in test set accuracy", but they do not provide any details (e.g. table of accuracy values on the train, test and adversarial sets before and after the fine-tuning). I would also expect to see an ablation study comparing the proposed method to simply including the adversarial examples found using Zhao et al (w/o GMM fitting and sampling) as additional training example - a standard adversarial defense approach (see e.g. [1]).Perhaps more importantly, the objective of the proposed method is not, in my opinion, clear. The title and abstract describe the goal as "debugging" a classifier and correcting fail regions, however the described method seems like a defense against on-manifold adversarial attack. If the method, as claimed, helps debugging and correcting the classifier, I would expect to see an improved accuracy on the (natural) unseen test set - not just on the synthetically generated adversarial examples. The quality and clarity of the writing can be improved as well. A lot of space is allocated to describing well-known methods (e.g. VAE, GMM), however, critical information about the experimental results are missing. I'm also not sure all the formally defined algorithms and equations actually help in the understanding (e.g. algorithm 1, equation 2). Some of the mathematical notations are not standard. Minor comment: The norm in definition 3.1 is a regular vector norm (l2?) and not a matrix norm. To summarize: pros: - interesting idea (clustering on-manifold failures, labeling them and then using them to improve the classifier)cons:- contribution over Zhao et al not well established- insufficient and inaccurate experimental results- general quality of writing - not sure actual work and experiments match the stated objective - significance *Update:* Following the authors' response, I upgraded my rating, but I still think there are critical issues with the paper. The most problematic point, in my opinion, is the only-marginal improvement on the test data, indicating that the suggested training method only improves the specific "failure scenarios", making it is similar to adversarial training methods used to gain adversarial robustness. However, the abstract and introduction indicates that the paper helps in debugging in fixing failures in general, which, I think should have been evident in improved test accuracy.[1] Zhang, Hongyang, et al. "Theoretically principled trade-off between robustness and accuracy."ICML 2019

Example output fragments in different lines:

Some of the mathematical notations are not standard.

Target input meta-review:

{{input_document}}

Final extracted fragments (follow the format above in different lines and if no resulted fragments just output "No related fragments"):

Figure 8: The prompt of *Aspect Identification* for the aspect of *Compliance*.

Aspect Identification: Soundness

You are good at understanding documents with scientific review opinions.
Below is a scientific meta-review for an academic manuscript, please extract fragments that are related to Soundness of the research work.

Definition of Soundness: There are usually two types of soundness: (1) Empirical: how well experiments are designed and executed to support the claims, whether methods used are appropriate, and how correctly the data and results are reported, analysed, and interpreted. (2) Theoretical: whether arguments or claims in the manuscript are well supported by theoretical analysis, i.e., completeness, and the methodology (e.g., mathematical approach) and the analysis is correct.

Example input meta-review:

The paper proposes to use the mirror descent algorithm for the binary network. It is easy to read. However, novelty over ProxQuant is somehow limited. The theoretical analysis is weak, in that there is no analysis on the convergence and neither how to choose the projection for mirror mapping construction. Experimental results can also be made more convincing, by adding comparisons with bigger datasets, STOA networks, and ablation study to demonstrate why mirror descent is better than proximal gradient descent in this application.

Example output fragments in different lines:

The theoretical analysis is weak, in that there is no analysis on the convergence and neither how to choose the projection for mirror mapping construction.

Experimental results can also be made more convincing, by adding comparisons with bigger datasets, STOA networks, and ablation study to demonstrate why mirror descent is better than proximal gradient descent in this application.

Target input meta-review:

{{input_document}}

Final extracted fragments (follow the format above in different lines and if no resulted fragments just output "No related fragments"):

Figure 9: The prompt of *Aspect Identification* for the aspect of *Soundness*.

Aspect Identification: Novelty

You are good at understanding documents with scientific review opinions.
Below is a scientific meta-review for an academic manuscript, please extract fragments that are related to Novelty of the research work.

Definition of Novelty:

How original the idea (e.g., tasks, datasets, or methods) is, and how clear where the problems and methods sit with respect to existing literature (i.e., meaningful comparison).

Example input meta-review:

The manuscript describes a method for identifying and correcting classifier performance when labels are assigned incorrectly. The identification is based on clustering classification failure regions in a VAE latent space and the correction phase is based on fine-tuning the classifier with additional synthetic samples from the VAE.

Reviewers agreed that the manuscript is not ready for publication. The main issue is that the suggested training method is similar to adversarial training methods used to gain adversarial robustness. The method does not help in debugging and fixing failures in general.

Example output fragments in different lines:

Reviewers appreciated the novelty, introducing a new simpler routing mechanism, and achieving good performance on real world datasets.

In particular, removing the squash function and experimenting with concurrent routing was highlighted as significant progress.

Alongside with them, I acknowledge the novelty of using layer norm and parallel execution, and recommend accept.

Target input meta-review:

{{input_document}}

Final extracted fragments (follow the format above in different lines and if no resulted fragments just output "No related fragments"):

Figure 10: The prompt of *Aspect Identification* for the aspect of *Novelty*.

Opinion Consolidation

You are good at writing summaries for opinionated texts. You are given some opinionated text fragments, please write a concise summary for them.
Example input review fragments:

"1) **Evaluating different explanation techniques:**",
"We thus believe that our results do *not* violate the surmise made in the shared reference, but rather support it.",
"We believe this makes our findings generalizable.",
"Although, the paper brings out the importance of analogies as explanations (which further motivates our work)",
"The proposed technique is flexible as it can provide two forms of explanations: feature and analogy-based.",
"Moreover, explanations in the form of analogies are intuitive for human users.",
"We feel that analogous examples do not need to share common words, content, or sentence structure. What is important is that they *point to latent factors* that may be responsible for the model's output.",
"**Purpose of analogies:**",
...
"The authors solved this problem by the use of a learned local distance matrix, in which interaction effects are clearly shown."

Example summary of the input fragments:

The proposed approach to explain similarity prediction is a relatively less explored area, which makes the problem addressed and the proposed method unique.

Example input review fragments:
"The paper is technically sound, and the claims are carefully developed and well supported.",
"The manuscript is well structured and very clearly written, with helpful introductions to the methodological ingredients that it builds upon.",
"The paper could be further improved with some reflection on the limitations of the approach.",
"I am not certain how large a contribution it will have to the field of Bayesian inference in general.",
...
"I'll use the rest of the section for high-level comments.",
"- In its current form, the paper convinces me that SHF decreases runtime and increases performance for datasets with low complexity."

Example summary of the input fragments:

Based on these, I recommend acceptance for this paper. All reviewers agree that the paper proposes an interesting approach to Bayesian inference incorporating coresets with Hamiltonian flows.

Target input review fragments:

{{review_fragments}}

The final summary of these target input text fragments (just output the answer without any other content):

Figure 11: The prompt of *Opinion Consolidation* for any aspect of scientific reviews.

Meta-review Synthesis

You are good at understanding documents with scientific review opinions.
Below are comments on different review aspects for an academic manuscript, please write a concise and natural meta-review which summaries the provided comments and covers all mentioned review aspects.

Comments on different aspects:

{{metas_generated}}

The meta-review is (directly output the answer without any other content):

Figure 12: The prompt of *Meta-Review Synthesis* for research articles.

Aspect Identification: Building

You are good at understanding documents with hotel review opinions.
Below is a business review for a hotel, please extract fragments that are related to Building of the hotel.

Definition of Building:
Analysis of how well the hotel was constructed, its design, functionality, and how these factors contribute to the success and satisfaction of its guests.

Target input review:

{{input_document}}

Final extracted fragments (follow the format above in different lines and if no resulted fragments just output "No related fragments"):

Figure 13: The prompt of *Aspect Identification* for the aspect of *Building*.

Aspect Identification: Cleanliness

You are good at understanding documents with hotel review opinions.
Below is a business review for a hotel, please extract fragments that are related to Cleanliness of the hotel.

Definition of Cleanliness:
Evaluation of how well the hotel maintains a clean, sanitary, and comfortable environment for its guests, impacting their overall experience and satisfaction.

Target input review:

{{input_document}}

Final extracted fragments (follow the format above in different lines and if no resulted fragments just output "No related fragments"):

Figure 14: The prompt of *Aspect Identification* for the aspect of *Cleanliness*.

Aspect Identification: Food

You are good at understanding documents with hotel review opinions.
Below is a business review for a hotel, please extract fragments that are related to Food of the hotel.

Definition of Food:
Evaluation of the dining experience including the quality and variety of the food, ultimately affecting guest satisfaction and the hotel's reputation.
Target input review:
{{input_document}}

Final extracted fragments (follow the format above in different lines and if no resulted fragments just output "No related fragments"):

Figure 15: The prompt of *Aspect Identification* with the aspect of *Food*.

F Implementation Details for Automatic Evaluation

Implementation details of G-Eval (Liu et al., 2023) are presented in Figures 38, 39, and 40 for the three domains, respectively. We use gpt-4o-2024-05-13 as the backbone LLM of G-Eval.

G Details of Human Evaluation on Quality of Generated Meta-Reviews

We conduct human evaluation based on pair-wise comparisons to verify the quality of our generated meta-reviews (in terms of aspect coverage and opinion faithfulness). We recruited crowdworkers through

Aspect Identification: Location

You are good at understanding documents with hotel review opinions.
Below is a business review for a hotel, please extract fragments that are related to Location of the hotel.

Definition of Location:
Analysis of how the hotel's location influences the guest experience, considering factors like convenience, safety, proximity to attractions, and the overall environment.

Target input review:

{{input_document}}

Final extracted fragments (follow the format above in different lines and if no resulted fragments just output "No related fragments"):

Figure 16: The prompt of *Aspect Identification* for the aspect of *Location*.

Aspect Identification: Rooms

You are good at understanding documents with hotel review opinions.
Below is a business review for a hotel, please extract fragments that are related to Rooms of the hotel.

Definition of Rooms:
Assessment of how well the room meets the guest's needs and expectations in terms of comfort, cleanliness, amenities, and overall experience.

Target input review:

{{input_document}}

Final extracted fragments (follow the format above in different lines and if no resulted fragments just output "No related fragments"):

Figure 17: The prompt of *Aspect Identification* for the review aspect of *Rooms*.

Aspect Identification: Service

You are good at understanding documents with hotel review opinions.
Below is a business review for a hotel, please extract fragments that are related to Service of the hotel.

Definition of Service:
Assessment of how well the hotel staff and management meet the needs of their guests, impacting their comfort, convenience, and overall experience.

Target input review:

{{input_document}}

Final extracted fragments (follow the format above in different lines and if no resulted fragments just output "No related fragments"):

Figure 18: The prompt of *Aspect Identification* with the aspect of *Service*.

Opinion Consolidation

You are good at writing summaries for opinionated texts. You are given some opinionated text fragments, please write a concise summary for them.

Target input review fragments:

{{review_fragments}}

The final summary of these target input text fragments (just produce the answer without any other content):

Figure 19: The prompt of *Opinion Consolidation* for any individual review aspect for hotels.

Meta-Review Synthesis

You are good at understanding documents with hotel review opinions.
Below are business reviews in different aspects for a hotel, please write a concise and natural meta-review which summaries the provided comments and covers all mentioned review aspects.

Comments on different aspects:

{{metas_generated}}
The meta-review is (directly output the answer without any other content):

Figure 20: The prompt of *Meta-Review Synthesis* for hotels.

Aspect Identification: Breathability

```
You are good at understanding documents with sports shoes review opinions.
Below is a product review for a pair of shoes, please extract fragments that are related to Breathability of shoes.

Definition of Breathability:
Evaluation about breathability of the shoes.

Target input review:

{{input_document}}

Final extracted fragments (follow the format above in different lines and if no resulted fragments just output "No related fragments"):
```

Figure 21: The prompt of *Aspect Identification* for the aspect of *Breathability*.

Aspect Identification: Comfort

```
You are good at understanding documents with sports shoes review opinions.
Below is a product review for a pair of shoes, please extract fragments that are related to Comfort of shoes.

Definition of Comfort:
Evaluation about comfort of the shoes, such as tongue padding, heel tab, and removable insole.

Target input review:
{{input_document}}

Final extracted fragments (follow the format above in different lines and if no resulted fragments just output "No related fragments"):
```

Figure 22: The prompt of *Aspect Identification* with the aspect of *Comfort*.

Aspect Identification: Cushioning

```
You are good at understanding documents with sports shoes review opinions.
Below is a product review for a pair of shoes, please extract fragments that are related to Cushioning of shoes.

Definition of Cushioning:
Evaluation about cushioning of the shoes, such as heel stack and forefoot stack.

Target input review:

{{input_document}}

Final extracted fragments (follow the format above in different lines and if no resulted fragments just output "No related fragments"):
```

Figure 23: The prompt of *Aspect Identification* for the review aspect of *Cushioning*.

Aspect Identification: Breathability

```
You are good at understanding documents with sports shoes review opinions.
Below is a product review for a pair of shoes, please extract fragments that are related to Durability of shoes.

Definition of Durability:
Evaluation about durability of the shoes, such as outsole hardness and thickness.

Target input review:

{{input_document}}

Final extracted fragments (follow the format above in different lines and if no resulted fragments just output "No related fragments"):
```

Figure 24: The prompt of *Aspect Identification* with the aspect of *Durability*.

Prolific[18] with compensation above the UK living wage at £12 per working hour.

For product reviews of sports shoes, we randomly select ten entities from the test data of AmaSum. Based on generated meta-reviews, for each entity we construct six pairs of comparisons between our modular approach with Llama-3.1-70B as a backbone and comparison baselines. There are originally about 400 source reviews in each entity and it is hard for humans to review all of them. To balance annotator workload, we present annotators with 20% reviews and randomly select reviews for three times to ensure experimental consistency. Therefore, there are 18 pairs of comparisons for each entity. Each

[18] www.prolific.com

Aspect Identification: Flexibility

You are good at understanding documents with sports shoes review opinions.
Below is a product review for a pair of shoes, please extract fragments that are related to Flexibility of shoes.

Definition of Flexibility:
Evaluation about flexibility of the shoes, such as stiffness, stiffness in the cold, and difference in stiffness in the cold.

Target input review:

{{input_document}}
Final extracted fragments (follow the format above in different lines and if no resulted fragments just output "No related fragments"):

Figure 25: The prompt of *Aspect Identification* with the review aspect of *Flexibility*.

Aspect Identification: Misc

You are good at understanding documents with sports shoes review opinions.
Below is a product review for a pair of shoes, please extract fragments that are related to Misc of shoes.

Definition of Misc:
Evaluation about reflective elements of the shoes.

Target input review:

{{input_document}}

Final extracted fragments (follow the format above in different lines and if no resulted fragments just output "No related fragments"):

Figure 26: The prompt of Aspect Identification with the review aspect of Misc.

Aspect Identification: Size and Fit

You are good at understanding documents with sports shoes review opinions.
Below is a product review for a pair of shoes, please extract fragments that are related to Size and Fit of shoes.

Definition of Size and Fit:
Evaluation about size and fit of the shoes, such as internal length, toebox width at the widest part, and gusset type.

Target input review:

{{input_document}}

Final extracted fragments (follow the format above in different lines and if no resulted fragments just output "No related fragments"):

Figure 27: The prompt of *Aspect Identification* for the aspect of *Size and Fit*.

Aspect Identification: Stability

You are good at understanding documents with sports shoes review opinions.
Below is a product review for a pair of shoes, please extract fragments that are related to Stability of shoes.

Definition of Stability:
Evaluation about stability of the shoes, such as torsional rigidity, heel counter stiffness, midsole width in the forefoot and midsole width in the heel.

Target input review:

{{input_document}}

Final extracted fragments (follow the format above in different lines and if no resulted fragments just output "No related fragments"):

Figure 28: The prompt of *Aspect Identification* for the aspect of *Stability*.

pair is rated by three different annotators and we obtain 540 annotations for the dataset.

