Event Understanding in Multi-Document Corpora

Adithya Pratapa

CMU-LTI-25-0111
September 2025

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Teruko Mitamura (Chair)
Graham Neubig
David R. Mortensen
Markus Dreyer (Amazon)

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Keywords: Multi-document NLP, Information Extraction, Summarization, Events

Abstract

Information about real-world events is spread over collections of text documents. For complex user queries, parsing and summarizing information from such collections has applications across the journalism, scientific, legal, and enterprise domains. Multi-document processing has been a long-standing focus for the NLP community through works on event tracking and temporal summarization. However, because of the large input lengths and task complexity, most of the progress has been limited to single-document tasks. With recent advances in large language models (LLMs) and long-context modeling, we are starting to see significant progress on these tasks. And with LLM-based systems being integrated into real-world user interfaces, it is critical to develop accurate and efficient methods for multi-document information synthesis. In this thesis, we contribute benchmarks, systems, and evaluation methods for multi-document event understanding.

In the first part of this thesis, we study coreference resolution and grounding as tools to model inter-document event relations. Driven by the progress in text summarization with LLMs, we then tackle multi-document summarization with a novel benchmark that requires synthesizing background information for complex news events. For long input tasks, we contrast compression methods with recent long-context LMs and present a hybrid retrieval-augmented generation system that provides optimal downstream performance. Finally, we study methods to improve preference-based summary evaluation through context augmentation from source documents.

Acknowledgments

I want to thank my family, friends and mentors who helped me over the years. I am forever grateful to my parents, Ramalakshmi and Saibaba, who have always supported and encouraged me to pursue my passions. I want to thank my brother Sasanka and my sister-in-law Sravya who helped me through the highs and lows of graduate school. I am blessed to have had the encouragement of my grandparents.

I would like to thank my thesis committee, Teruko Mitamura (Chair), Graham Neubig, David R. Mortensen, and Markus Dreyer (Amazon), for their support, collaborations, and thoughtful discussions during my time at CMU. I am especially grateful to Teruko for giving me an opportunity when I joined MLT and believing in me for the past six years. This thesis would not have been possible without her. She always encouraged me to pursue my research passion and helped shape my research interests. I was fortunate to have worked with Graham and David during my MLT and Ph.D., and their guidance over the years had a significant impact on my thesis. Working with Markus on my internship project and on my thesis has been truly wonderful. Our collaboration has been pivotal in helping shape my thesis.

I am really lucky to have made valuable friends through my time at CMU and Pittsburgh. I know that these paragraphs would not be able to do justice to the support I received from them, but I will try my best to acknowledge their valuable presence in my life. Spending time with my friends was what helped me through my difficult times in the Ph.D. program. I am grateful to have shared my time at LTI with these kind and amazing people, Amanda, Jared, Clara, Akhila, Lindia, Sanket, Jimin, Kimi, Kundan, Vijay, Sireesh, Simran, Hwijeen, Vidhi, Gayatri, Shruti, Aditi, Vidhisha, Jimin, Saujas, Xuhui, Sai, Khyathi, Sang, Lintang, Prakhar, Rishabh, Amrit, Ritam. I was lucky to have shared my office space with Torsten, Aman, Zhenqiao, Mingqian, Pranjal, and Weihua. I would like to thank all the staff and faculty members at LTI for their support throughout my time in the MLT and Ph.D. programs.

I want to thank my housemates and friends, Sandeep, Rawal, Ankush, for the time we spent over the last many years and especially for getting me through the pandemic. I made so many memories with my friends in Pittsburgh, Pranav, Philip, Sophia, Joseph, Jon, Nirav, Don, Suhas, Jovina, Anup, Chirag, Aditya, Arvind, DJ. A shoutout to some of my favorite spots in town, Schenley Oval, Frick, Redhawk Cafe, Page's, Washington's Landing, Carnegie Library, La Prima, Squirrel Hill, and the three rivers. Finally, I want to give a very special thanks to all my running, tennis and biking buddies who made the last six years truly fun and memorable. The Saturday morning long runs, the GAP trail bike rides and the tennis sessions in Schenley and Frick really made my time in Pittsburgh special.

Contents

1 Introduction **1**

2 Cross-document Event Identity via Dense Annotation **5**

2.1 Introduction 5
2.2 Related Work 7
2.3 Corpus Preparation 8
2.4 Annotating Coreference via Crowdsourcing 10
2.4.1 Annotation Task 10
2.4.2 Annotation Tool 11
2.4.3 Collecting CDEC annotations 12
2.4.4 Dataset Validation 13
2.5 Studying Quasi-Identity of Events 13
2.6 Baselines 16
2.7 Conclusion & Future Work 17
2.8 Appendix 17
2.8.1 Ethical Considerations 17
2.8.2 Annotation Guidelines 18
2.8.3 MTurk Consent Form 18
2.8.4 MTurk Qualification Test 18
2.8.5 HIT Template 18
2.8.6 Follow-up Questions 19

3 Multilingual Event Linking to Wikidata **31**

3.1 Introduction 31
3.2 Related Work 33
3.2.1 Entity Linking 34
3.2.2 Event Linking 34
3.2.3 Event Coreference 34
3.2.4 Multilingual Tasks 34
3.3 Multilingual Event Linking Dataset 35
3.3.1 Dataset Compilation 35
3.3.2 Task Definition 38
3.4 Modeling 39
3.4.1 BM25 39

3.4.2 Retrieve+Rank . . . 40
3.5 Evaluation . . . 41
3.5.1 Results . . . 42
3.5.2 Analysis . . . 42
3.5.3 Discussion . . . 43
3.6 Conclusion & Future Work . . . 43
3.7 Appendix . . . 44
3.7.1 Ethical Considerations . . . 44
3.7.2 Dataset . . . 44
3.7.3 Modeling . . . 44

4 Background Summarization of Event Timelines **57**
4.1 Introduction . . . 57
4.2 Related Work . . . 59
4.3 Background Summarization . . . 61
4.3.1 Task . . . 61
4.3.2 Dataset Construction . . . 62
4.3.3 Dataset Splits . . . 63
4.3.4 Inter-annotator Agreement . . . 63
4.3.5 Background Utility Score (BUS) . . . 63
4.4 Experiments . . . 64
4.5 Results . . . 66
4.5.1 Automatic Evaluation . . . 66
4.5.2 Human Evaluation . . . 66
4.6 BUS Analysis . . . 67
4.6.1 BUS–GPT . . . 68
4.6.2 BUS–Human . . . 68
4.6.3 Comparison of BUS methods . . . 69
4.7 Conclusion & Future Work . . . 69
4.8 Appendix . . . 70
4.8.1 Annotation Guidelines for Writing Background Summaries . . . 70
4.8.2 Details on the MTurk BWS Evaluation . . . 70
4.8.3 Details on the MTurk BUS Evaluation . . . 71
4.8.4 Experiment Setup . . . 71
4.8.5 Additional Results . . . 71