We recruited 27 annotators from Prolific with L1 English from the US or UK, with a minimum approval rate of 100% in more than 100 studies. In addition to the attention check question for each annotation instance, we also included quality control instances, asking participants to distinguish human-written

Aspect Identification: Traction

You are good at understanding documents with sports shoes review opinions.
Below is a product review for a pair of shoes, please extract fragments that are related to Traction of shoes.

Definition of Traction:
Evaluation about traction of the shoes, such as lug depth.

Target input review:

{{input_document}}

Final extracted fragments (follow the format above in different lines and if no resulted fragments just output "No related fragments"):

Figure 29: The prompt of *Aspect Identification* for the review aspect of *Traction*.

Aspect Identification: Weight

You are good at understanding documents with sports shoes review opinions.
Below is a product review for a pair of shoes, please extract fragments that are related to Weight of shoes.

Definition of Weight:
Evaluation about weight of the shoes.

Target input review:

{{input_document}}

Final extracted fragments (follow the format above in different lines and if no resulted fragments just output "No related fragments"):

Figure 30: The prompt of *Aspect Identification* for the review aspect of *Weight*.

Opinion Consolidation

You are good at writing summaries for opinionated texts. You are given some opinionated text fragments, please write a concise summary for them.
Target input review fragments:

{{review_fragments}}

The final summary of these target input text fragments (just produce the answer without any other content):

Figure 31: The prompt of *Opinion Consolidation* for any individual review aspect for sports shoes.

Meta-Review Synthesis

You are good at understanding documents with sports shoes review opinions.
Below are product reviews in different aspects for a pair of shoes, please write a concise and natural meta-review which summaries the provided comments and covers all mentioned review aspects.

Comments on different aspects:

{{metas_generated}}

The meta-review is (directly output the answer without any other content):

Figure 32: The prompt of *Meta-Review Synthesis* for the product reviews of sports shoes.

reference meta-reviews from random meta-reviews (taken from other entities). Each annotator worked on 20 annotation instances for the main study and another 4 quality control instances. Raters were asked five questions about review aspects and opinion faithfulness. Our annotation instructions and interface are shown in Figure 41, Figure 42, and Figure 43. After filtering out annotators failing more than one quality control annotation pair, the annotators have reasonable agreement and the average Krippendorff's α of 0.335.

We follow the same setting for the evaluation of meta-reviews for hotels. There are also 540 annotations,

Naive Aspect-Aware Prompt

Please write a summary for the reviews on a scientific article, focused on the review aspects below.

Review aspects:

(1) Advancement: importance of the manuscript to discipline, significance of the contributions of the manuscript, and its potential impact to the field.

(2) Clarity: the readability of the writing (e.g., structure and language), reproducibility of details, and how accurately what the research question is, what was done and what was the conclusion are presented.

(3) Compliance: whether the manuscript fits the venue, and all ethical and publication requirements are met.

(4) Soundness: there are usually two types of soundness, empirical (how well experiments are designed and executed to support the claims, whether methods used are appropriate, and how correctly the data and results are reported, analysed, and interpreted.) and theoretical (whether arguments or claims in the manuscript are well supported by theoretical analysis, i.e., completeness, and the methodology, e.g., mathematical approach and the analysis is correct.)

(5) Novelty: how original the idea (e.g., tasks, datasets, or methods) is, and how clear where the problems and methods sit with respect to existing literature (i.e., meaningful comparison).

Reviews on a scientific article:
{{source_documents}}

The output summary:

Figure 33: The prompt with aspects in scientific reviews of research articles for *naive aspect-aware prompting*.

Automatic Decomposition Prompt

You are requested to write the steps. Please output the final answer with only the steps in different lines, no other useless content.

Please give me sequential steps to write a summary specific for the following reviews on an academic paper.
Reviews on a paper: {source_text}
The steps to write a summary in different lines:

Figure 34: The prompt for automatic decomposition to generate intermediate reasoning steps to write the meta-review for scientific reviews.

Prompt to Follow Reasoning Steps from Automatic Decomposition

You are requested to follow the instruction and only generate the requested output.

{output_from_last_step}
Please follow the instruction below and give your output.
{current_step}
The output:

Figure 35: The prompt to follow automatically predicted steps by *automatic decomposition* to generate the final meta-review.

and we obtain 27 annotators from Prolific. The annotation instructions and experimental interface are shown in Figure 44, Figure 45, and Figure 46. After filtering out annotators who failed on more than one quality control instances, the average Krippendorff's α is 0.622.

For scientific reviews of research articles, we randomly select ten entities from the test data of PeerSum. There are also six pairs of comparisons between our modular approach with Llama-3.1-70B as a backbone

Chunk Summarization Prompt

You are requested to do summarization. Please output the final answer with only the summary, no other useless content.

Please write a summary for the following review on an academic paper.
The review: {the_text_chunk}
The output summary:

Figure 36: The prompt of *chunk-wise decomposition* to summarize individual chunks of texts for scientific reviews of research articles.

Summary Aggregation Prompt

You are requested to do summarization. Please output the final answer with only the summary, no other useless content.

Please write a summary for the following texts.
The texts to be summarized:
{the_concatenation_of_small_meta_reviews_of_chunks}
The output summary:

Figure 37: The Prompt for aggregating chunk-specific meta-reviews into the global meta-review.

G-Eval for Sports Shoes

Here are several review documents that contain opinions from different people about a pair of shoes, along with a candidate summary of these reviews.

You are required to evaluate how accurately the given summary reflects the overall opinions for review aspects expressed in the original reviews.

Please read all opinions in the summary and calculate the percentage of faithful opinions that are clearly supported by the source review documents.

Review documents:

{{source_documents}}

The candidate summary:

{{generation_summary}}

The percentage of faithful opinions (only output a decimal like 0.12, no other content):

Figure 38: The G-Eval prompt for evaluating meta-reviews for sports shoes.

and comparison baselines. As there are only about 15 reviews on average, we show annotators all reviews. Therefore, there are 6 pairs of comparisons for each entity. Each pair gets annotated by three different annotators and we have 180 annotations for the dataset. We elicited 9 annotators from Prolific with required L1 English from the US or UK, and a minimum approval rate of 100% in more than 100 studies. We also required that they are pursuing a PhD in computer science or engineering. In addition to the attention check question for each annotation instance, we also included quality control instances, same as before. Therefore, each annotator worked on 20 pairs of comparisons for the main study and another 4 quality control instances. In each annotation, participants are asked 5 questions about review aspects and opinion faithfulness. The annotation instructions and interface are shown in Figure 47, Figure 48, and Figure 49. After filtering out annotators failing more than one quality control instances, the annotators, the average Krippendorff's α is 0.463.

G-Eval for Research Articles

Here are several review documents that contain opinions from different people about a scientific paper, along with a candidate summary of these reviews.

You are required to evaluate how accurately the given summary reflects the overall opinions for review aspects expressed in the original reviews.

Please read all opinions in the summary and calculate the percentage of faithful opinions that are clearly supported by the source review documents.

Review documents:

{{source_documents}}

The candidate summary:

{{generation_summary}}

The percentage of faithful opinions (only output a decimal like 0.12, no other content):

Figure 39: The G-Eval prompt for evaluating meta-reviews on research articles.

G-Eval for Hotels

Here are several review documents that contain opinions from different people about a hotel, along with a candidate summary of these reviews.

You are required to evaluate how accurately the given summary reflects the overall opinions for review aspects expressed in the original reviews.

Please read all opinions in the summary and calculate the percentage of faithful opinions that are clearly supported by the source review documents.

Review documents:

{{source_documents}}

The candidate summary:

{{generation_summary}}

The percentage of faithful opinions (only output a decimal like 0.12, no other content):

Figure 40: The G-Eval prompt for evaluating meta-reviews on hotels.

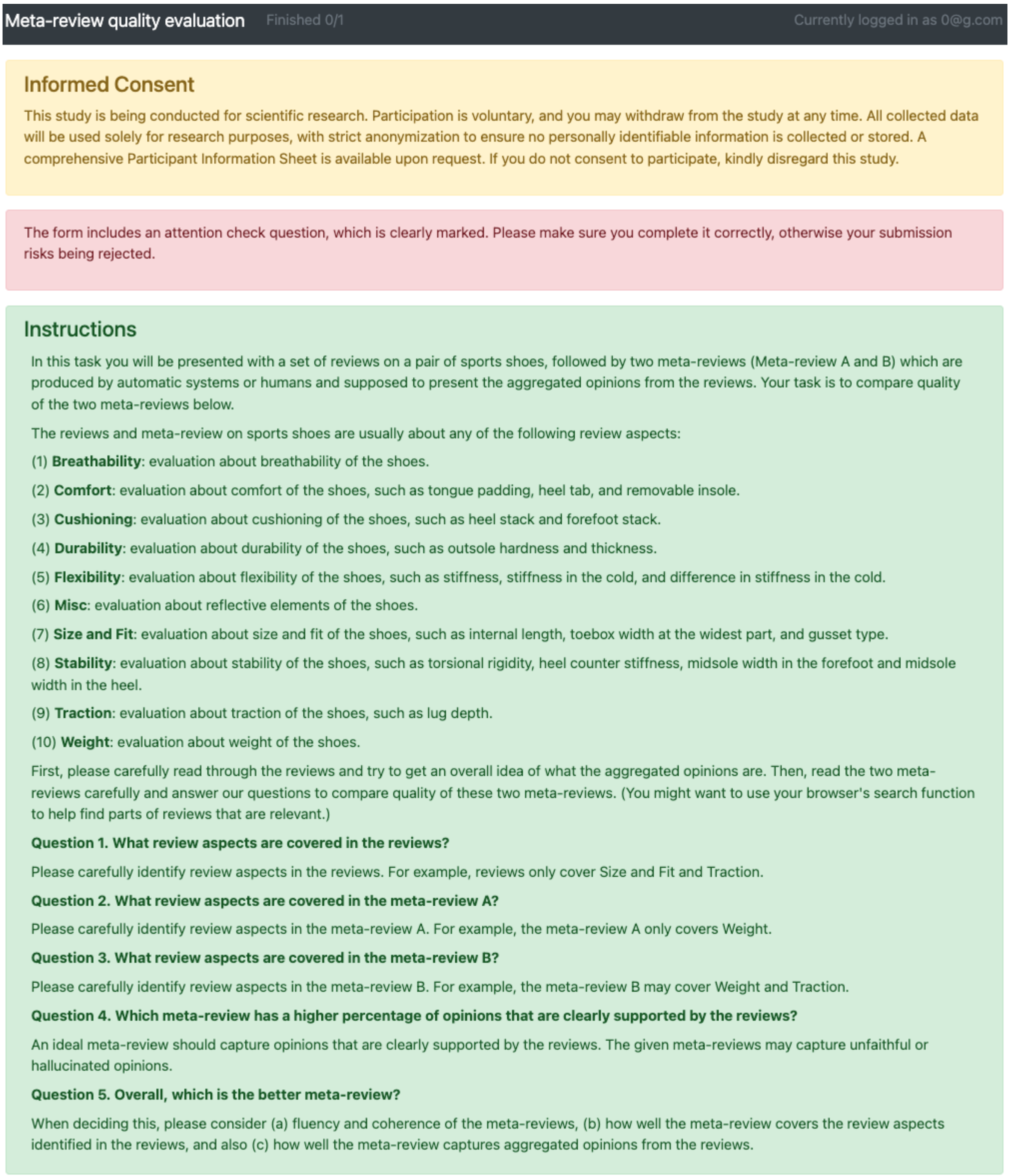
H Details of Human Evaluation on Usefulness of Intermediate Outputs

To record the time that humans spend to write meta-reviews with different reasoning steps, we conduct the experiments also with Prolific and present annotators interfaces with instructions in Figure 50, Figure 51 and Figure 52.We recruited five crowdworkers through Prolific[19] with compensation above the UK living wage at £12 per working hour. These annotators are required to be experienced in L1 English from the US or UK, with a minimum approval rate of 100% in more than 100 studies. Annotators are required to focus on the annotation task and finish the writing task in a continuous period of time. The study is conducted on ten entities and there are three meta-reviews for each (according to the three conditions described in Section 5). To avoid memorization, each annotator must write a meta-review for each entity only once. We find that all our annotators passed our attention check question present in our instructions Figure 52. We calculate the average time that the participants take for the ten instances in each condition from the

[19] www.prolific.com

five annotators.

To compare the quality of written meta-reviews in the three different conditions, we run another human evaluation in the same setting as the one to compare model-generated meta-reviews in Section 5. This was also based on pair-wise comparison and there were 30 pairs of comparison. We recruited three annotators and each pair of comparison was annotated for three times. The agreement among the three annotators is high (Krippendorff's α is 0.542).



Meta-review quality evaluation Finished 0/1 Currently logged in as 0@g.com

Informed Consent

This study is being conducted for scientific research. Participation is voluntary, and you may withdraw from the study at any time. All collected data will be used solely for research purposes, with strict anonymization to ensure no personally identifiable information is collected or stored. A comprehensive Participant Information Sheet is available upon request. If you do not consent to participate, kindly disregard this study.

The form includes an attention check question, which is clearly marked. Please make sure you complete it correctly, otherwise your submission risks being rejected.

Instructions

In this task you will be presented with a set of reviews on a pair of sports shoes, followed by two meta-reviews (Meta-review A and B) which are produced by automatic systems or humans and supposed to present the aggregated opinions from the reviews. Your task is to compare quality of the two meta-reviews below.

The reviews and meta-review on sports shoes are usually about any of the following review aspects:

(1) **Breathability**: evaluation about breathability of the shoes.

(2) **Comfort**: evaluation about comfort of the shoes, such as tongue padding, heel tab, and removable insole.

(3) **Cushioning**: evaluation about cushioning of the shoes, such as heel stack and forefoot stack.

(4) **Durability**: evaluation about durability of the shoes, such as outsole hardness and thickness.

(5) **Flexibility**: evaluation about flexibility of the shoes, such as stiffness, stiffness in the cold, and difference in stiffness in the cold.

(6) **Misc**: evaluation about reflective elements of the shoes.

(7) **Size and Fit**: evaluation about size and fit of the shoes, such as internal length, toebox width at the widest part, and gusset type.

(8) **Stability**: evaluation about stability of the shoes, such as torsional rigidity, heel counter stiffness, midsole width in the forefoot and midsole width in the heel.

(9) **Traction**: evaluation about traction of the shoes, such as lug depth.

(10) **Weight**: evaluation about weight of the shoes.

First, please carefully read through the reviews and try to get an overall idea of what the aggregated opinions are. Then, read the two meta-reviews carefully and answer our questions to compare quality of these two meta-reviews. (You might want to use your browser's search function to help find parts of reviews that are relevant.)

Question 1. What review aspects are covered in the reviews?

Please carefully identify review aspects in the reviews. For example, reviews only cover Size and Fit and Traction.

Question 2. What review aspects are covered in the meta-review A?

Please carefully identify review aspects in the meta-review A. For example, the meta-review A only covers Weight.

Question 3. What review aspects are covered in the meta-review B?

Please carefully identify review aspects in the meta-review B. For example, the meta-review B may cover Weight and Traction.

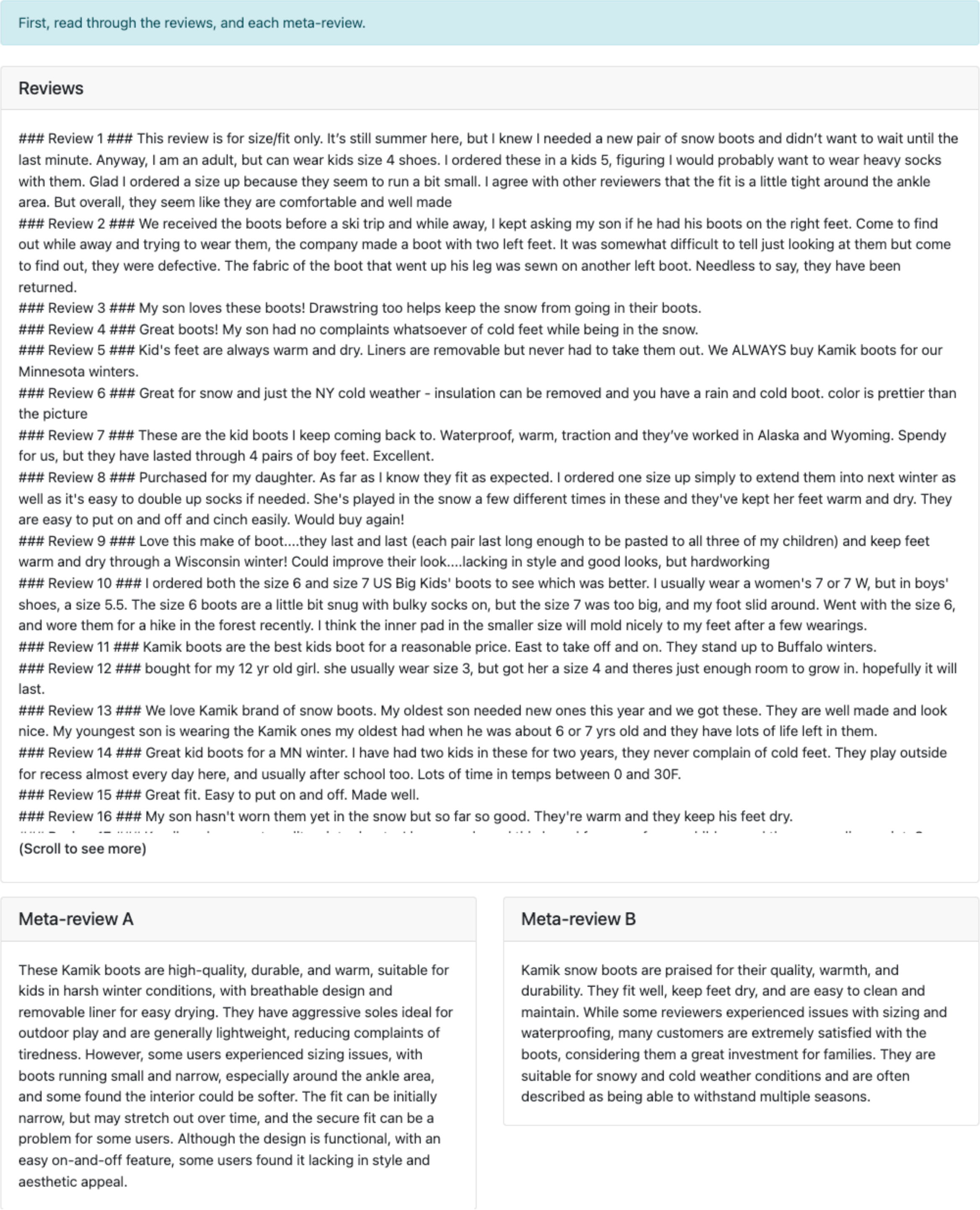
Question 4. Which meta-review has a higher percentage of opinions that are clearly supported by the reviews?

An ideal meta-review should capture opinions that are clearly supported by the reviews. The given meta-reviews may capture unfaithful or hallucinated opinions.

Question 5. Overall, which is the better meta-review?

When deciding this, please consider (a) fluency and coherence of the meta-reviews, (b) how well the meta-review covers the review aspects identified in the reviews, and also (c) how well the meta-review captures aggregated opinions from the reviews.

Figure 41: Experimental instructions and interface for human evaluation study on sports shoes reviews (part 1).



First, read through the reviews, and each meta-review.

Reviews

Review 1 ### This review is for size/fit only. It's still summer here, but I knew I needed a new pair of snow boots and didn't want to wait until the last minute. Anyway, I am an adult, but can wear kids size 4 shoes. I ordered these in a kids 5, figuring I would probably want to wear heavy socks with them. Glad I ordered a size up because they seem to run a bit small. I agree with other reviewers that the fit is a little tight around the ankle area. But overall, they seem like they are comfortable and well made
Review 2 ### We received the boots before a ski trip and while away, I kept asking my son if he had his boots on the right feet. Come to find out while away and trying to wear them, the company made a boot with two left feet. It was somewhat difficult to tell just looking at them but come to find out, they were defective. The fabric of the boot that went up his leg was sewn on another left boot. Needless to say, they have been returned.
Review 3 ### My son loves these boots! Drawstring too helps keep the snow from going in their boots.
Review 4 ### Great boots! My son had no complaints whatsoever of cold feet while being in the snow.
Review 5 ### Kid's feet are always warm and dry. Liners are removable but never had to take them out. We ALWAYS buy Kamik boots for our Minnesota winters.
Review 6 ### Great for snow and just the NY cold weather - insulation can be removed and you have a rain and cold boot. color is prettier than the picture
Review 7 ### These are the kid boots I keep coming back to. Waterproof, warm, traction and they've worked in Alaska and Wyoming. Spendy for us, but they have lasted through 4 pairs of boy feet. Excellent.
Review 8 ### Purchased for my daughter. As far as I know they fit as expected. I ordered one size up simply to extend them into next winter as well as it's easy to double up socks if needed. She's played in the snow a few different times in these and they've kept her feet warm and dry. They are easy to put on and off and cinch easily. Would buy again!
Review 9 ### Love this make of boot....they last and last (each pair last long enough to be pasted to all three of my children) and keep feet warm and dry through a Wisconsin winter! Could improve their look....lacking in style and good looks, but hardworking
Review 10 ### I ordered both the size 6 and size 7 US Big Kids' boots to see which was better. I usually wear a women's 7 or 7 W, but in boys' shoes, a size 5.5. The size 6 boots are a little bit snug with bulky socks on, but the size 7 was too big, and my foot slid around. Went with the size 6, and wore them for a hike in the forest recently. I think the inner pad in the smaller size will mold nicely to my feet after a few wearings.
Review 11 ### Kamik boots are the best kids boot for a reasonable price. East to take off and on. They stand up to Buffalo winters.
Review 12 ### bought for my 12 yr old girl. she usually wear size 3, but got her a size 4 and theres just enough room to grow in. hopefully it will last.
Review 13 ### We love Kamik brand of snow boots. My oldest son needed new ones this year and we got these. They are well made and look nice. My youngest son is wearing the Kamik ones my oldest had when he was about 6 or 7 yrs old and they have lots of life left in them.
Review 14 ### Great kid boots for a MN winter. I have had two kids in these for two years, they never complain of cold feet. They play outside for recess almost every day here, and usually after school too. Lots of time in temps between 0 and 30F.
Review 15 ### Great fit. Easy to put on and off. Made well.
Review 16 ### My son hasn't worn them yet in the snow but so far so good. They're warm and they keep his feet dry.

(Scroll to see more)

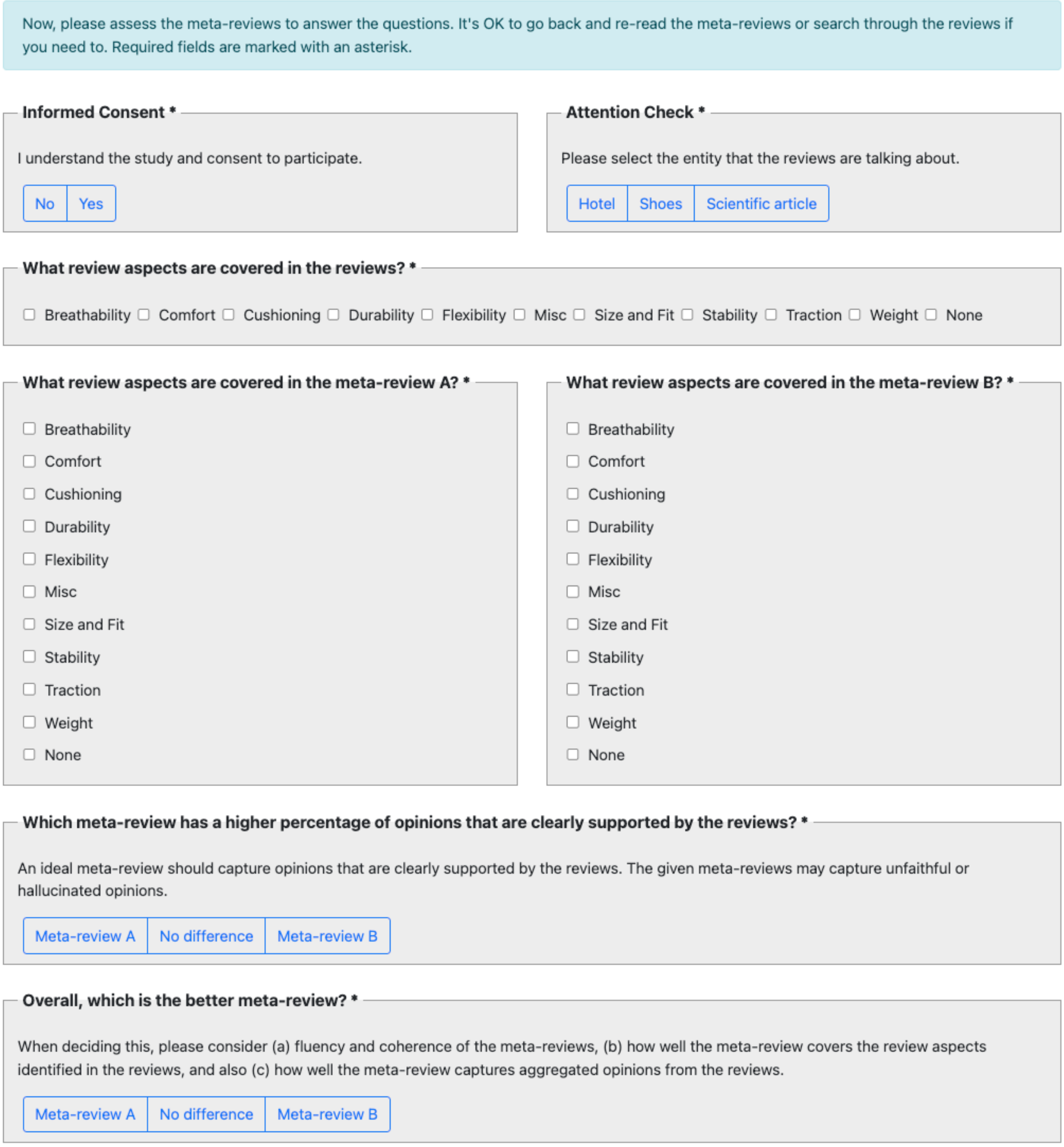
Meta-review A

These Kamik boots are high-quality, durable, and warm, suitable for kids in harsh winter conditions, with breathable design and removable liner for easy drying. They have aggressive soles ideal for outdoor play and are generally lightweight, reducing complaints of tiredness. However, some users experienced sizing issues, with boots running small and narrow, especially around the ankle area, and some found the interior could be softer. The fit can be initially narrow, but may stretch out over time, and the secure fit can be a problem for some users. Although the design is functional, with an easy on-and-off feature, some users found it lacking in style and aesthetic appeal.

Meta-review B

Kamik snow boots are praised for their quality, warmth, and durability. They fit well, keep feet dry, and are easy to clean and maintain. While some reviewers experienced issues with sizing and waterproofing, many customers are extremely satisfied with the boots, considering them a great investment for families. They are suitable for snowy and cold weather conditions and are often described as being able to withstand multiple seasons.

Figure 42: Experimental instructions and interface for human evaluation study on sports shoes (part 2).



Now, please assess the meta-reviews to answer the questions. It's OK to go back and re-read the meta-reviews or search through the reviews if you need to. Required fields are marked with an asterisk.

Informed Consent *

I understand the study and consent to participate.

Attention Check *

Please select the entity that the reviews are talking about.

What review aspects are covered in the reviews? *

□ Breathability □ Comfort □ Cushioning □ Durability □ Flexibility □ Misc □ Size and Fit □ Stability □ Traction □ Weight □ None

What review aspects are covered in the meta-review A? *

□ Breathability
□ Comfort
□ Cushioning
□ Durability
□ Flexibility
□ Misc
□ Size and Fit
□ Stability
□ Traction
□ Weight
□ None

What review aspects are covered in the meta-review B? *

□ Breathability
□ Comfort
□ Cushioning
□ Durability
□ Flexibility
□ Misc
□ Size and Fit
□ Stability
□ Traction
□ Weight
□ None

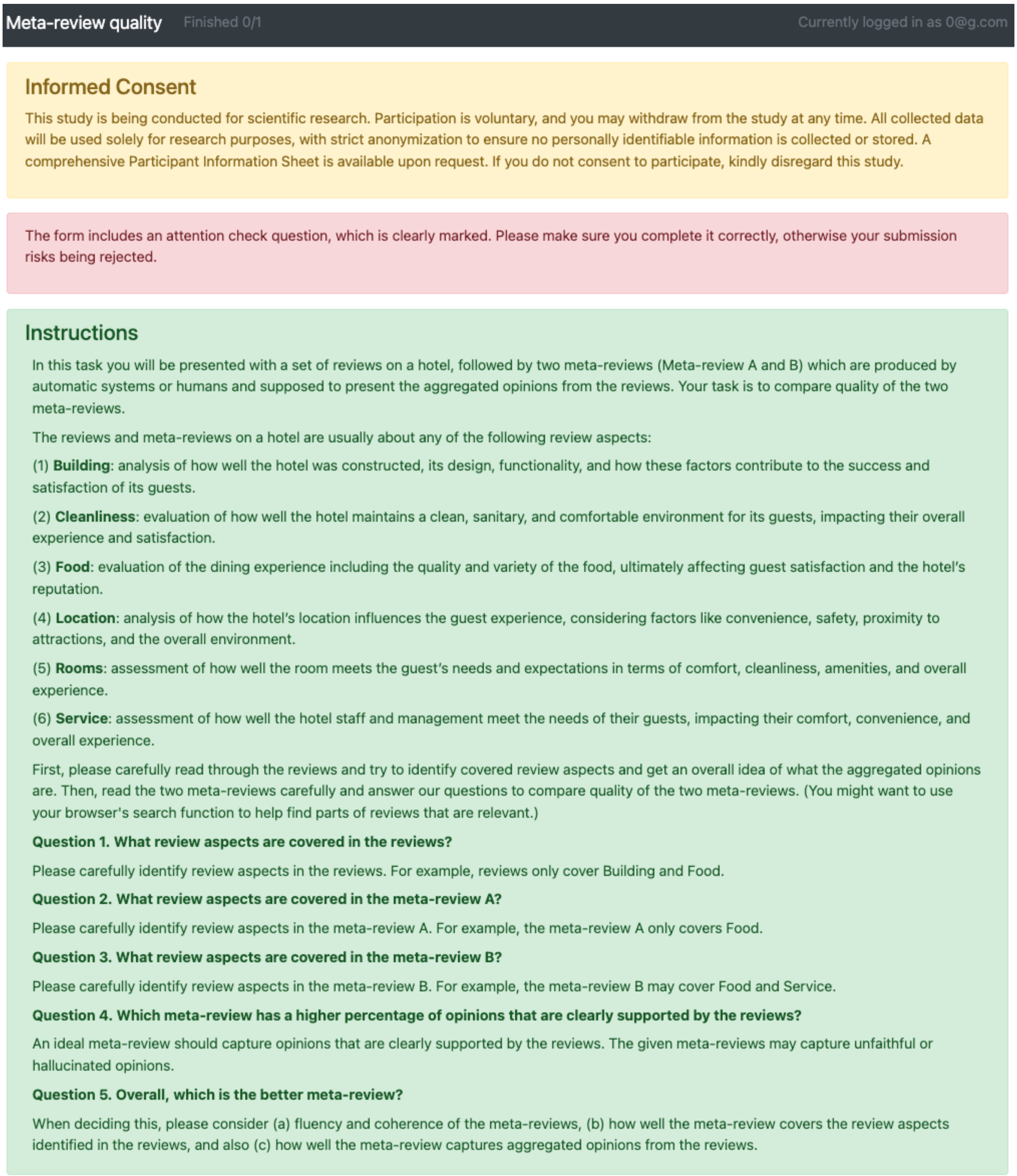
Which meta-review has a higher percentage of opinions that are clearly supported by the reviews? *

An ideal meta-review should capture opinions that are clearly supported by the reviews. The given meta-reviews may capture unfaithful or hallucinated opinions.