5 Scaling Multi-Document Event Summarization: Evaluating Compression vs. Full-Text Approaches **87**
5.1 Introduction . . . 87
5.2 Experimental Setup . . . 88
5.2.1 Datasets . . . 88
5.2.2 Methods . . . 88
5.2.3 Transformers . . . 90
5.3 Results . . . 90

5.3.1 Metrics . . . 90
5.3.2 Overall Results . . . 91
5.3.3 Analysis: Full-context & Transformer . . . 91
5.3.4 Analysis: Full-context vs. Compression . . . 92
5.3.5 Human Evaluation . . . 92
5.3.6 Recommendations for Future Work . . . 93
5.4 Conclusion . . . 94
5.5 Limitations . . . 94
5.6 Appendix . . . 94
5.6.1 Datasets . . . 94
5.6.2 Experimental Setup . . . 95
5.6.3 Full Metrics . . . 96
5.6.4 Example-level Trends . . . 96
5.6.5 Ablations . . . 96

6 Estimating Optimal Context Length for Hybrid Retrieval-augmented Multi-document Summarization 105
6.1 Introduction . . . 105
6.2 Estimating Optimal Context Length for Retrieval . . . 107
6.2.1 LM panel . . . 108
6.2.2 Generating silver references . . . 108
6.2.3 Search for optimal retrieval length . . . 109
6.3 Experimental Setup . . . 109
6.3.1 Dataset & Metric . . . 109
6.3.2 Retrieval systems . . . 110
6.3.3 Summarization systems . . . 110
6.3.4 Baselines . . . 110
6.4 Results . . . 111
6.5 Discussion & Analysis . . . 111
6.5.1 Very long-context LMs . . . 112
6.5.2 Generalization to new models . . . 112
6.5.3 Effectiveness of system pooling . . . 112
6.5.4 Effect of sample size . . . 112
6.5.5 Performance & Efficiency . . . 113
6.5.6 Limitations of Optimal Estimate . . . 113
6.5.7 Retrieval vs Iterative Compression . . . 113
6.6 Conclusion & Future Work . . . 114
6.7 Limitations . . . 114
6.8 Ethics Statement . . . 115
6.9 Appendix . . . 115
6.9.1 Context length estimates . . . 115
6.9.2 Experiment details . . . 115

7 Context Augmented LM Judges for Long Text Summarization 123
7.1 Introduction 123
7.2 Augmenting LM Judges with Compressed Context 124
7.2.1 Standard Arena: Pairwise Evaluation Task 125
7.2.2 Contextual Arena: Context-Augmented Pairwise Evaluation Task 125
7.2.3 LM judges 126
7.3 Experimental Setup 126
7.3.1 Metrics 126
7.3.2 Preference Dataset: SciArena-Eval 127
7.3.3 Synthetic Preference Dataset: SQuALITY-pref 127
7.3.4 LM judges 129
7.4 Results 129
7.5 Analysis 130
7.6 Conclusion & Future Work 131
7.7 Appendix 131
7.7.1 Computing ratings and metrics 131
7.7.2 Additional Metrics 131
7.7.3 Compression 132
7.7.4 Synthetic preference dataset 132
7.7.5 LM judges 132

8 Conclusion & Future Work 137

Bibliography 141

Chapter 1

Introduction

Information about real-world events is spread across document collections. Automatically extracting and summarizing relevant information from these collections has numerous downstream applications. For example, domain experts perform document research to synthesize findings in journalism, business intelligence, legal, and scientific studies. For critical news events, such as natural disasters, timely identification of salient content from a collection of news reports benefits news readers (Guo et al., 2013). Recent advances in large language models (LLMs) have led to significant improvements in several natural language processing (NLP) tasks. With the rapid integration of LLM-based assistants into real-world user interfaces, quantifying their ability to parse and summarize document collections is critical. In this thesis, we study the task of multi-document event understanding through the lens of datasets, systems, and evaluation methods.

The NLP community has long been interested in multi-document event understanding. Early efforts in this direction include work on event tracking (Allan et al., 1998), coreference resolution (Bejan and Harabagiu, 2008; Cybulska and Vossen, 2014), and text summarization (Dang, 2007; Dang and Owczarzak, 2009) on small collections of documents. Later efforts expanded the scope to study complex news events through work on temporal summarization (Chieu and Lee, 2004; Binh Tran et al., 2013; Aslam et al., 2015) and narrative modeling (Liu et al., 2017). A common goal of this line of research is to extract and summarize information relevant to an event from a collection of text documents. Despite these pioneering efforts, most progress in event-centric NLP has been limited to the single-document setting (Chen et al., 2021). This covered tasks such as event extraction, coreference resolution, and temporal relation extraction (Mitamura et al., 2017). While some learnings from developing single-document systems could be transferred, a multi-document setting presents unique challenges in various parts of the NLP system development stack.

First, long inputs increase the reading time and cognitive load for dataset annotators. For cross-document coreference resolution, there is also a lack of annotation toolkits that facilitate simultaneous annotation over multiple documents (chapter 2). For complex tasks such as event summarization, the annotation process must be segmented into feasible subtasks (chapter 4). For some tasks, the automatic collection of datasets is a potential solution (chapter 3). Second, the long inputs test the limits of LLM's context window. Downstream applications of these systems involve hundreds of documents, often with high information redundancy. Retrieval-augmented systems or other compression-based solutions could offer a viable path (chapter 5, chapter 6).

Third, evaluating multi-document summarization systems can be a formidable challenge to human experts. Although human evaluation is the gold standard, scaling it up to hundreds of documents might be infeasible (chapter 7). In this thesis, we make technical contributions to address these gaps in the literature around dataset curation, system development, and evaluation.

We tackle multi-document event understanding through a series of increasingly complex tasks. We start with the fundamental question of identifying the coreference relationship between event mentions across documents. Determining if two mentions are identical requires inspecting their participants and temporal and spatial attributes and often leads to cases of partial identity (Recasens et al., 2011; Hovy et al., 2013). In chapter 2, we study cross-document event coreference resolution that explicitly includes partial identity labels between mentions. In the follow-up work, we expand the scope of inter-document event relations via a knowledge-base grounding task (chapter 3). Coreference helps connect events within and across a few documents, and grounding helps scale these connections to large document collections. Building on this foundation of event identity, we then tackle the temporal aspects of events through the temporal summarization task.

Temporal summarization of news events (Chieu and Lee, 2004) has many downstream applications, but the difficulty of the task has limited progress. With recent advances in text summarization with LLMs (Goyal et al., 2022a), we can tackle these complex summarization tasks. In chapter 4, we present a novel background summarization task that provides a concise summary of an event's past to help readers keep track of day-to-day news updates. The background summarization work highlights two key challenges facing current large-scale summarization systems. First, long inputs push the limits of even the most recent long-context systems. Second, evaluating the content selection aspects of system-generated summaries remains critical but increasingly challenging in these settings. We address these specific issues in the following chapters. In chapter 5 we contrast compression-based and full-text approaches for the large-scale multi-document summarization task. Our results highlight the need for a hybrid approach that combines long-context and compression for optimal task performance. In chapter 6, we propose a hybrid RAG method that combines RAG and long-context by identifying the optimal retrieval length for a given experiment configuration. Finally, in chapter 7, we explore compression as a tool to help improve Arena-style evaluation benchmarks for complex search and synthesis tasks.