Overall, which is the better meta-review? *

When deciding this, please consider (a) fluency and coherence of the meta-reviews, (b) how well the meta-review covers the review aspects identified in the reviews, and also (c) how well the meta-review captures aggregated opinions from the reviews.

Figure 43: Experimental instructions and interface for human evaluation study on sports shoes (part 3).



Meta-review quality Finished 0/1 Currently logged in as 0@g.com

Informed Consent

This study is being conducted for scientific research. Participation is voluntary, and you may withdraw from the study at any time. All collected data will be used solely for research purposes, with strict anonymization to ensure no personally identifiable information is collected or stored. A comprehensive Participant Information Sheet is available upon request. If you do not consent to participate, kindly disregard this study.

The form includes an attention check question, which is clearly marked. Please make sure you complete it correctly, otherwise your submission risks being rejected.

Instructions

In this task you will be presented with a set of reviews on a hotel, followed by two meta-reviews (Meta-review A and B) which are produced by automatic systems or humans and supposed to present the aggregated opinions from the reviews. Your task is to compare quality of the two meta-reviews.

The reviews and meta-reviews on a hotel are usually about any of the following review aspects:

(1) **Building**: analysis of how well the hotel was constructed, its design, functionality, and how these factors contribute to the success and satisfaction of its guests.

(2) **Cleanliness**: evaluation of how well the hotel maintains a clean, sanitary, and comfortable environment for its guests, impacting their overall experience and satisfaction.

(3) **Food**: evaluation of the dining experience including the quality and variety of the food, ultimately affecting guest satisfaction and the hotel's reputation.

(4) **Location**: analysis of how the hotel's location influences the guest experience, considering factors like convenience, safety, proximity to attractions, and the overall environment.

(5) **Rooms**: assessment of how well the room meets the guest's needs and expectations in terms of comfort, cleanliness, amenities, and overall experience.

(6) **Service**: assessment of how well the hotel staff and management meet the needs of their guests, impacting their comfort, convenience, and overall experience.

First, please carefully read through the reviews and try to identify covered review aspects and get an overall idea of what the aggregated opinions are. Then, read the two meta-reviews carefully and answer our questions to compare quality of the two meta-reviews. (You might want to use your browser's search function to help find parts of reviews that are relevant.)

Question 1. What review aspects are covered in the reviews?

Please carefully identify review aspects in the reviews. For example, reviews only cover Building and Food.

Question 2. What review aspects are covered in the meta-review A?

Please carefully identify review aspects in the meta-review A. For example, the meta-review A only covers Food.

Question 3. What review aspects are covered in the meta-review B?

Please carefully identify review aspects in the meta-review B. For example, the meta-review B may cover Food and Service.

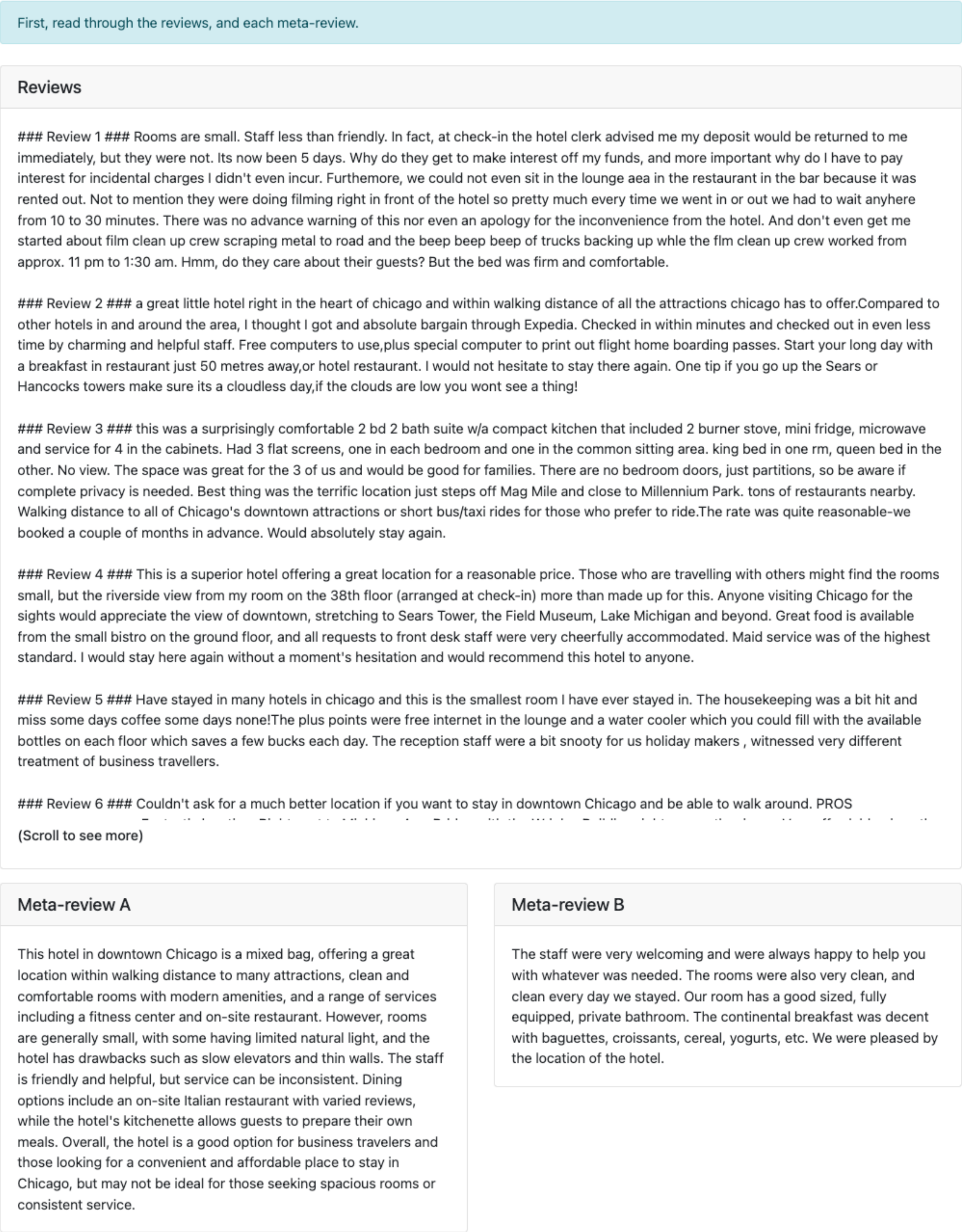
Question 4. Which meta-review has a higher percentage of opinions that are clearly supported by the reviews?

An ideal meta-review should capture opinions that are clearly supported by the reviews. The given meta-reviews may capture unfaithful or hallucinated opinions.

Question 5. Overall, which is the better meta-review?

When deciding this, please consider (a) fluency and coherence of the meta-reviews, (b) how well the meta-review covers the review aspects identified in the reviews, and also (c) how well the meta-review captures aggregated opinions from the reviews.

Figure 44: Experimental instructions and interface for human evaluation study on hotels (part 1).



First, read through the reviews, and each meta-review.

Reviews

Review 1 ### Rooms are small. Staff less than friendly. In fact, at check-in the hotel clerk advised me my deposit would be returned to me immediately, but they were not. Its now been 5 days. Why do they get to make interest off my funds, and more important why do I have to pay interest for incidental charges I didn't even incur. Furthemore, we could not even sit in the lounge aea in the restaurant in the bar because it was rented out. Not to mention they were doing filming right in front of the hotel so pretty much every time we went in or out we had to wait anyhere from 10 to 30 minutes. There was no advance warning of this nor even an apology for the inconvenience from the hotel. And don't even get me started about film clean up crew scraping metal to road and the beep beep beep of trucks backing up whle the flm clean up crew worked from approx. 11 pm to 1:30 am. Hmm, do they care about their guests? But the bed was firm and comfortable.

Review 2 ### a great little hotel right in the heart of chicago and within walking distance of all the attractions chicago has to offer.Compared to other hotels in and around the area, I thought I got and absolute bargain through Expedia. Checked in within minutes and checked out in even less time by charming and helpful staff. Free computers to use,plus special computer to print out flight home boarding passes. Start your long day with a breakfast in restaurant just 50 metres away,or hotel restaurant. I would not hesitate to stay there again. One tip if you go up the Sears or Hancocks towers make sure its a cloudless day,if the clouds are low you wont see a thing!

Review 3 ### this was a surprisingly comfortable 2 bd 2 bath suite w/a compact kitchen that included 2 burner stove, mini fridge, microwave and service for 4 in the cabinets. Had 3 flat screens, one in each bedroom and one in the common sitting area. king bed in one rm, queen bed in the other. No view. The space was great for the 3 of us and would be good for families. There are no bedroom doors, just partitions, so be aware if complete privacy is needed. Best thing was the terrific location just steps off Mag Mile and close to Millennium Park. tons of restaurants nearby. Walking distance to all of Chicago's downtown attractions or short bus/taxi rides for those who prefer to ride.The rate was quite reasonable-we booked a couple of months in advance. Would absolutely stay again.

Review 4 ### This is a superior hotel offering a great location for a reasonable price. Those who are travelling with others might find the rooms small, but the riverside view from my room on the 38th floor (arranged at check-in) more than made up for this. Anyone visiting Chicago for the sights would appreciate the view of downtown, stretching to Sears Tower, the Field Museum, Lake Michigan and beyond. Great food is available from the small bistro on the ground floor, and all requests to front desk staff were very cheerfully accommodated. Maid service was of the highest standard. I would stay here again without a moment's hesitation and would recommend this hotel to anyone.

Review 5 ### Have stayed in many hotels in chicago and this is the smallest room I have ever stayed in. The housekeeping was a bit hit and miss some days coffee some days none!The plus points were free internet in the lounge and a water cooler which you could fill with the available bottles on each floor which saves a few bucks each day. The reception staff were a bit snooty for us holiday makers , witnessed very different treatment of business travellers.

Review 6 ### Couldn't ask for a much better location if you want to stay in downtown Chicago and be able to walk around. PROS

(Scroll to see more)

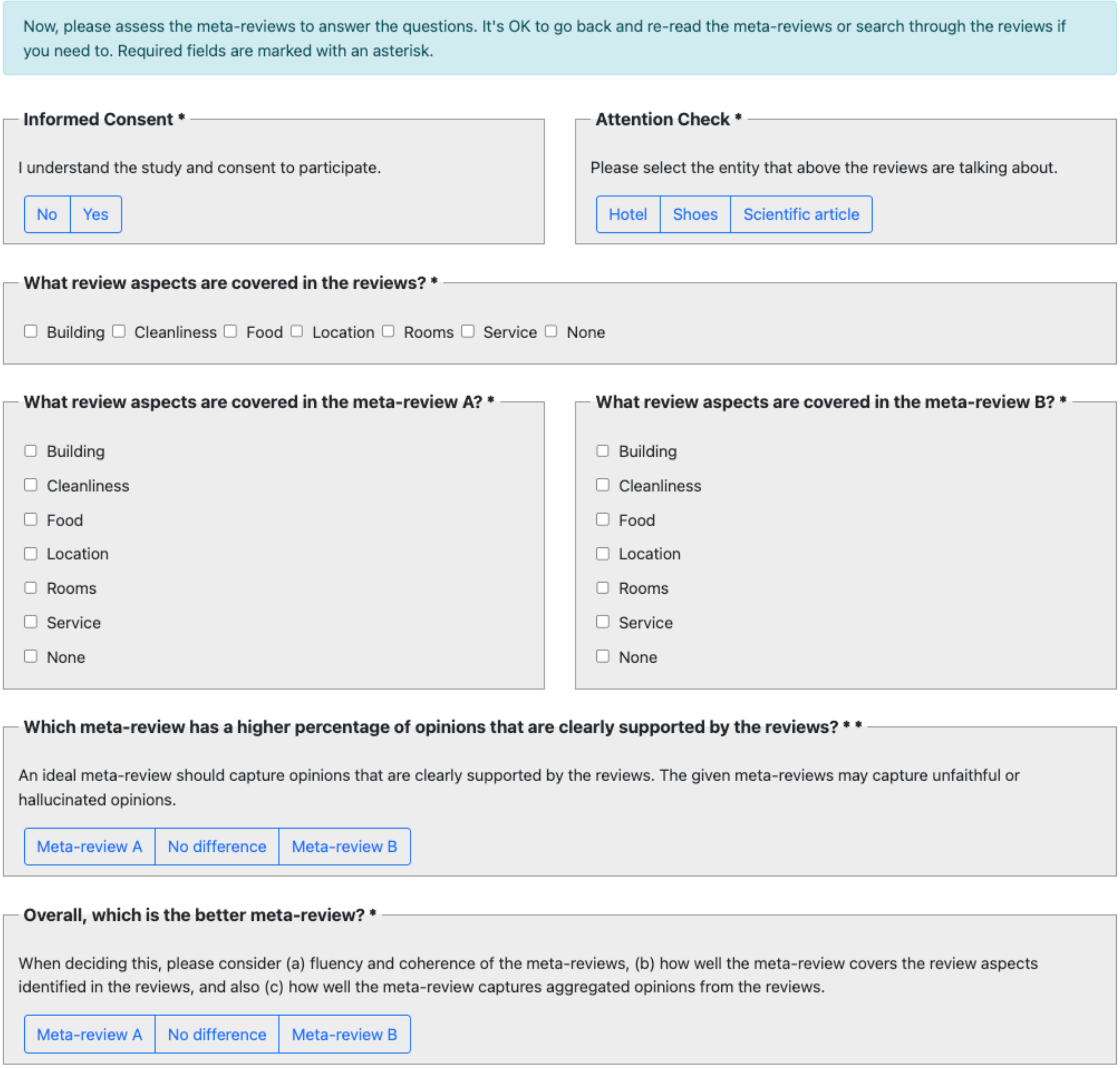
Meta-review A

This hotel in downtown Chicago is a mixed bag, offering a great location within walking distance to many attractions, clean and comfortable rooms with modern amenities, and a range of services including a fitness center and on-site restaurant. However, rooms are generally small, with some having limited natural light, and the hotel has drawbacks such as slow elevators and thin walls. The staff is friendly and helpful, but service can be inconsistent. Dining options include an on-site Italian restaurant with varied reviews, while the hotel's kitchenette allows guests to prepare their own meals. Overall, the hotel is a good option for business travelers and those looking for a convenient and affordable place to stay in Chicago, but may not be ideal for those seeking spacious rooms or consistent service.

Meta-review B

The staff were very welcoming and were always happy to help you with whatever was needed. The rooms were also very clean, and clean every day we stayed. Our room has a good sized, fully equipped, private bathroom. The continental breakfast was decent with baguettes, croissants, cereal, yogurts, etc. We were pleased by the location of the hotel.

Figure 45: Experimental instructions and interface for human evaluation study on hotels (part 2).



Now, please assess the meta-reviews to answer the questions. It's OK to go back and re-read the meta-reviews or search through the reviews if you need to. Required fields are marked with an asterisk.

Informed Consent *

I understand the study and consent to participate.

No | Yes

Attention Check *

Please select the entity that above the reviews are talking about.

Hotel | Shoes | Scientific article

What review aspects are covered in the reviews? *

□ Building □ Cleanliness □ Food □ Location □ Rooms □ Service □ None

What review aspects are covered in the meta-review A? *

□ Building
□ Cleanliness
□ Food
□ Location
□ Rooms
□ Service
□ None

What review aspects are covered in the meta-review B? *

□ Building
□ Cleanliness
□ Food
□ Location
□ Rooms
□ Service
□ None

Which meta-review has a higher percentage of opinions that are clearly supported by the reviews? * *

An ideal meta-review should capture opinions that are clearly supported by the reviews. The given meta-reviews may capture unfaithful or hallucinated opinions.

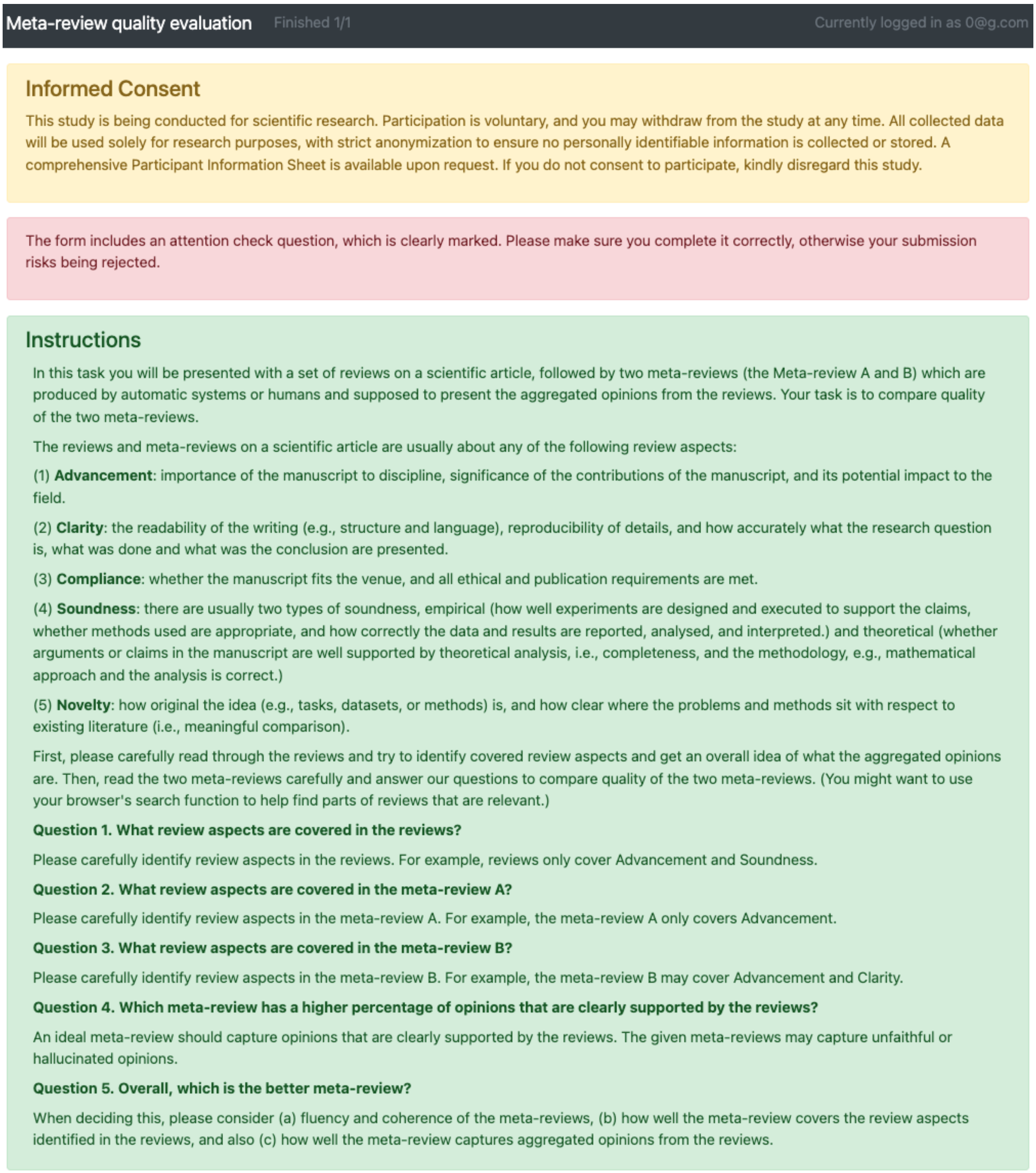
Meta-review A | No difference | Meta-review B

Overall, which is the better meta-review? *

When deciding this, please consider (a) fluency and coherence of the meta-reviews, (b) how well the meta-review covers the review aspects identified in the reviews, and also (c) how well the meta-review captures aggregated opinions from the reviews.

Meta-review A | No difference | Meta-review B

Figure 46: Experimental instructions and interface for human evaluation study on hotels (part 3).



Meta-review quality evaluation
Finished 1/1
Currently logged in as 0@g.com

Informed Consent

This study is being conducted for scientific research. Participation is voluntary, and you may withdraw from the study at any time. All collected data will be used solely for research purposes, with strict anonymization to ensure no personally identifiable information is collected or stored. A comprehensive Participant Information Sheet is available upon request. If you do not consent to participate, kindly disregard this study.

The form includes an attention check question, which is clearly marked. Please make sure you complete it correctly, otherwise your submission risks being rejected.

Instructions

In this task you will be presented with a set of reviews on a scientific article, followed by two meta-reviews (the Meta-review A and B) which are produced by automatic systems or humans and supposed to present the aggregated opinions from the reviews. Your task is to compare quality of the two meta-reviews.

The reviews and meta-reviews on a scientific article are usually about any of the following review aspects:

(1) **Advancement**: importance of the manuscript to discipline, significance of the contributions of the manuscript, and its potential impact to the field.

(2) **Clarity**: the readability of the writing (e.g., structure and language), reproducibility of details, and how accurately what the research question is, what was done and what was the conclusion are presented.

(3) **Compliance**: whether the manuscript fits the venue, and all ethical and publication requirements are met.

(4) **Soundness**: there are usually two types of soundness, empirical (how well experiments are designed and executed to support the claims, whether methods used are appropriate, and how correctly the data and results are reported, analysed, and interpreted.) and theoretical (whether arguments or claims in the manuscript are well supported by theoretical analysis, i.e., completeness, and the methodology, e.g., mathematical approach and the analysis is correct.)

(5) **Novelty**: how original the idea (e.g., tasks, datasets, or methods) is, and how clear where the problems and methods sit with respect to existing literature (i.e., meaningful comparison).

First, please carefully read through the reviews and try to identify covered review aspects and get an overall idea of what the aggregated opinions are. Then, read the two meta-reviews carefully and answer our questions to compare quality of the two meta-reviews. (You might want to use your browser's search function to help find parts of reviews that are relevant.)

Question 1. What review aspects are covered in the reviews?

Please carefully identify review aspects in the reviews. For example, reviews only cover Advancement and Soundness.

Question 2. What review aspects are covered in the meta-review A?

Please carefully identify review aspects in the meta-review A. For example, the meta-review A only covers Advancement.

Question 3. What review aspects are covered in the meta-review B?

Please carefully identify review aspects in the meta-review B. For example, the meta-review B may cover Advancement and Clarity.

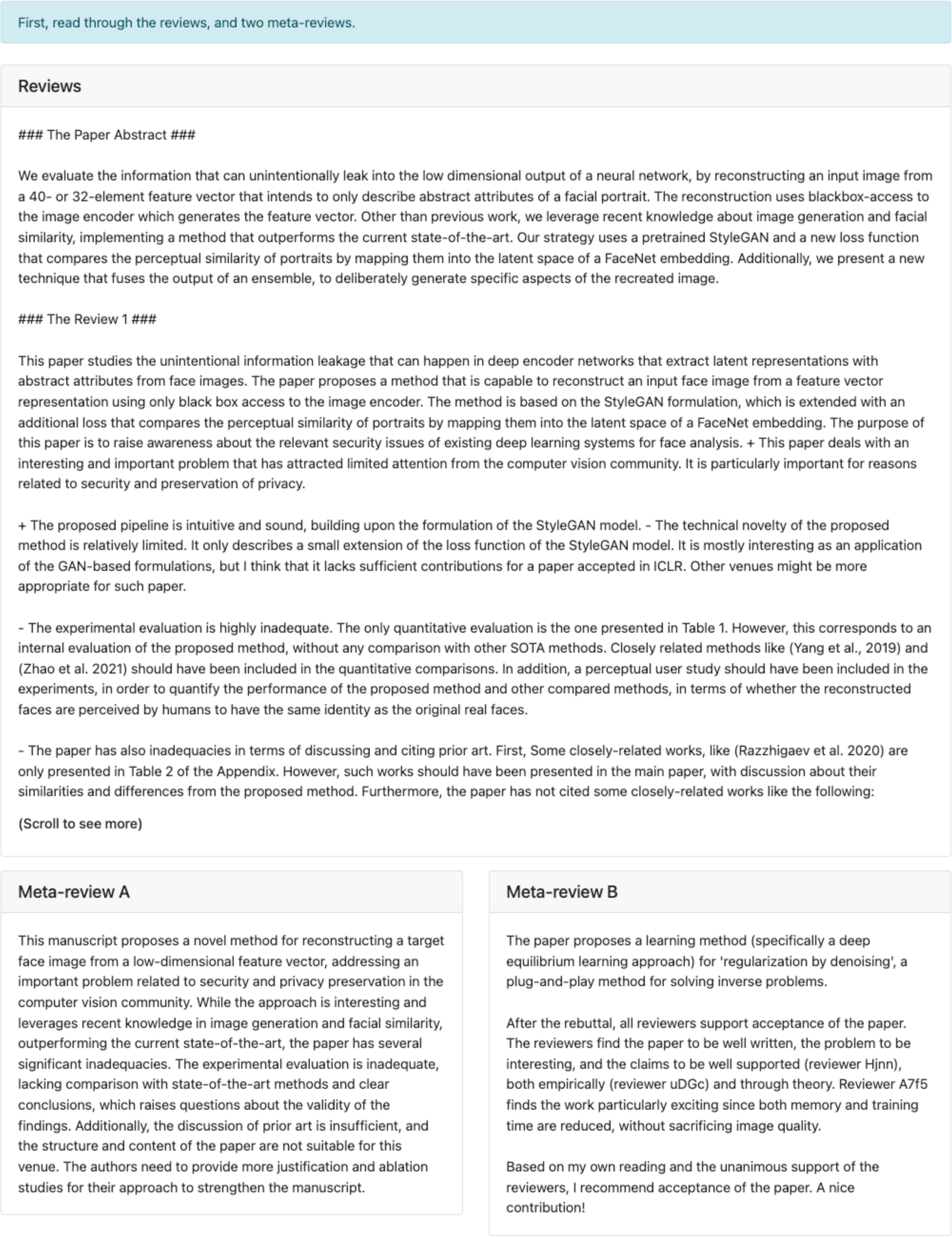
Question 4. Which meta-review has a higher percentage of opinions that are clearly supported by the reviews?

An ideal meta-review should capture opinions that are clearly supported by the reviews. The given meta-reviews may capture unfaithful or hallucinated opinions.

Question 5. Overall, which is the better meta-review?

When deciding this, please consider (a) fluency and coherence of the meta-reviews, (b) how well the meta-review covers the review aspects identified in the reviews, and also (c) how well the meta-review captures aggregated opinions from the reviews.

Figure 47: Experimental instructions and interface for human evaluation study on article reviews (part 1).



First, read through the reviews, and two meta-reviews.

Reviews

The Paper Abstract

We evaluate the information that can unintentionally leak into the low dimensional output of a neural network, by reconstructing an input image from a 40- or 32-element feature vector that intends to only describe abstract attributes of a facial portrait. The reconstruction uses blackbox-access to the image encoder which generates the feature vector. Other than previous work, we leverage recent knowledge about image generation and facial similarity, implementing a method that outperforms the current state-of-the-art. Our strategy uses a pretrained StyleGAN and a new loss function that compares the perceptual similarity of portraits by mapping them into the latent space of a FaceNet embedding. Additionally, we present a new technique that fuses the output of an ensemble, to deliberately generate specific aspects of the recreated image.

The Review 1

This paper studies the unintentional information leakage that can happen in deep encoder networks that extract latent representations with abstract attributes from face images. The paper proposes a method that is capable to reconstruct an input face image from a feature vector representation using only black box access to the image encoder. The method is based on the StyleGAN formulation, which is extended with an additional loss that compares the perceptual similarity of portraits by mapping them into the latent space of a FaceNet embedding. The purpose of this paper is to raise awareness about the relevant security issues of existing deep learning systems for face analysis. + This paper deals with an interesting and important problem that has attracted limited attention from the computer vision community. It is particularly important for reasons related to security and preservation of privacy.

+ The proposed pipeline is intuitive and sound, building upon the formulation of the StyleGAN model. - The technical novelty of the proposed method is relatively limited. It only describes a small extension of the loss function of the StyleGAN model. It is mostly interesting as an application of the GAN-based formulations, but I think that it lacks sufficient contributions for a paper accepted in ICLR. Other venues might be more appropriate for such paper.

- The experimental evaluation is highly inadequate. The only quantitative evaluation is the one presented in Table 1. However, this corresponds to an internal evaluation of the proposed method, without any comparison with other SOTA methods. Closely related methods like (Yang et al., 2019) and (Zhao et al. 2021) should have been included in the quantitative comparisons. In addition, a perceptual user study should have been included in the experiments, in order to quantify the performance of the proposed method and other compared methods, in terms of whether the reconstructed faces are perceived by humans to have the same identity as the original real faces.

- The paper has also inadequacies in terms of discussing and citing prior art. First, Some closely-related works, like (Razzhigaev et al. 2020) are only presented in Table 2 of the Appendix. However, such works should have been presented in the main paper, with discussion about their similarities and differences from the proposed method. Furthermore, the paper has not cited some closely-related works like the following:

(Scroll to see more)

Meta-review A

This manuscript proposes a novel method for reconstructing a target face image from a low-dimensional feature vector, addressing an important problem related to security and privacy preservation in the computer vision community. While the approach is interesting and leverages recent knowledge in image generation and facial similarity, outperforming the current state-of-the-art, the paper has several significant inadequacies. The experimental evaluation is inadequate, lacking comparison with state-of-the-art methods and clear conclusions, which raises questions about the validity of the findings. Additionally, the discussion of prior art is insufficient, and the structure and content of the paper are not suitable for this venue. The authors need to provide more justification and ablation studies for their approach to strengthen the manuscript.