We organize the thesis into three parts: event identity, summarization, and evaluation. The first part models event identity in a cross-document setting for coreference resolution and knowledge base grounding. The second part tackles the more complex event summarization task and associated challenges with benchmark curation and system development. The third part covers content selection evaluation for event summarization. In the following, we provide an overview of each part and the corresponding chapters.

Part 1: Event Identity

Chapter 2: We develop systems for cross-document event coreference resolution (CDEC; Pratapa et al. (2021)). Unlike the within-document resolution task, the mentions from the two documents don't share any explicit context. This lack of shared context creates some ambiguity in determining the identity relationship between the mentions. Our work focuses on defining this *identity* relationship and teases apart special cases of partial identity. We curate a densely annotated dataset spanning 198 document pairs across 55 subtopics from English Wikinews, averaging 41 mentions per document. We collect evidence for partial identity between event mentions through a novel annotation workflow. We designed a custom annotation interface that is extensible to other cross-document tasks. Our dataset identifies three cases of partial identity between events: membership, sub-event, and spatio-temporal continuity. Our dataset facilitates a fine-grained coreference resolution task. We open-source our dataset, annotation interface, and systems for future work on CDEC.

Chapter 3: After coreference resolution, we look at the complementary task of event grounding (or linking). Coreference resolution requires a pairwise comparison between event mentions, being more expensive and can lead to ambiguities with the event identity. We postulate linking as a complementary task to coreference resolution, where the first mention of an event in the document is grounded in a knowledge base, and coreference resolution captures its relationship with other mentions. In Pratapa et al. (2022), we curate a new large-scale dataset and develop systems for multilingual and cross-lingual event linking to Wikidata. Our dataset captures 1.8M Wikipedia mentions in 44 languages that connect to over 10K events. We tested out-of-domain performance using a small test from Wikinews. In a follow-up work (Ou et al., 2023), we expand our event dictionary to include hierarchical event structures from Wikidata.

Part 2: Event Summarization

Chapter 4: In the first part of this thesis, we looked at fundamental information extraction tasks such as coreference resolution and knowledge-base grounding. We now expand the temporal scope of our events by studying complex news events. In Pratapa et al. (2023), we present a novel task of background summarization. For news events such as natural disasters and political conflicts, journalists curate retrospective timelines that highlight key sub-events. However, for a new observer, catching up on the historical context needed to understand a news update can be a challenging ordeal. Background summaries help alleviate this problem. We used expert annotators to curate a dataset for 14 complex news events with more than 1100 background summaries. We identify that existing metrics do not measure the downstream utility of background summaries to the readers. To this end, we propose a background utility score that measures the utility of a background summary to contextualize an update. We computed this score using an LLM-based QA-style metric and found a high correlation with human evaluation.

Chapter 5: In the follow-up work (Pratapa and Mitamura, 2025b), we contrast two classes of systems for large-scale multi-document summarization (MDS): compression and full-text.

Compression-based methods use a multi-stage pipeline and often lead to lossy summaries. Full-text methods promise a lossless summary by relying on recent advances in long-context reasoning. To understand their utility on large-scale MDS, we evaluated them on three datasets, each containing approximately one hundred documents per summary. Our experiments cover diverse long-context transformers (Llama-3.1, Command-R, Jamba-1.5-Mini) and compression methods (retrieval-augmented, hierarchical, incremental). Overall, we find that full-text and retrieval methods perform the best in most settings. With an additional analysis of the salient information retention patterns, we show that compression-based methods show strong promise at intermediate stages, even outperforming full context. However, they suffer information loss due to their multi-stage pipeline and lack of global context. Our results highlight the need to develop hybrid approaches that combine compression and full-text approaches for optimal performance on large-scale multi-document summarization.

Chapter 6: Building on our analysis in Chapter 5, we further explore the possibility of combining long-context models and retrieval-augmented systems (Pratapa and Mitamura, 2025a). Previous work has shown that these long-context models are not effective at their claimed context windows. To this end, retrieval-augmented systems provide an efficient and effective alternative. However, their performance can be highly sensitive to the choice of retrieval context length. In this work, we present a hybrid method that combines retrieval-augmented systems with long-context windows supported by recent language models. Our method first estimates the optimal retrieval length as a function of the retriever, summarizer, and dataset. On a randomly sampled subset of the dataset, we use a panel of LMs to generate a pool of silver references. We use these silver references to estimate the optimal context length for a given RAG system configuration. Our results on the multi-document summarization task showcase the effectiveness of our method across model classes and sizes. We compare against length estimates from strong long-context benchmarks such as RULER and HELMET. Our analysis also highlights the effectiveness of our estimation method for very long-context LMs and its generalization to new classes of LMs.

Part 3: Evaluating Summarization Systems

Chapter 7: In the final chapter, we revisit system evaluation for complex multi-document summarization systems. Recent advances in LLMs lead to novel applications in knowledge discovery in domains such as news, literature, and science. These LLM-based systems are being used to answer general user queries on the Web, as well as to search and synthesize findings from the scientific literature. For such complex knowledge discovery tasks, reliable evaluation of system-generated responses can be challenging and expensive. Recent Arena-style benchmarks simulate model battles by comparing model responses for a given user question. Although these methods are effective at ranking models, their application to search-and-synthesis tasks is rather limited. In this chapter, we explore the idea of augmenting context from source documents into the standard pairwise evaluations in Arena-style benchmarks. We evaluated our method on the SciArena-Eval benchmark using a variety of LM judges. Our results highlight improvements in model separability.

Chapter 2

Cross-document Event Identity via Dense Annotation

This work was published at CoNLL 2021 (Pratapa et al., 2021).[1]

2.1 Introduction

Coreference resolution is the task of identifying events (or entities) that refer to the same underlying activity (or objects). Accurately resolving coreference is a prerequisite for many NLP tasks, such as question answering, summarization, and dialogue understanding. For instance, to get a holistic view of an ongoing natural disaster, we need to aggregate information from various sources (newswire, social media, public communication, etc.) over an extended period. Often this requires resolving coreference between mentions across documents.[2]

Recasens et al. (2011) defines coreference as "identity of reference". Therefore, modeling event coreference requires understanding the extent of the shared identity between event mentions. Numerous factors determine this identity, including the semantics of the event mention, its arguments, and the document context. Resolving coreference across documents is more challenging, as it requires modeling identity over a much longer context. To this end, we identify two major issues with existing cross-document event coreference (CDEC) datasets that limit the progress on this task. First, many prior datasets often annotate coreference *only on a restricted set of event types*, limiting the coverage of mentions in the dataset. Second, many datasets and models *insufficiently tackle the concept of event identity*. As highlighted by Hovy et al. (2013), the decision of whether two mentions refer to the same event is often non-trivial. Occasionally, event mentions only share a partial identity (*quasi-identity*). In this work, we present a new dataset for CDEC that attempts to overcome both issues.

Earlier efforts on CDEC dataset collection were limited to specific pre-defined event types, restricting the scope of event mentions that could be studied. In this work, we instead annotate mentions of all types, i.e., open-domain events (Araki and Mitamura, 2018), and provide a *dense annotation* (Cassidy et al., 2014) by checking for coreference relationship between every men-

[1]Code and data are available at: `https://github.com/adithya7/cdec-wikinews`.

[2]A mention is a linguistic expression in text that denotes a specific instance of an event.

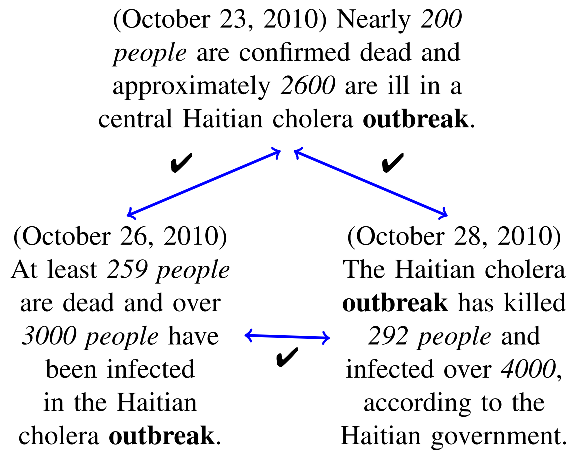


Figure 2.1: An illustration of the quasi-identity nature of events. The event [Haitian cholera] 'outbreak' is expressed by instances with varying counts of infections and deaths. The identity of this event continuously evolves over space and time, attributed to a new type of quasi-identity, spatiotemporal continuity.

tion pair in all underlying document pairs. We compile documents from the publicly available English Wikinews.[3] To facilitate our goal of dense annotation of mentions and their coreference, we develop and release a new easy-to-use annotation tool that allows linking text spans across documents. We crowdsource coreference annotations on Mechanical Turk.[4]

Prior work has attributed the quasi-identity behavior of events to two specific phenomena, membership and subevent (Hovy et al., 2013). However, its implications in cross-document settings remain unclear. In this work, we specifically focus on a cross-document setup. As highlighted by Recasens et al. (2012), a direct annotation of quasi-identity relations is hard because annotators might not be familiar with the phenomenon. Therefore, we propose a new annotation workflow that allows for easy determination of quasi-identity links. To this end, we collect evidence for time, location, and participant(s) overlap between corefering mentions. We also collect information regarding any potential inclusion relationship between the mention pair.

Our workflow allowed us to empirically identify a new type of quasi-identity, *spatiotemporal continuity*, in addition to the existing types defined by Hovy et al. (2013). Figure 2.1 illustrates this phenomenon using the case of [Haitian cholera] outbreak. The event gradually evolves over space and time, leading to cases of partial coreference. Additionally, traditional coreference annotations cluster mentions together. However, this methodology can be misleading when dealing with cases of quasi-identity (see §2.5). To overcome this limitation, we frame our annotation task as a (cross-document) mention pair linking. The proposed task simplifies the annotation process by avoiding merging quasi-identical mentions into a single cluster.

The main contributions of our work can be summarized as follows,

- We present an empirical study of the quasi-identity of events in the context of CDEC. In addition to providing evidence for previously studied types of quasi-identity (membership,

[3] https://en.wikinews.org/
[4] https://www.mturk.com/

subevent), we identify a novel type relating to the spatiotemporal continuity of events.

- We release a densely annotated CDEC dataset, CDEC-WN, spanning 198 document pairs across 55 subtopics from English Wikinews. The dataset is available under an open license. To serve as a benchmark for future work, we provide two baselines, lemma-match, and a BERT-based cross-encoder.
- To efficiently collect evidence for quasi-identity, we develop a novel annotation workflow built upon a custom-designed annotation tool. We deploy the workflow to crowdsource CDEC annotations from Mechanical Turk.

In the upcoming sections, we first position our work within the existing CDEC literature (§2.2). We then describe our methodology for preparing the source corpus (§2.3), and our crowdsourcing setup for collecting coreference annotations on this corpus (§2.4). In §2.5, we present a study of quasi-identity of events in our dataset. Finally, in §2.6, we present two baselines models for the proposed dataset.

2.2 Related Work

Event Coreference: Widely studied in the literature, with datasets curated for both within and cross-document tasks. ACE 2005 (Walker et al., 2006), OntoNotes (Weischedel et al., 2013), and TAC-KBP (Mitamura et al., 2017) are commonly used benchmarks for within-document coreference. For cross-document coreference, ECB+ (Cybulska and Vossen, 2014) is a widely popular benchmark and is an extended version of the original ECB dataset (Bejan and Harabagiu, 2008). ECB+ suffers from a major limitation with coreference annotations restricted to only the first few sentences in the documents. However, CDEC is a long-range phenomenon, and there is a need for more densely annotated datasets.

Many other datasets have since been curated for the task of CDEC. Some related works include, MEANTIME (Minard et al., 2016), Event hoppers (Song et al., 2018), Gun Violence Corpus (GVC) (Vossen et al., 2018), Football Coreference Corpus (FCC) (Bugert et al., 2020), and Wikipedia Event Coreference (WEC) (Eirew et al., 2021). However, most CDEC systems are still evaluated primarily on ECB+. Additionally, all of these datasets do not account for the quasi-identity nature of events.

Though compiled from Wikinews, CDEC annotations in the MEANTIME corpus were limited to events with participants from a pre-defined list of 44 seed entities. While the FCC corpus was also crowdsourced, the annotation unit was an entire sentence instead of a single event mention. WEC corpus uses hyperlinks from Wikipedia but primarily handles referential events. In this work, we use open-domain events and treat an event mention as our annotation unit. We collect coreference links across all the mention pairs from all the underlying document pairs.

Event Identity: Recasens et al. (2011) postulated entity coreference as a continuum, with identity, non-identity and near-identity relations. In a follow-up work (Recasens et al., 2012), they identify near-identity relations using the disagreement between annotators. They say subjects are not fully aware of the near-identity behavior, therefore making direct annotation collection hard. The continuum idea has since extended to events (Hovy et al., 2013). Determining if two event mentions are identical is not a trivial decision. It depends on the arguments of the mentions (often underspecified in the local context), the semantics of the mention, and the document context. In this work, we are specifically interested in cross-document coreference. Wright-Bettner et al. (2019) studied the impact of the subevent relationship on quasi-identity, but a more general annotation framework is missing. Accurately capturing event identity is critical to CDEC dataset construction and the subsequent modeling. Therefore, we qualitatively study this phenomenon by collecting supplementary information with each coreference link.

2.3 Corpus Preparation

In our goal of curating a CDEC dataset, we first needed to identify documents that exhibit cross-document coreference. We now describe our document collection process and our methodology for annotating event mentions in these documents.