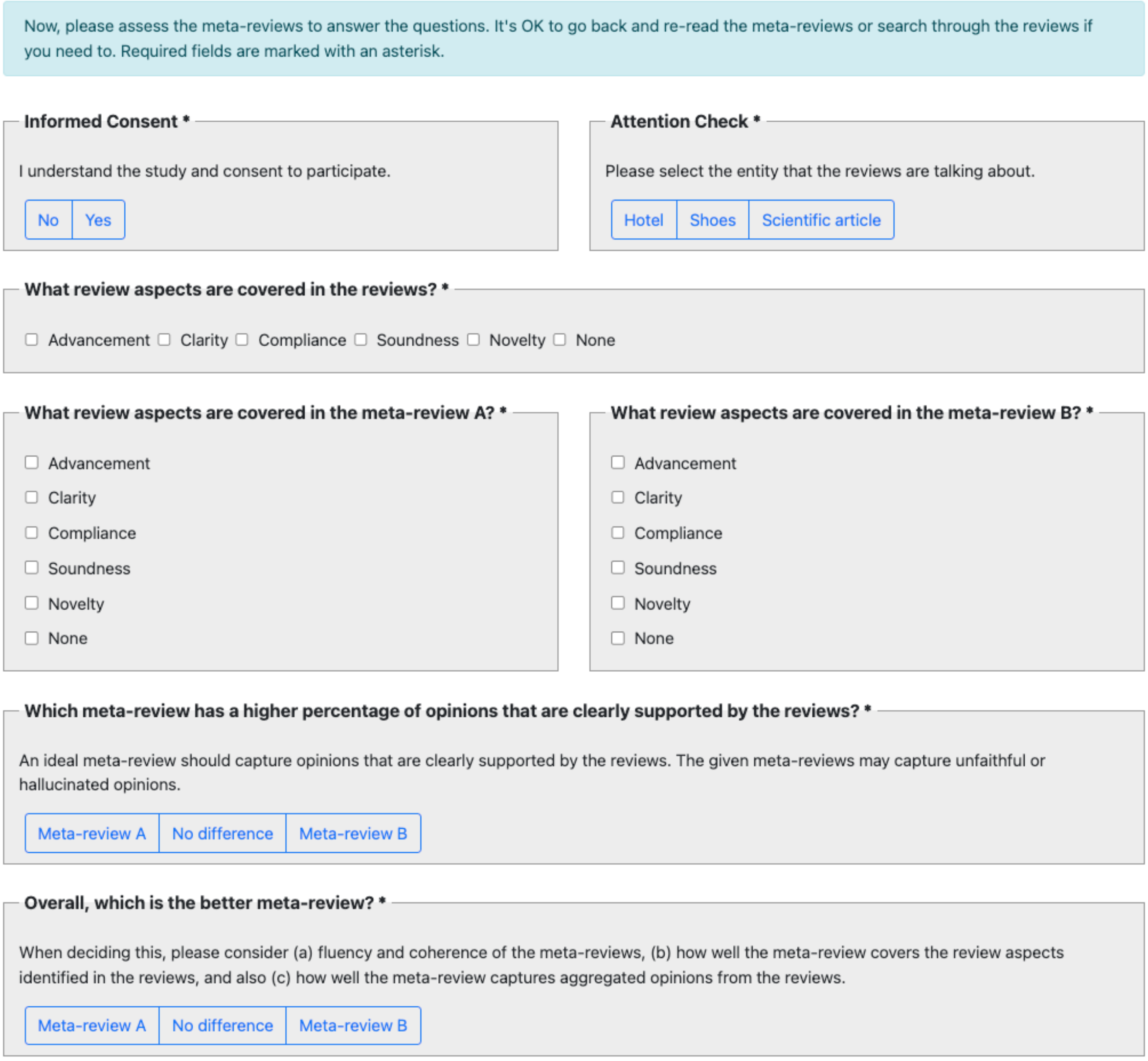
Meta-review B

The paper proposes a learning method (specifically a deep equilibrium learning approach) for 'regularization by denoising', a plug-and-play method for solving inverse problems.

After the rebuttal, all reviewers support acceptance of the paper. The reviewers find the paper to be well written, the problem to be interesting, and the claims to be well supported (reviewer Hjnn), both empirically (reviewer uDGc) and through theory. Reviewer A7f5 finds the work particularly exciting since both memory and training time are reduced, without sacrificing image quality.

Based on my own reading and the unanimous support of the reviewers, I recommend acceptance of the paper. A nice contribution!

Figure 48: Experimental instructions and interface for human evaluation study on article reviews (part 2).



Now, please assess the meta-reviews to answer the questions. It's OK to go back and re-read the meta-reviews or search through the reviews if you need to. Required fields are marked with an asterisk.

Informed Consent *

I understand the study and consent to participate.

No | Yes

Attention Check *

Please select the entity that the reviews are talking about.

Hotel | Shoes | Scientific article

What review aspects are covered in the reviews? *

□ Advancement □ Clarity □ Compliance □ Soundness □ Novelty □ None

What review aspects are covered in the meta-review A? *

□ Advancement
□ Clarity
□ Compliance
□ Soundness
□ Novelty
□ None

What review aspects are covered in the meta-review B? *

□ Advancement
□ Clarity
□ Compliance
□ Soundness
□ Novelty
□ None

Which meta-review has a higher percentage of opinions that are clearly supported by the reviews? *

An ideal meta-review should capture opinions that are clearly supported by the reviews. The given meta-reviews may capture unfaithful or hallucinated opinions.

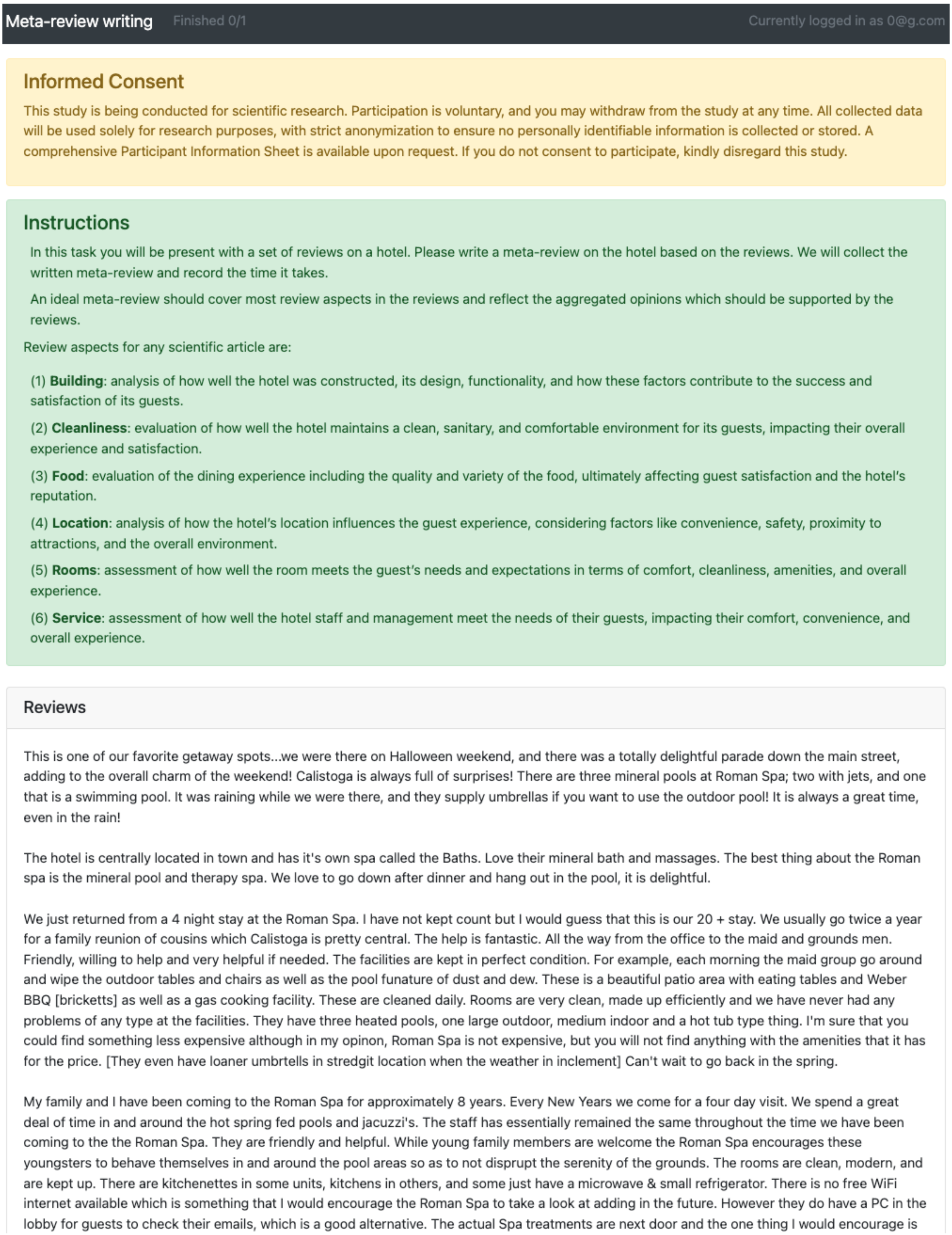
Meta-review A | No difference | Meta-review B

Overall, which is the better meta-review? *

When deciding this, please consider (a) fluency and coherence of the meta-reviews, (b) how well the meta-review covers the review aspects identified in the reviews, and also (c) how well the meta-review captures aggregated opinions from the reviews.

Meta-review A | No difference | Meta-review B

Figure 49: Experimental instructions and interface for human evaluation study on article reviews (part 3).



Informed Consent

This study is being conducted for scientific research. Participation is voluntary, and you may withdraw from the study at any time. All collected data will be used solely for research purposes, with strict anonymization to ensure no personally identifiable information is collected or stored. A comprehensive Participant Information Sheet is available upon request. If you do not consent to participate, kindly disregard this study.

Instructions

In this task you will be present with a set of reviews on a hotel. Please write a meta-review on the hotel based on the reviews. We will collect the written meta-review and record the time it takes.

An ideal meta-review should cover most review aspects in the reviews and reflect the aggregated opinions which should be supported by the reviews.

Review aspects for any scientific article are:

(1) **Building**: analysis of how well the hotel was constructed, its design, functionality, and how these factors contribute to the success and satisfaction of its guests.

(2) **Cleanliness**: evaluation of how well the hotel maintains a clean, sanitary, and comfortable environment for its guests, impacting their overall experience and satisfaction.

(3) **Food**: evaluation of the dining experience including the quality and variety of the food, ultimately affecting guest satisfaction and the hotel's reputation.

(4) **Location**: analysis of how the hotel's location influences the guest experience, considering factors like convenience, safety, proximity to attractions, and the overall environment.

(5) **Rooms**: assessment of how well the room meets the guest's needs and expectations in terms of comfort, cleanliness, amenities, and overall experience.

(6) **Service**: assessment of how well the hotel staff and management meet the needs of their guests, impacting their comfort, convenience, and overall experience.

Reviews

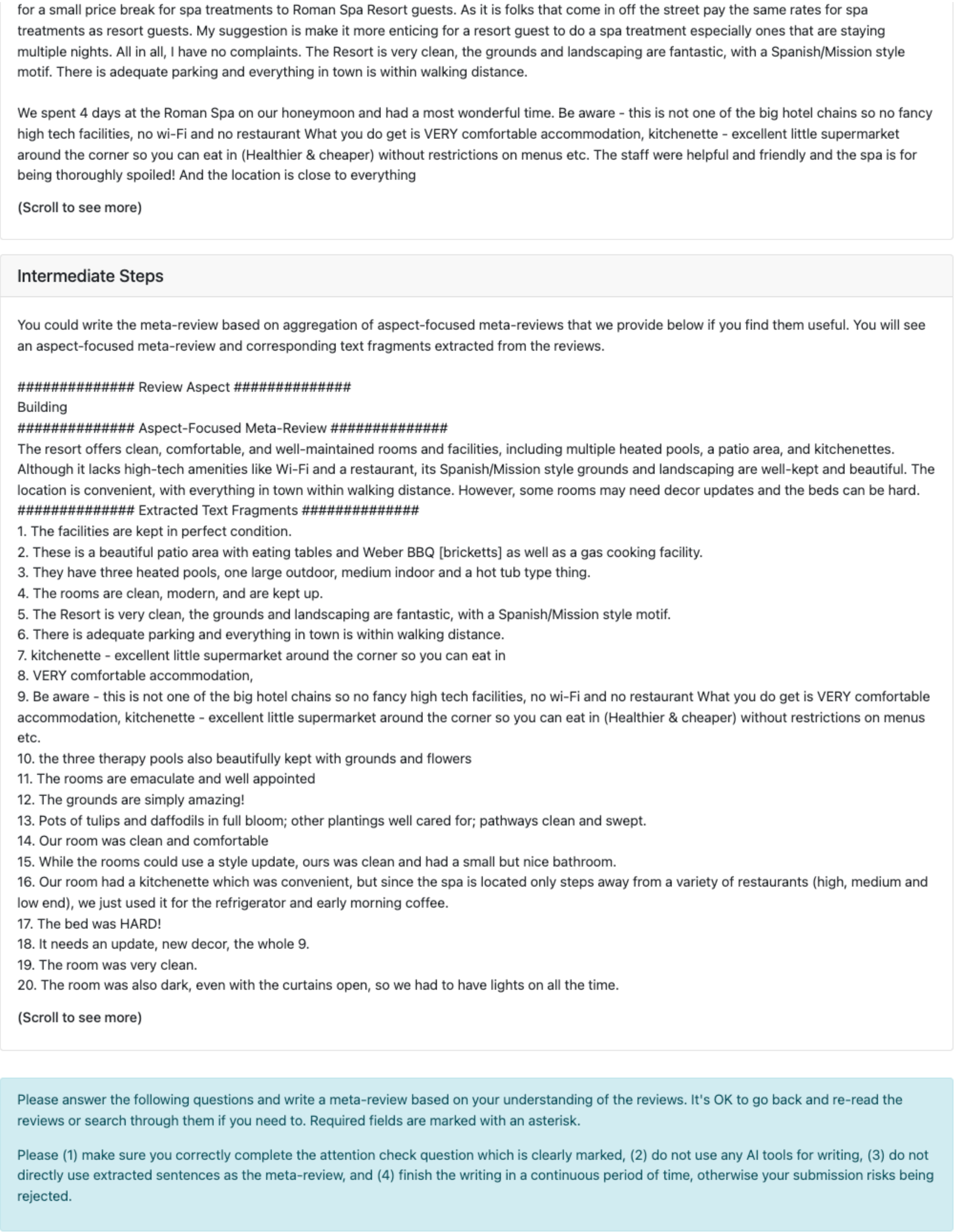
This is one of our favorite getaway spots...we were there on Halloween weekend, and there was a totally delightful parade down the main street, adding to the overall charm of the weekend! Calistoga is always full of surprises! There are three mineral pools at Roman Spa; two with jets, and one that is a swimming pool. It was raining while we were there, and they supply umbrellas if you want to use the outdoor pool! It is always a great time, even in the rain!

The hotel is centrally located in town and has it's own spa called the Baths. Love their mineral bath and massages. The best thing about the Roman spa is the mineral pool and therapy spa. We love to go down after dinner and hang out in the pool, it is delightful.

We just returned from a 4 night stay at the Roman Spa. I have not kept count but I would guess that this is our 20 + stay. We usually go twice a year for a family reunion of cousins which Calistoga is pretty central. The help is fantastic. All the way from the office to the maid and grounds men. Friendly, willing to help and very helpful if needed. The facilities are kept in perfect condition. For example, each morning the maid group go around and wipe the outdoor tables and chairs as well as the pool funature of dust and dew. These is a beautiful patio area with eating tables and Weber BBQ [bricketts] as well as a gas cooking facility. These are cleaned daily. Rooms are very clean, made up efficiently and we have never had any problems of any type at the facilities. They have three heated pools, one large outdoor, medium indoor and a hot tub type thing. I'm sure that you could find something less expensive although in my opinon, Roman Spa is not expensive, but you will not find anything with the amenities that it has for the price. [They even have loaner umbrtells in stredgit location when the weather in inclement] Can't wait to go back in the spring.

My family and I have been coming to the Roman Spa for approximately 8 years. Every New Years we come for a four day visit. We spend a great deal of time in and around the hot spring fed pools and jacuzzi's. The staff has essentially remained the same throughout the time we have been coming to the the Roman Spa. They are friendly and helpful. While young family members are welcome the Roman Spa encourages these youngsters to behave themselves in and around the pool areas so as to not disprupt the serenity of the grounds. The rooms are clean, modern, and are kept up. There are kitchenettes in some units, kitchens in others, and some just have a microwave & small refrigerator. There is no free WiFi internet available which is something that I would encourage the Roman Spa to take a look at adding in the future. However they do have a PC in the lobby for guests to check their emails, which is a good alternative. The actual Spa treatments are next door and the one thing I would encourage is

Figure 50: Interface for annotators to write meta-reviews based on different intermediate outputs (part 1).



for a small price break for spa treatments to Roman Spa Resort guests. As it is folks that come in off the street pay the same rates for spa treatments as resort guests. My suggestion is make it more enticing for a resort guest to do a spa treatment especially ones that are staying multiple nights. All in all, I have no complaints. The Resort is very clean, the grounds and landscaping are fantastic, with a Spanish/Mission style motif. There is adequate parking and everything in town is within walking distance.