Document Selection: To facilitate the redistribution of the documents under an open license, we prioritized collecting the documents from publicly available news sources. We chose Wikinews for three key reasons. First, the news articles were sourced from trusted news outlets and reported impartially. Second, these articles are available under an open license (CC BY 2.5), allowing easy redistribution. Finally, each article is human-labeled with categories (e.g., Disaster and accidents, Health, Sports, etc.),[5] as we describe later, this meta-information plays a significant role in our dataset collection. We use the July 1st, 2020 dump of English Wikinews, which contains a total of 21k titles (or articles/documents). These news articles are timestamped from November 2004 to July 2020. Annotating coreference between every document pair in Wikinews is infeasible. Therefore, we first identify groups of related news articles. Articles within a given group usually describe a part of a developing news story or storyline.

Identifying Storylines: To identify these latent storylines, we first construct an undirected Wikinews graph (W) with articles as nodes and add an edge between two nodes if one is mentioned under the "Related News" section in the other. We then identify cliques (C_W) (i.e., fully connected sub-graphs) in the Wikinews graph, which constitute our potential set of storylines. While the articles within each clique are related, we also want to minimize the relatedness of articles across cliques. Therefore, we construct a new graph (M), where each clique ($\in C_W$) is a node, and an edge is added between two nodes if the two cliques are not disjoint or if any two articles in the two cliques share an edge in the Wikinews graph (W). Finally, we extract maximal independent sets from M that correspond to separate storylines. Among the multiple feasible maximal independent sets, we optimize for maximum overlap in Wikinews categories of articles within each clique.

This algorithm satisfies two requirements of a CDEC dataset. First, within each storyline, all articles are related to each other. Second, articles from different storylines aren't adjacent in the Wikinews graph (W); thereby, they are very likely unrelated.

[5] https://en.wikinews.org/wiki/Wikinews:Categories_and_topic_pages

# topics	1
# subtopics	55
# documents	176
# sentences per doc (avg.)	14.6
# tokens per doc (avg.)	344
# event mentions	7220
# mentions per doc (avg.)	41
# document pairs	198
# CDEC links	4282
# CDEC links per document pair	21.6
# full coreference links	2914
# partial coreference links	1368

Table 2.1: An overview of the compiled CDEC dataset.

For this work, we narrow our focus only to articles in the "Disaster and Accidents" category on Wikinews.[6] Following the terminology of prior work, our dataset constitutes of a single topic (Disaster and accidents) and 55 subtopics (individual storylines). We restrict CDEC annotations to subtopics that contain 3 or 4 documents. Our algorithm aims for completeness of the CDEC dataset by maximizing for intra-subtopic and minimizing inter-subtopic coreference.

Event Mention Identification: To annotate the event mentions in the above-collected documents, we first run a combination of mention detection systems. Specifically, we use the OpenIE system (Stanovsky et al., 2018) from AllenNLP (Gardner et al., 2018) and an open-domain event extraction system (Araki and Mitamura, 2018). The former is effective at extracting verbal events, whereas the latter is good at nominal events. In contrast to most prior work, we do not restrict the mentions to specific event types or salient events. We believe it is important to study all underlying events to achieve a complete understanding of the corpus. Since the quality of mention identification is critical to our CDEC dataset, we ask an expert to go through the automatically identified mentions and add/edit/delete mentions using the Stave annotation tool (Liu et al., 2020).[7]

Table 2.1 presents the overall statistics of our document corpus. Our documents are ∼14.6 sentences long, comparable to prior work, ECB+ (16.6), GVC (19.2), and FCC (34.4). However, our documents are significantly more dense in terms of event mentions. Our documents contain ∼41 mentions (on avg.), much higher compared to prior work, ECB+ (15.3), GVC (14.3), FCC (5.8). Given the dense nature of our documents, we appropriately design our annotation task and interface.

[6] https://en.wikinews.org/wiki/Category:Disasters_and_accidents

[7] the expert annotator is an author of this work.

2.4 Annotating Coreference via Crowdsourcing

Corefering event mentions share their identity. However, the extent of sharing for them to be considered coreferential is unclear. To empirically study this behavior, we crowdsource annotations on Mechanical Turk. We use the crowd workers' responses to analyze the influence of quasi-identity on coreference decisions.

2.4.1 Annotation Task

The input to our annotation task constitutes a pair of documents, with all event mentions pre-identified. Annotator iterates through every mention on the left document and select corefering mentions from the right document. We also provide the document titles and publication dates to help set the context for the articles. Note that we focus solely on cross-document coreference in this work and leave the addition of within-document links to future work.

Prior work has highlighted the difficulty in capturing event coreference, specifically in cases where the mentions are only quasi-identical (Hovy et al., 2013). Notably, Recasens et al. (2012) found direct annotation of partial identity to be a difficult task. Therefore, we propose to analyze this behavior by collecting supplementary information from the annotators. For each coreference link created by an annotator, we ask them four *follow-up questions*, 1. overlap in location, 2. overlap in time, 3. overlap in participants, and 4. potential inclusion relationship.[8] Annotators implicitly consider these aspects when making a coreference decision; therefore, responding to these questions won't increase the annotators' cognitive load significantly. As we show in §2.5, the responses to these questions help us tease apart the cases of partial identity.

Unlike within-document coreference, disjoint narratives between documents often complicate CDEC annotation tasks. Wright-Bettner et al. (2019) analyzed this behavior in detail and proposed a new `contains-subevent` label for within-document links that improved annotator agreement and reduced inconsistencies. However, they rely on experts to create the within-doc `contains-subevent` label beforehand. Instead, we focus solely on cross-document links and frame the task as a simple pair-wise classification. Our framing allows non-expert annotators to make decisions without concern for complex granularity issues. Our follow-up question regarding inclusion facilitates a post hoc analysis of the event granularities in our dataset.

To ensure completeness of our CDEC dataset, we collect annotations for each pair of documents in a given subtopic (§2.3). As highlighted earlier, the quasi-identity of events may or may not allow for the application of transitivity property. Therefore, in our dataset, we cannot expand coreference links using transitivity. So collecting annotations between each pair in a given subtopic is necessary.

Annotation Guidelines: Events are commonplace in the newswire; therefore, it is feasible to explain the concept of events and their coreference via simple example-based guidelines. In our guidelines, we first define *events* and then provide numerous examples of identical and non-identical event mentions, with detailed explanations. Following prior work (Song et al., 2018),

[8] see Table 2.14 in Appendix for the exact formulation of these follow-up questions.

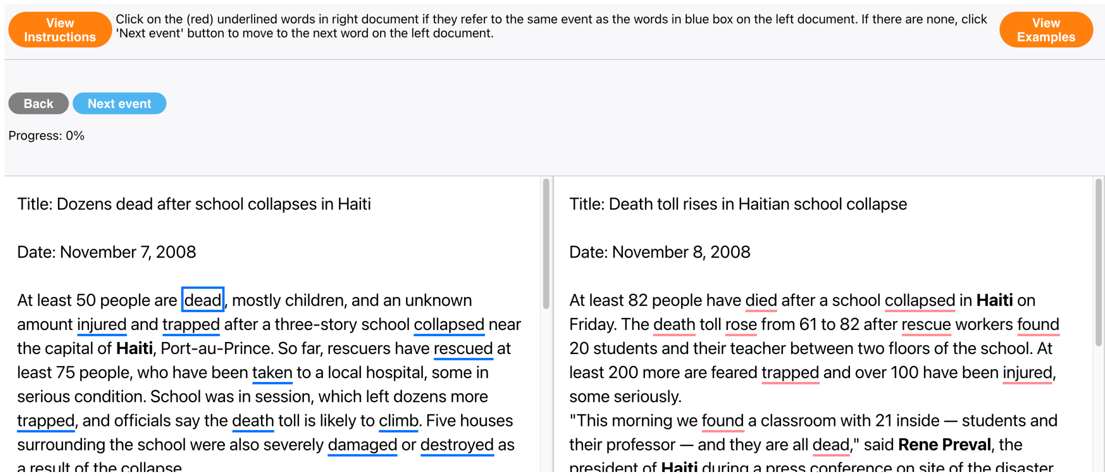


Figure 2.2: Tool for annotating cross-document event coreference. The two documents are shown side-by-side, with event mentions pre-highlighted. We provide on-screen instructions as well as dedicated pages for viewing detailed instructions and examples. As seen in the example here, we allow annotation of every pair of mentions in the given document pair. In our annotation effort, we present every pair of related documents on this tool, leading to a *densely* annotated dataset.

we rely on the annotator's intuition to decide coreference.[9]

2.4.2 Annotation Tool

To efficiently crowdsource annotations, we require a tool that is both easy-to-use and customizable to our workflow. For this purpose, we build upon the Forte[10] and Stave[11] toolkits (Liu et al., 2020). We extend both the toolkits to support cross-document linking as required by our annotation task. Figure 2.2 presents a snapshot of our annotation interface. We highlight event mentions in both the documents and allow the annotator to iterate through each mention on the left document. In addition to dedicated links to instructions and examples, we provide on-screen instructions to assist the annotator in real-time. We also use an English NER tool (Ma and Hovy, 2016) to highlight the named entities in the documents. These entities help the annotator keep track of various event participants in the two documents.

We utilize this tool for our entire dataset collection. While we show an application of our annotation tool for CDEC, we believe it's adaptable to other cross-document tasks like entity coreference and event/entity relation labeling tasks. We release our toolkit to encourage future work on cross-document NLP tasks.[12]

[9] see 2.8.2 in Appendix for complete guidelines.
[10] `https://github.com/asyml/forte`
[11] `https://github.com/asyml/stave`
[12] `https://github.com/adithya7/cdec-ann-tool`

2.4.3 Collecting CDEC annotations

We crowdsource annotations for CDEC using Amazon Mechanical Turk (MTurk). Each Human Intelligence Task (HIT) constitutes annotating cross-document links for one pair of documents. We obtained IRB approval and set our HIT price based on preliminary studies.[13] On MTurk, we restricted our HITs to crowd workers from the US and set our qualification thresholds for % HITs, and total HITs approved as 95% and 1000 respectively. We paid a fair compensation of $10.9/hour on average.[14] Our annotation task requires proficiency in English, as well as a good understanding of event coreference. To this end, we attach a qualification test with eight yes/no questions regarding event coreference, with a qualification threshold of 75%.[15]

For each document pair, we collected annotations from three different crowd workers. In each task, crowd workers go through the two documents and develop a high-level understanding of the news story. They then iterate through the mentions in the left document, in the narrative order, to identify potential cross-document coreference links. From our preliminary studies, we found that annotators spend considerable time reading the two documents. Therefore, to make the best use of the crowd workers' time and effort, we group HITs that constitutes document pairs from the same subtopic. This way, if the crowd worker chooses to, they can annotate the entire subtopic in one sitting, sharing their understanding of a document from one HIT to the next. In total, we collected annotations for 198 document pairs, spanning 176 unique documents and 55 subtopics from 46 crowd workers.

Inter Annotator Agreement (IAA): For each pair of documents, we collect annotations from three crowd workers. Our setup allows the annotator to decide coreference for every mention pair. To measure IAA, we associate a value to each mention pair (corefering or non-corefering) and compute Krippendorff's α. For coreference links, we observed an α of 0.46, indicating moderate agreement (Artstein and Poesio, 2008).[16] Additionally, we compare the impact of the quasi nature of coreference on the annotator agreement. In our dataset, 31% of the full-coreference links have a perfect majority (3/3 annotators). However, only 13% of the partial-coreference links have the same (see section 2.5 for the methodology used to determine partial coreference). This sharp contrast illustrates the difficulty in capturing partial coreference links.

Selecting CDEC links: For each pair of mentions, we take a majority vote on the three crowd-sourced annotations. In our preliminary analysis, we found many valid coreference links annotated by just one crowd worker. While we encourage the crowd workers to annotate every pair of corefering mentions, they occasionally miss links. Therefore, to ensure completeness of our dataset, we use an adjudicator to go through the single-annotator links to decide if they are in-fact corefering or not.[17]

[13] see 2.8.1 in Appendix for more details.

[14] The median pay was slightly higher at $16.3/hour. Both mean and median pay are above the current minimum wage requirements in the United States.

[15] see 2.8.4 in Appendix for the test format and the questions.

[16] It's important to note that we compute IAA on our entire dataset. Our IAA score is comparable to those of quasi-relations from Hovy et al. (2013).

[17] the adjudicator is an author of this paper.

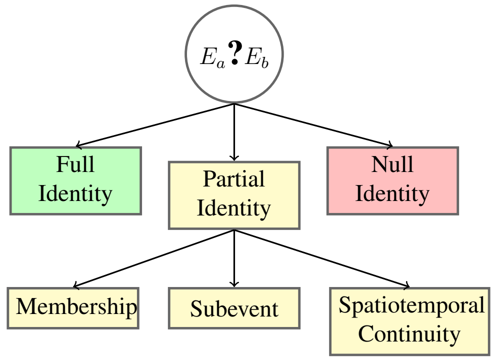


Figure 2.3: A taxonomy of event identity. While full and null identities are well understood, the definition of partial identity is still evolving. We present the three types of partial identity found in our dataset.

Table 2.1 presents an overview of the compiled CDEC dataset. Unlike prior work, we do not create mention clusters by expanding the links via transitive closure. As we show in §2.5, quasi-identity of events warrants the need to analyze coreference at the level of mention pairs instead of clusters.