We spent 4 days at the Roman Spa on our honeymoon and had a most wonderful time. Be aware - this is not one of the big hotel chains so no fancy high tech facilities, no wi-Fi and no restaurant What you do get is VERY comfortable accommodation, kitchenette - excellent little supermarket around the corner so you can eat in (Healthier & cheaper) without restrictions on menus etc. The staff were helpful and friendly and the spa is for being thoroughly spoiled! And the location is close to everything

(Scroll to see more)

Intermediate Steps

You could write the meta-review based on aggregation of aspect-focused meta-reviews that we provide below if you find them useful. You will see an aspect-focused meta-review and corresponding text fragments extracted from the reviews.

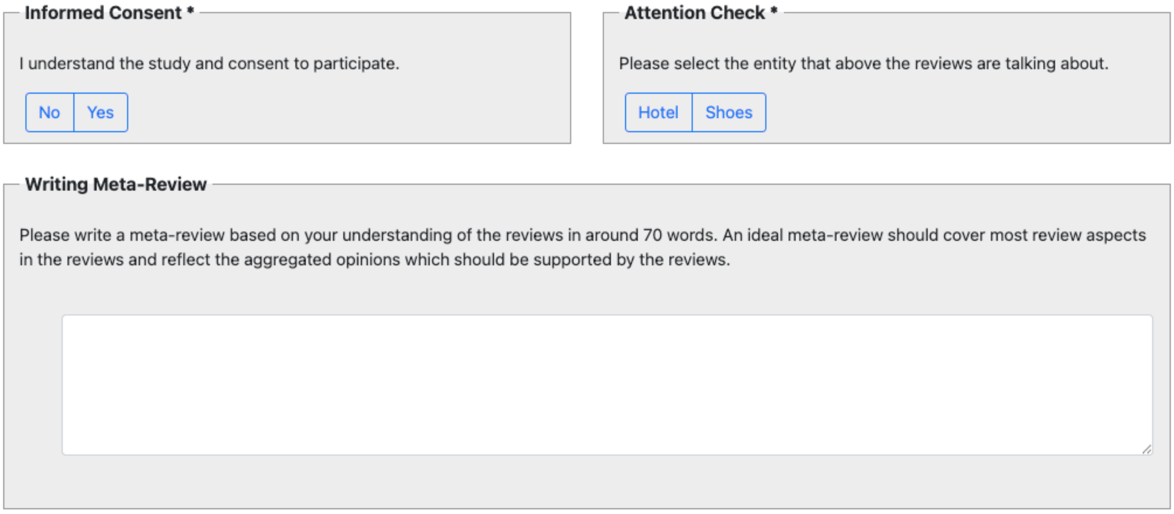
############## Review Aspect ##############
Building
############## Aspect-Focused Meta-Review ##############
The resort offers clean, comfortable, and well-maintained rooms and facilities, including multiple heated pools, a patio area, and kitchenettes. Although it lacks high-tech amenities like Wi-Fi and a restaurant, its Spanish/Mission style grounds and landscaping are well-kept and beautiful. The location is convenient, with everything in town within walking distance. However, some rooms may need decor updates and the beds can be hard.
############## Extracted Text Fragments ##############
1. The facilities are kept in perfect condition.
2. These is a beautiful patio area with eating tables and Weber BBQ [bricketts] as well as a gas cooking facility.
3. They have three heated pools, one large outdoor, medium indoor and a hot tub type thing.
4. The rooms are clean, modern, and are kept up.
5. The Resort is very clean, the grounds and landscaping are fantastic, with a Spanish/Mission style motif.
6. There is adequate parking and everything in town is within walking distance.
7. kitchenette - excellent little supermarket around the corner so you can eat in
8. VERY comfortable accommodation,
9. Be aware - this is not one of the big hotel chains so no fancy high tech facilities, no wi-Fi and no restaurant What you do get is VERY comfortable accommodation, kitchenette - excellent little supermarket around the corner so you can eat in (Healthier & cheaper) without restrictions on menus etc.
10. the three therapy pools also beautifully kept with grounds and flowers
11. The rooms are emaculate and well appointed
12. The grounds are simply amazing!
13. Pots of tulips and daffodils in full bloom; other plantings well cared for; pathways clean and swept.
14. Our room was clean and comfortable
15. While the rooms could use a style update, ours was clean and had a small but nice bathroom.
16. Our room had a kitchenette which was convenient, but since the spa is located only steps away from a variety of restaurants (high, medium and low end), we just used it for the refrigerator and early morning coffee.
17. The bed was HARD!
18. It needs an update, new decor, the whole 9.
19. The room was very clean.
20. The room was also dark, even with the curtains open, so we had to have lights on all the time.

(Scroll to see more)

Please answer the following questions and write a meta-review based on your understanding of the reviews. It's OK to go back and re-read the reviews or search through them if you need to. Required fields are marked with an asterisk.

Please (1) make sure you correctly complete the attention check question which is clearly marked, (2) do not use any AI tools for writing, (3) do not directly use extracted sentences as the meta-review, and (4) finish the writing in a continuous period of time, otherwise your submission risks being rejected.

Figure 51: Interface for annotators to write meta-reviews based on different intermediate outputs (part 2).



Informed Consent *

I understand the study and consent to participate.

Attention Check *

Please select the entity that above the reviews are talking about.

Writing Meta-Review

Please write a meta-review based on your understanding of the reviews in around 70 words. An ideal meta-review should cover most review aspects in the reviews and reflect the aggregated opinions which should be supported by the reviews.

Figure 52: Interface for annotators to write meta-reviews based on different intermediate outputs (part 3).

5.2 Reflections

Our experimental results demonstrate that our proposed opinion summarization approach based on aspect-aware decomposition produced better meta-reviews than other strong baselines in multiple domains in terms of automatic and human evaluation. Our work differs from existing research in that we aim to render the generative process less opaque across domains while generating useful aspect-aware reasoning chains that could assist humans with summarizing and aggregating reviews. Although the idea of incorporating aspects is not new, prior works either focus on extractive summarization or they do not provide interpretable intermediate steps (Amplayo et al., 2021; Bhaskar et al., 2023a). Based on our ablation study, we find that the opinion consolidation module is the most important among the three modules. Ultimately, it is interesting to find that our aspect-aware decomposition could guide LLMs to generate better intermediate steps for summarization, and that these intermediate steps can help humans write meta-reviews.

However, our work for transparent opinion summarization across different domains still has some limitations especially in aspect detection, prompt optimization, and experimental evaluation.

Our approach requires pre-defined aspects. Defining an aspect set for any domain is challenging. Our work focuses on improving the grounding and transparency of opinion summarization and investigate reasoning trajectories of prompted LLMs with given aspects, while leaving the aspect engineering for the future work. Based on our experiments, when we have a reasonable aspect set, our approach works very well in different domains. In the future work, we could automatically detect the aspects for any domain and combine the aspect detection with our current approach to make it more applicable to new domains. We could directly use the aspect detection approaches in the existing work, such as topic modelling (Zhang et al., 2023). As LLMs shows impressive performance in sentiment analysis (Zhang et al., 2024c), another straightforward idea could be prompting LLMs to detect aspects from input reviews.

Moreover, our prompts could be further optimized. For example, we can insert constraints like word count into the prompts for our modules to constrain generations. We use similar prompts across domains in the current work; however, prompt optimization could bring further benefits for specific domains.

Lastly, we could do more experiments to validate our approach. We only have experiments on datasets in English. As a future work it would be good to extend our approach to other languages, e.g., Chinese and Spanish. The current work focuses on aspect coverage and faithfulness of generated meta-reviews in evaluation. However, our approach does not explicitly address the potential of biased or harmful generations. For example, the model might under-represent opinions from female users, particularly when their opinions are expressed less frequently or in more subtle terms, thereby introducing demographic bias. We could incorporate another module to check the biases in the generated meta-reviews before presented to human users. This could reduce the risk of generating biased meta-reviews when in real-world deployment.

Chapter 6

Conclusion and Future Directions

6.1 Thesis Summary

The thesis focuses on developing computational models to integrate information from multiple sources to generate better summaries on both ideational and opinionated documents. We now summarize the contributions of the thesis by highlighting the key findings and insights based on our studies in Chapter 3, Chapter 4 and Chapter 5.

For summarization of ideational documents, which primarily contain factual information, we represent input document clusters using heterogeneous graphs. These graphs are integrated into encoder–decoder pre-trained language models (PLMs) through graph compression, where the graph of input documents is compressed into a graph representing the ground truth summary. This heterogeneous representation enables the model to better capture semantic relationships across documents. Empirical results demonstrate that our approach achieves state-of-the-art performance on widely used datasets in both news and scientific domains. These findings highlight the potential of incorporating heterogeneous graph representations for broader multi-document tasks that require integration of information from multiple sources.

For summarizing opinionated documents, we construct a dataset based on scientific meta-reviews, which can be interpreted as abstractive summaries of reviews and multi-turn discussion during the peer-review process. Unlike existing MDS datasets, our dataset features rich inter-document relationships, an explicit conversational structure, and occasional conflicting information. It thus provides a novel test bed for research on MDS of opinionated sources. To model the explicit conversational structure, we design a customized PLM that incorporates structural inductive biases derived from the conversational structure by manipulating Transformer attention mechanisms. Experiments show that our model generates higher-quality meta-reviews than strong baselines. That said, we find that all models struggle to resolve conflicts across input documents. Further human evaluation highlights the difficulty of this task.

We conduct these two summarization studies for ideational and opinionated sources when PLMs are the foundation of text summarization research. With the rapid development of LLMs, however, they have demonstrated remarkable instruction-following ability, revolutionizing nearly all NLP tasks, including text summarization, and marking a paradigm shift in 2023. To understand whether these models are truly capable of consolidating information for summary generation, we conduct an in-depth study of the scientific meta-reviewing process. We propose a three-layer framework for sentiment consolidation, guided by review aspects commonly considered by human meta-reviewers (e.g., novelty). The emergent capabilities of LLMs in chain-of-thought and in-context learning make it straightforward to integrate any symbolic 'reasoning structures' into prompts, while challenging for PLM-based approaches to integrate the reasoning steps based on on graph neural networks or attention manipulation. We implement our framework through prompting LLMs and show that it produces higher-quality meta-reviews than other strong prompting approaches.

We extend the idea of sentiment consolidation to make meta-review generation more grounded, transparent, and adaptable across domains. To this end, we propose an aspect-aware decomposition framework for opinion summarization that operates in multiple domains. The summarization process is decomposed into three sub-tasks: aspect identification, opinion consolidation, and meta-review synthesis, each implemented through prompting LLMs. This decomposition is different from our previous work even though both are guided by review aspects. Unlike our previous work that extracts sentiments based on the predefined format for scientific opinions and then predicts the overall sentiments based on them, this new approach extracts text fragments and generates aspect-focused meta-reviews directly from them. The fragment length is dynamic, allowing the model to incorporate richer opinion information, such as justifications, whereas the earlier method considered only sentiments. Experiments show that this approach produces better meta-reviews in most domains, with higher aspect coverage and more faithful generations than strong baselines, while remaining competitive with our prior method in the scientific domain. Moreover, the framework generates intermediate steps that assist human meta-reviewers in producing higher-quality reviews more efficiently. Overall, this work demonstrates that integrating reasoning steps by prompting enhances the effectiveness of LLMs in MDS.

6.2 Future Directions

Following our work on benchmark datasets, modelling methodologies and evaluation metrics in the thesis, we share insights and future directions for better multi-document summarization (MDS).

6.2.1 Reasoning-Driven Summarization

To make MDS truly useful and reliable for real-world utilities, it should go beyond simple pattern recognition or surface-level processing. MDS models should ideally not only generate high-quality summaries but also reasoning steps or explanations, i.e., how they synthesize the summarized information from the source documents. This requires the summarization models to possess the capability of language reasoning to logically analyse relationships, draw inferences, and synthesize information from inputs to produce meaningful and consistent summaries. For example, language reasoning would explain how MDS models integrate information from different source documents and solve the conflicts in the documents. This will enhance the explainability of MDS models.

As humans do not write down the reasoning processes of how they get the summarized information, in the future we could use reinforcement learning to make the models to automatically discover the reasoning processes of MDS. We could compare the difference between model reasoning processes and human reasoning processes with human evaluation to help us understand more about how humans correlate and integrate dispersed information across multiple sources.

6.2.2 Process-Aware Evaluation

Most evaluation for MDS focuses on the quality of generated summaries instead of the information integration process across source documents.

We argued previously that a useful summarization system should have the capability to not only generate a good summary but also present their reasoning steps to justify the generated summary. Therefore, we also need to design evaluation methodologies to check not only quality of generated summaries but also how well the models follow reasonable steps to get the outputs, such as how they deal with conflicts in generation. This requires deep understanding and diagnosis of model behaviours based on the reasoning steps. To achieve this, we could design evaluations on the generated reasoning steps in future work. We need to first make the models generate a summary and corresponding reasoning steps, then check the logical correctness of the reasoning steps, and lastly evaluate the faithfulness of the generated summary to the reasoning steps. For example, the reasoning steps may not logically connected with repeated or missing steps.

To this end, we could build adversarial evaluations to understand more about the model behaviours in MDS. We could change critical information in source documents to break the existing cross-document relationships, such as on human names and numerical values, and check whether the models could generate summaries having the corresponding reflection on the updated information.

6.2.3 Multimodal Documents as Input

Documents could inherently contain information in multiple modalities, not only texts but also tables, images, and even videos. For example, models must understand texts, tables, and images to write a summary of multiple scientific articles and reviews with opinionated emojis. However, existing models for MDS only focus on language information rather than visual information. When processing multimodal inputs, MDS needs to effectively integrate information from different sources in different modalities. Large vision-language models could be a simple solution to this and we could prompt them to generate the summary for multimodal documents (Liu et al., 2023a; Reid et al., 2024). To make them useful, we have to understand if they truly integrate multimodal information from different sources. However, there are few benchmark datasets and it is expensive to have human annotations with multimodal inputs. We need to develop benchmarks for the future work. The collaboration between human annotators and LLMs could be able to accelerate the annotation process. For example, we could first get annotators to write questions based on a cluster of documents, then get LLMs to extract the related elements (e.g., sentences or images) from the documents, and finally annotators could get the answers based on the extracted content by LLMs. This can make the annotation process more efficient as annotators do not need to read the entire documents.

To make the summarization on multimodal documents more transparent and grounded, we could try to decompose the process of multimodal summarization into different modules similar to what we have done in the thesis. Specifically, to fuse information from different modalities we could first transform related information from other modalities into texts, and use a large language model to summarize the texts from different sources like what we have done in the thesis. Once we could extract related information from images to texts, we can get reasonable summaries like what we can get in the thesis.

6.2.4 Large-Scale Synthetic Datasets

For MDS data scarcity is a crucial problem to train and test LLMs which serve as the modelling foundation of the task. This is because it is not cost-effective to get large-scale data for training and testing. In professional domains, e.g., medical documents, it is expensive to get human annotated data as it requires professional knowledge. For scenarios where the number of input documents is large, it is time-consuming to get humans write summaries. To solve the data scarcity issue, we could resort to synthetic datasets. For professional domains, we could get LLMs to curate data themselves and then iteratively select high-quality samples for continuous training and testing based on quality heuristics. When the number of input documents is large, they could fill up the context window of any LLMs. Although current LLMs have extended context windows even handling up to 10 million tokens, they still have limited performance on MDS when the number of input documents is large (Bai et al., 2024; Yen et al., 2025). To improve the performance of long-context models on MDS over large numbers of documents, we could run an retrieval-augmented approach to synthesize datasets. We could first retrieve critical information from individual documents, e.g., paragraphs, and then generate a summary using LLMs taking the small number of retrieved paragraphs as input. We can use the original document clusters as input and the generated summaries as output to train or test long-context capabilities of LLMs.

References

Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. 2021. Unsupervised opinion summarization with content planning. In *AAAI*, pages 12489–12497.

Stefanos Angelidis, Reinald Kim Amplayo, Yoshihiko Suhara, Xiaolan Wang, and Mirella Lapata. 2021. Extractive opinion summarization in quantized transformer spaces. *TACL*, 9:277–293.

Alan Baddeley. 2003. Working memory: looking back and looking forward. *Nature reviews neuroscience*, 4(10):829–839.

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

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. Longbench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 3119–3137. Association for Computational Linguistics.

Regina Barzilay and Kathleen R. McKeown. 2005. Sentence fusion for multidocument news summarization. *Computational Linguistics*, 31(3):297–328.

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

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *JMLR*, 3:1137–1155.

Adithya Bhaskar, Alex Fabbri, and Greg Durrett. 2023a. Prompted opinion summarization with GPT-3.5. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9282–9300, Toronto, Canada. Association for Computational Linguistics.

Adithya Bhaskar, Alexander R. Fabbri, and Greg Durrett. 2023b. Prompted opinion summarization with GPT-3.5. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 9282–9300. Association for Computational Linguistics.

Chaitanya Bhatia, Tribikram Pradhan, and Sukomal Pal. 2020. Metagen: An academic meta-review generation system. In *SIGIR*, pages 1653–1656.

Arthur Brazinskas, Mirella Lapata, and Ivan Titov. 2021. Learning opinion summarizers by selecting informative reviews. In *EMNLP*, pages 9424–9442.

Peter F. Brown, John Cocke, Stephen Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Comput. Linguistics*, 16(2):79–85.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *NeurIPS*.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *EMNLP*, pages 1724–1734.

Janara Christensen, Mausam, Stephen Soderland, and Oren Etzioni. 2013a. Towards coherent multi-document summarization. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pages 1163–1173. The Association for Computational Linguistics.

Janara Christensen, Mausam, Stephen Soderland, and Oren Etzioni. 2013b. Towards coherent multi-document summarization. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1163–1173, Atlanta, Georgia. Association for Computational Linguistics.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4299–4307.

Eric Chu and Peter J. Liu. 2019. Meansum: A neural model for unsupervised multi-document abstractive summarization. In *ICML*, pages 1223–1232.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *NAACL-HLT*, pages 615–621.

Peng Cui and Le Hu. 2021. Topic-guided abstractive multi-document summarization. In *Findings of EMNLP*, pages 1463–1472.

Tri Dao. 2024. Flashattention-2: Faster attention with better parallelism and work partitioning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Ian J Deary, Lars Penke, and Wendy Johnson. 2010. The neuroscience of human intelligence differences. *Nature reviews neuroscience*, 11(3):201–211.

DeepSeek-AI. 2024. DeepSeek-V3 technical report. *CoRR*, abs/2412.1943.

Mingkai Deng, Bowen Tan, Zhengzhong Liu, Eric P. Xing, and Zhiting Hu. 2021. Compression, transduction, and creation: A unified framework for evaluating natural language generation. In *EMNLP*, pages 7580–7605.

Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Lu Wang. 2021. Ms\^2: Multi-document summarization of medical studies. In *EMNLP*, pages 7494–7513.

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

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

Günes Erkan and Dragomir R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *JAIR*, 22:457–479.

Alexander R. Fabbri, Wojciech Kryscinski, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir R. Radev. 2021. Summeval: Re-evaluating summarization evaluation. *TACL*, 9:391–409.

Alexander R. Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir R. Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *ACL*, pages 1074–1084.

Angela Fan, Claire Gardent, Chloé Braud, and Antoine Bordes. 2019. Using local knowledge graph construction to scale seq2seq models to multi-document inputs. In *EMNLP-IJCNLP*, pages 4184–4194.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. Gptscore: Evaluate as you desire. In *NAACL*, pages 6556–6576.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2021. The pile: An 800gb dataset of diverse text for language modeling. *CoRR*, abs/2101.00027.

Yang Gao, Wei Zhao, and Steffen Eger. 2020. SUPERT: towards new frontiers in unsupervised evaluation metrics for multi-document summarization. In *ACL*, pages 1347–1354.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2023a. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *JAIR*, 77:103–166.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2023b. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *JAIR*, 77:103–166.

Gemini Team. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Technical Report*.

Demian Gholipour Ghalandari, Chris Hokamp, Nghia The Pham, John Glover, and Georgiana Ifrim. 2020. A large-scale multi-document summarization dataset from the wikipedia current events portal. In *ACL*, pages 1302–1308.

Mandy Guo, Joshua Ainslie, David C. Uthus, Santiago Ontañón, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2022. Longt5: Efficient text-to-text transformer for long sequences. In *Findings of NAACL*, pages 724–736.

M.A.K. Halliday. 1970. Functional diversity in language, as seen from a consideration of modality and mood in english. *Foundations of Language*, 6(3):322–361.

Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Liang Zhang, Wentao Han, Minlie Huang, Qin Jin, Yanyan Lan, Yang Liu, Zhiyuan Liu, Zhiwu Lu, Xipeng Qiu, Ruihua Song, Jie Tang, Ji-Rong Wen, Jinhui Yuan, Wayne Xin Zhao, and Jun Zhu. 2021. Pre-trained models: Past, present and future. *CoRR*, abs/2106.07139.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Tom Hosking, Hao Tang, and Mirella Lapata. 2023. Attributable and scalable opinion summarization. In *ACL*, pages 8488–8505.

Tom Hosking, Hao Tang, and Mirella Lapata. 2024. Hierarchical indexing for retrieval-augmented opinion summarization. *Transactions of the Association for Computational Linguistics*, 12:1533–1555.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, August 22-25, 2004*, pages 168–177. ACM.

Raghav Jain, Anubhav Jangra, Sriparna Saha, and Adam Jatowt. 2022. A survey on medical document summarization. *CoRR*, abs/2212.01669.

Hanqi Jin, Tianming Wang, and Xiaojun Wan. 2020. Multi-granularity interaction network for extractive and abstractive multi-document summarization. In *ACL*, pages 6244–6254.

Daniel Jurafsky and James H. Martin. 2025. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd edition. Online manuscript released January 12, 2025.

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

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2023. Decomposed prompting: A modular approach for solving complex tasks. In *ICLR*.

Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The efficient transformer. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *EMNLP*, pages 9332–9346.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018*, pages 66–71. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *TACL*, 10:163–177.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*, pages 7871–7880.

Junyi Li, Tianyi Tang, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024a. Pre-trained language models for text generation: A survey. *ACM Computing Survey*, 56(9).

Miao Li, Eduard Hovy, and Jey Han Lau. 2023a. Summarizing multiple documents with conversational structure for meta-review generation. In *Findings of EMNLP*.

Miao Li, Jey Han Lau, and Eduard H. Hovy. 2024b. A sentiment consolidation framework for meta-review generation. In *ACL*, pages 10158–10177.

Miao Li, Jianzhong Qi, and Jey Han Lau. 2023b. Compressed heterogeneous graph for abstractive multi-document summarization. In *AAAI*.

Wei Li, Xinyan Xiao, Jiachen Liu, Hua Wu, Haifeng Wang, and Junping Du. 2020. Leveraging graph to improve abstractive multi-document summarization. In *ACL*, pages 6232–6243.

Wei Li and Hai Zhuge. 2021. Abstractive multi-document summarization based on semantic link network. *TKDE*, 33(1):43–54.

Chin-Yew Lin and Eduard H. Hovy. 2002. From single to multi-document summarization. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 457–464. ACL.

Chin-Yew Lin and Eduard H. Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *HLT-NAACL*, pages 71–78.

Hui Lin and Vincent Ng. 2019. Abstractive summarization: A survey of the state of the art. In *AAAI*, pages 9815–9822.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. In *ICLR*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: NLG evaluation using GPT-4 with better human alignment. *CoRR*, abs/2303.16634.

Yang Liu and Mirella Lapata. 2019. Hierarchical transformers for multi-document summarization. In *ACL*, pages 5070–5081.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Llama Team. 2025. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. *Technical Report*.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 22631–22648. PMLR.

Yao Lu, Yue Dong, and Laurent Charlin. 2020. Multi-xscience: A large-scale dataset for extreme multi-document summarization of scientific articles. In *EMNLP*, pages 8068–8074.

Congbo Ma, Wei Emma Zhang, Mingyu Guo, Hu Wang, and Quan Z. Sheng. 2020. Multi-document summarization via deep learning techniques: A survey. *CoRR*, abs/2011.04843.

Tomás Mikolov, Martin Karafiát, Lukás Burget, Jan Cernocký, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010, Makuhari, Chiba, Japan, September 26-30, 2010*, pages 1045–1048. ISCA.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

OpenAI. 2025. Introducing gpt-4.1 in the api. *Technical Report*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*

35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2024. Yarn: Efficient context window extension of large language models. In *ICLR*.

Joseph Peper, Wenzhao Qiu, and Lu Wang. 2024a. PELMS: pre-training for effective low-shot multi-document summarization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 7652–7674. Association for Computational Linguistics.

Joseph Peper, Wenzhao Qiu, and Lu Wang. 2024b. PELMS: pre-training for effective low-shot multi-document summarization. In *NAACL*, pages 7652–7674.

Jason Phang, Yao Zhao, and Peter J. Liu. 2022. Investigating efficiently extending transformers for long input summarization. *CoRR*, abs/2208.04347.

Yiwei Qin, Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2022. T5score: Discriminative fine-tuning of generative evaluation metrics. *CoRR*, abs/2212.05726.

Dragomir R. Radev. 2000. A common theory of information fusion from multiple text sources step one: Cross-document structure. In *SIGDIAL*, pages 74–83.

Dragomir R. Radev and Kathleen R. McKeown. 1998. Generating natural language summaries from multiple on-line sources. *Computational Linguistics*, 24(3):469–500.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 21:140:1–140:67.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: memory optimizations toward training trillion parameter models. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, SC 2020, Virtual Event / Atlanta, Georgia, USA, November 9-19, 2020*, page 20. IEEE/ACM.

Nir Ratner, Yoav Levine, Yonatan Belinkov, Ori Ram, Inbal Magar, Omri Abend, Ehud Karpas, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. Parallel context windows for large language models. In *ACL*, pages 6383–6402.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew M. Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, James Molloy, Jilin Chen, Michael Isard, Paul Barham, Tom Hennigan, Ross McIlroy, Melvin Johnson, Johan Schalkwyk, Eli Collins, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Clemens Meyer, Gregory Thornton, Zhen Yang, Henryk Michalewski, Zaheer Abbas, Nathan Schucher, Ankesh Anand, Richard Ives, James Keeling, Karel Lenc, Salem Haykal, Siamak Shakeri, Pranav Shyam, Aakanksha Chowdhery, Roman Ring, Stephen Spencer, Eren Sezener, and et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *CoRR*, abs/2403.05530.

Nils Reimers and Iryna Gurevych. 2019a. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP*, pages 3980–3990.

Nils Reimers and Iryna Gurevych. 2019b. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP-IJCNLP*, pages 3980–3990.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Darsh J. Shah, Lili Yu, Tao Lei, and Regina Barzilay. 2021. Nutri-bullets hybrid: Consensual multi-document summarization. In *NAACL-HLT*, pages 5213–5222.

Claude E. Shannon. 1948. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27(3):379–423.

Chenhui Shen, Liying Cheng, Ran Zhou, Lidong Bing, Yang You, and Luo Si. 2022. Mred: A meta-review dataset for structure-controllable text generation. In *Findings of ACL*, pages 2521–2535.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Souza Moura, Dominik Krzeminski, Hakimeh Fadaei, Irem Ergün, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Minh Vu Chien, Sebastian Ruder, Surya Guthikonda, Emad A. Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 11521–11567. Association for Computational Linguistics.

Aviv Slobodkin, Ori Shapira, Ran Levy, and Ido Dagan. 2024. Multi-review fusion-in-context. In *Findings of NAACL*, pages 3003–3021.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Raghavi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Pete Walsh, Luke Zettlemoyer, Noah A. Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. 2024. Dolma: an open corpus of three trillion tokens for language model pretraining research. *CoRR*, abs/2402.00159.

Jianlin Su, Murtadha H. M. Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Yutao Sun, Li Dong, Barun Patra, Shuming Ma, Shaohan Huang, Alon Benhaim, Vishrav Chaudhary, Xia Song, and Furu Wei. 2023. A length-extrapolatable transformer. In *ACL*, pages 14590–14604.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *NeurIPS*, pages 3104–3112.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*, pages 5998–6008.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *ICLR*.

Danqing Wang, Pengfei Liu, Yining Zheng, Xipeng Qiu, and Xuanjing Huang. 2020. Heterogeneous graph neural networks for extractive document summarization. In *ACL*, pages 6209–6219.

Haifeng Wang, Jiwei Li, Hua Wu, Eduard Hovy, and Yu Sun. 2023. Pre-trained language models and their applications. *Engineering*, 25:51–65.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2022. PRIMERA: pyramid-based masked sentence pre-training for multi-document summarization. In *ACL*, pages 5245–5263.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *NAACL*, pages 483–498.

An Yang, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoyan Huang, Jiandong Jiang, Jianhong Tu, Jianwei Zhang, Jingren Zhou, Junyang Lin, Kai Dang, Kexin Yang, Le Yu, Mei Li, Minmin Sun, Qin Zhu, Rui Men, Tao He, Weijia Xu, Wenbiao Yin, Wenyuan Yu, Xiafei Qiu, Xingzhang Ren, Xinlong Yang, Yong Li, Zhiying Xu, and Zipeng Zhang. 2025. Qwen2.5-1m technical report. *CoRR*, abs/2501.15383.

Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. 2025. HELMET: How to evaluate long-context models effectively and thoroughly. In *The Thirteenth International Conference on Learning Representations*.

Wenpeng Yin, Dragomir R. Radev, and Caiming Xiong. 2021. Docnli: A large-scale dataset for document-level natural language inference. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 4913–4922. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *NeurIPS*, volume 34, pages 27263–27277.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. Alignscore: Evaluating factual consistency with A unified alignment function. In *ACL*, pages 11328–11348.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020a. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In *ICML*, pages 11328–11339.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with BERT. In *ICLR*.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen R. McKeown, and Tatsunori B. Hashimoto. 2024a. Benchmarking large language models for news summarization. *Trans. Assoc. Comput. Linguistics*, 12:39–57.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen R. McKeown, and Tatsunori B. Hashimoto. 2024b. Benchmarking large language models for news summarization. *Trans. Assoc. Comput. Linguistics*, 12:39–57.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. 2024c. Sentiment analysis in the era of large language models: A reality check. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906, Mexico City, Mexico. Association for Computational Linguistics.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2023. A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. *IEEE Trans. Knowl. Data Eng.*, 35(11):11019–11038.

Jinming Zhao, Ming Liu, Longxiang Gao, Yuan Jin, Lan Du, He Zhao, He Zhang, and Gholamreza Haffari. 2020. Summpip: Unsupervised multi-document summarization with sentence graph compression. In *SIGIR*, pages 1949–1952.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *CoRR*, abs/2303.18223.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *EMNLP*, pages 2023–2038.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V. Le, and Ed H. Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Hao Zhou, Weidong Ren, Gongshen Liu, Bo Su, and Wei Lu. 2021. Entity-aware abstractive multi-document summarization. In *ACL*, *Findings of ACL*, pages 351–362.

Dawei Zhu, Nan Yang, Liang Wang, Yifan Song, Wenhao Wu, Furu Wei, and Sujian Li. 2024. Pose: Efficient context window extension of llms via positional skip-wise training. In *ICLR*.