2.4.4 Dataset Validation

To facilitate benchmarking future coreference resolution models, we split our dataset into train and test. Of the 55 subtopics, 40 are for model training and development, and 15 are for the unseen test set. Given the importance of the test set quality, we perform expert validation on a randomly selected subset of 18 document pairs from our test set. The expert inspected the annotated coreference links in the subset and found 97.5% precision (549/563 were corefering). On the other hand, measuring the recall is hard due to a large number of mention pairs. Therefore, we specifically focus on two types of potentially missing coreference links, 1. mention pairs that share the same head lemma (but not annotated as corefering), 2. mention pairs that are part of a non-transitive triplet.[18] Upon inspection by the expert, we find that majority of lemma-match links are non-corefering (50/565 were corefering), while a majority of non-transitive pairs are corefering (149/173 were corefering). This result indicates the scope for improvement in tackling missing coreference links. We leave this extension to future work.

2.5 Studying Quasi-Identity of Events

Numerous factors determine the identity of an event mention, including the semantics of the mention, arguments (place, time, and participants), and the overall document context. Therefore, overlap in these factors determines the extent of coreference between two given mentions. This overlap leads to cases of partial (quasi-) identity. Our annotation workflow allows for empirical investigation of this phenomenon, and we summarize our observations through a taxonomy of

[18] (E_A, E_B, E_C) is a non-transitive event triplet if E_A corefers with E_B, E_B corefers with E_C, but E_A and E_C are non-corefering.

Membership	
1a	The **fire** has burned about 4400 acres so far and 15 homes have been lost, however there have been no reported injuries or deaths.
1b	Reports say that the amount of people fleeing from their homes in California located in the United States due to **wildfires** has reached the 1,000,000 mark as the fires continue to grow.
2a	Several aftershocks have rocked the same area, the latest measuring 7.1, had a depth of 10 km. It was first reported to be a 7.3 **aftershock**.
2b	Some smaller **aftershocks** with magnitudes between 5.2 and 5.7 were also reported in the region.
2c	That quake was followed by as many as 60 **aftershocks** for at least a week, with some ranging as high as magnitude 7.8.
Subevent	
3a	A freight train in Lviv, Ukraine derailed, caught fire, and spilled a toxic chemical, releasing dangerous fumes into the air early Tuesday morning (local time), and people who live near the site of the **crash** are still becoming sick.
3b	The available information about the phosphorous cloud following the railway **accident** in the Ukraine last Monday is becoming more and more cryptic.
4a	During the fifteen days of the trial, the prosecutors called 92 witnesses to testify as to the chaotic scenes following the **bombing**.
4b	Two **explosions** within seconds of each other tore through the finish line at the Boston Marathon, approximately four hours after the start of the men's race.
Spatiotemporal Continuity	
5a	Tropical **storm** Richard is nearing hurricane strength with winds of 70 mph (115 kph) as it lashes Honduras with heavy rains
5b	**Hurricane** Richard made landfall in Belize about 20 mi (35 km) south-southeast of Belize City with winds of 90 mph (150 kph) at approximately 6:45 local time (0045 UTC) according to the National Hurricane Center (NHC)

Table 2.2: An illustration of quasi-identity of event mentions across documents. These examples cover the three identified types of quasi-identity, membership, subevent, and spatiotemporal continuity.

event identity in Figure 2.3. Except for Wright-Bettner et al. (2019), prior CDEC datasets do not account for the partial identity during the annotation process. Hovy et al. (2013) have previously proposed two types of partial identity, membership, and subevent. In addition to providing evidence for these two types in our dataset, we also identify a novel type of partial identity termed as *spatiotemporal continuity*.

Collecting Partial Identity: We use the responses to follow-up questions for qualitatively analyzing cases of partial identity. We consider a link to be a case of partial identity if a strict majority of annotators indicate one of the following. First, there is an inclusion relationship between corefering mentions. Second, the two overlap in place, time, or participants. With this screening methodology, we found ~32% of the total CDEC links to be candidates for partial identity (Table 2.1). We qualitatively analyze the dataset and identify three types of partial identity, 1. Membership, 2. Subevent, and 3. Spatiotemporal continuity. Table 2.2 illustrates each type with examples from our compiled dataset.

Membership: An event mention E_a is a member of event mention E_b. Consider the two sentences, 1a, and 1b. The mention 'fire' (1a) denotes a specific wildfire, whereas 'wildfires' (1b) denotes a group of wildfires, including the one in 1a. The concept of partial identity often challenges the transitivity assumption of coreference. For instance, the mentions [smaller] 'aftershocks' (2b) and [7.1] 'aftershock' (2a) share no identity, thereby, non-coreferential. However, both the mentions partially corefer with [60] 'aftershocks' from 2c.

Subevent: An event mention E_a is a subevent of event mention E_b. This behavior can be seen in the coreference between the 'crash' event from 3a, and the 'accident' event from 3b. While the 'accident' event involves many individual events, derailed, caught fire, spill chemical, and release fumes, it partially corefers with the event 'crash' from 3a that likely refers only to the derailment. Similarly, consider the case of the Boston Marathon Bombing in examples 4a and 4b. The 'bombing' event from 4a refers to the whole incident, whereas the 'explosions' in 4b refers to specific subevents of the 'bombing'.

Spatiotemporal Continuity: The identity of an event can *continuously* evolve over space and time. Consider the two mentions, 'storm' and 'Hurricane' from Table 2.2 (5a, 5b). At a high level, these mentions are corefering because they denote the same event (storm Richard). However, the expressions of this event differ slightly across the two documents. In the former, it's a storm (with 70mph winds) having an impact in Honduras, whereas, in the latter, it's a hurricane (with 90mph winds) impacting Belize. Similar behavior is visible with the [Haitian cholera] 'outbreak' event from Figure 2.1. The outbreak gradually evolves, with growing infection (2600 $\rightarrow$ 3000 $\rightarrow$ 4000) and deaths (200 $\rightarrow$ 259 $\rightarrow$ 292). In both of these examples, we observe the event changes gradually and is always continuous in both space and time dimensions.[19]

In line with prior work on entities (Recasens et al., 2011), we believe identity and coreference of events to be a continuum. Our dataset already includes many instances of partial identity to support this hypothesis. The above-described cases of partial identity (membership, subevent, and spatiotemporal continuity) will pose new challenges to future dataset collection efforts. We believe our annotation workflow and guidelines will be of use to future work.

In this section, we establish a clear case for tackling partial identity within the coreference resolution task. However, in practical settings, the boundaries between full, partial, and null identities remain fuzzy. As seen in our analysis on the inter-annotator agreement, humans find it hard to identify cases of partial coreference. In the downstream coreference resolution task, users are primarily interested in knowing if two given mentions share an identity or not. Therefore, we propose to view both full and partial identity under a single coreference label ('coreference') and contrast them against cases with no shared identity ('non-coreference'). Compared to prior datasets, this presents new challenges in tackling partial identity within the 'coreference' label.

[19]We borrow the term spatiotemporal continuity from the Philosophy literature. It describes the properties of well-behaved objects (Wiggins, 1967). A similar treatment for entities is presented in Recasens et al. (2011).

2.6 Baselines

We define the task as a mention pair classification problem. Due to the quasi-identity nature of event mentions (§2.5), we do not cluster mentions in coreference groups. Additionally, we consider both full and partial identity under the coreference label. We present two baseline models, lemma-match, and a cross-encoder model. We split the dataset of 55 subtopics into train and test, with 40 subtopics for training and development, and 15 subtopics for the held-out test set. For our experiments, we assume gold mentions and subtopic information.[20]

Lemma-match: For our first baseline, we implement the traditional lemma-match baseline. We use spacy's large model[21] to extract the head lemma of the event mentions, and consider two mentions corefering if the lemma's match. Following Upadhyay et al. (2016), we also experiment with a Lemma-δ baseline. In our experiments, we found the best dev performance with δ=0, resolving to a simple lemma baseline. This could be due to our assumption of access to gold subtopic information.

Cross-Encoder: As a second baseline, we implement BERT-based cross-encoder model. The input consists of a pair of sentences with both mentions highlighted using special tokens to indicate the start and end of mention spans (<E>, </E>). We first concatenate the two event-tagged sentences (with [SEP] token) and pass it through a bert-base-uncased encoder. We then perform mean pooling on the event start tags (<E>), and pass the pooled embedding through a linear classification layer to predict coreference vs. non-coreference. For training the cross-encoder, in addition to the positive coreference pairs, we generate two types of negative mention pairs. For the first type, we collect non-coreference mention pairs from sentences that have a coreference link between a different mention pair. For the second type, we extract non-coreference mention pairs from random sentence pairs between the documents. During training, we use a dataset ratio of 1:5:5 (positive:negative-I:negative-II). We use huggingface transformers (Wolf et al., 2020), and train the model using AdamW (Loshchilov and Hutter, 2019) with an initial learning rate of 2e-5. We also use a linear warmup scheduler, with 10% of training steps for warmup. We finetune the # epochs and positive:negative dataset ratio during the development stage (5-fold cross-validation) and use the best configuration when training on the entire train set.

Results: Table 2.3 presents the results of our baselines. For model development, we perform 5-fold cross-validation on the training set (40 subtopics). To report the results on the held-out test set (15 subtopics), we train the model's best configuration on the entire training set. We report precision, recall, and F1 scores of the coreference label averaged on five different runs. The lemma baseline only achieves an F1 score of 48.2, indicating that the proposed dataset is lexically diverse. The cross-encoder improves upon the lemma baseline, especially on the recall. Upon inspection of development set predictions, we observe two possible error cases for the cross-encoder model. First, the model struggles at the cases of partial identity ('explosion' vs. 'incident' and 'evacuate' vs. 'evacuations'). This drawback of cross-encoder indicates that the

[20] topic-level performance (Cattan et al., 2021)

[21] `en_core_web_lg` from `https://spacy.io`

Model	Dev			Test		
	P	R	F1	P	R	F1
Lemma-match	46.6	54.9	49.9	42.3	56.0	48.2
Cross-Encoder	43.1 $\pm$ 0.6	75.4 $\pm$ 0.5	54.3 $\pm$ 0.5	45.9 $\pm$ 0.8	77.3 $\pm$ 1.1	57.6 $\pm$ 0.6

Table 2.3: Baseline results on development and test sets. For cross-encoder, we report the average scores and their standard deviation across five runs.

model requires a deeper understanding of event identity. Second, the cross-encoder model is often limited by the information available in a single sentence. It is known the event arguments are often underspecified in the local context (Ebner et al., 2020); therefore, increasing the context to a paragraph or the entire document might help improve the performance.

2.7 Conclusion & Future Work

In this work, we present a study of the identity of events through annotation of cross-document event coreference. We use a custom-designed annotation tool to collect coreference annotations on a subset of English Wikinews articles. We release our dataset, CDEC-WN, under an open license to encourage further research on event coreference. By collecting evidence for the extent of shared identity between events, we identify three types of partial-identity, membership, subevent, and spatiotemporal continuity. To serve as a benchmark for future coreference resolution systems, we provide results on two baseline models, lemma-match and BERT-based cross-encoder. We believe that our work will encourage further research on the identity of events in the context of CDEC. Potential future directions include expanding CDEC-WN to include within-document coreference links, designing coreference resolution systems that account for cases of partial identity between mentions, and expanding the study of the partial identity of event coreference to new domains.

2.8 Appendix

2.8.1 Ethical Considerations

In our dataset construction, we follow the standard norms for ethical research involving human participants. We obtained IRB approval before starting our study. Our pilot study indicated that each HIT takes ∼10-15 minutes; therefore, we set the price of individual HIT to be $2.3. Overall, we paid a fair compensation of $10.9/hour (with median pay of $16.3/hour). For each HIT, the crowd workers on Mechanical Turk have signed the informed consent form before starting the task (see 2.8.3 in Appendix). We provided clear instructions for using our annotation tool, both within and through an instructional video. We provide positive and negative examples to illustrate event coreference to the crowd workers (see 2.8.2 in Appendix). Our dataset is limited to the English language, specifically for text documents relating to Disasters and accidents. While we have taken specific steps to improve the quality of our dataset, there might be incorrect or missing coreference links. However, we believe that such incorrect/missing links will not create additional risks to the models trained on our dataset.

2.8.2 Annotation Guidelines

To explain the task of cross-document event coreference to crowd workers on Mechanical Turk, we present detailed example-based guidelines (Table 2.6, Table 2.7). Additionally, we provide crowd workers with detailed instructions to our annotation interface (Table 2.4, Table 2.5). Workers view these instructions before the start of each task and optionally during the task. In our HIT, we also link to a 1-minute video tour of our annotation interface.

In our guidelines, we only present examples of full and null coreference. While we consider membership a form of coreference (partial), we don't train the crowd workers on full and partial identity.

2.8.3 MTurk Consent Form

A consent form is attached to the start of each HIT. Crowd workers are required to go through the form and provide their consent before starting the task. Anonymized version of the consent form is presented in Table 2.8 and Table 2.9. We anonymize the document for the conference review process.

2.8.4 MTurk Qualification Test

To identify high-quality crowd workers, we design a qualification test and add it as an additional requirement to solving our HITs.

Test Questions

In the qualification test on MTurk, we randomly select eight questions from a pool of 20 questions. Table 2.10 and Table 2.11 list all the questions.

Test Format

Table 2.12 presents the format of the qualification test used for screening crowd workers.

2.8.5 HIT Template

Table 2.13 presents our HIT layout. Our layout is simple, and all of our annotations are collected using our custom-designed annotation tool.

2.8.6 Follow-up Questions

Table 2.14 lists the four follow-up questions. We present these questions for each coreference link annotated by the crowd worker.

Instructions for using the tool

This tool can be used to select events that are the same across the two given documents.

How to open instructions
embedded GIF
How to annotate one pair of events
embedded GIF
How to delete previous annotations
embedded GIF
How to proceed to the next event
embedded GIF

At any point during the task, you can click on the "View Instructions" button to read these instructions.

What is this task about?

- Two related documents are presented side-by-side on the tool.
- A few words in both the documents are underlined and these are referred to as events.
- The task is to select events from the right document that are the same as the currently highlighted event in the left document.

How should I solve this task?

- When you first start the task, make sure you read through both the left and right documents to get an overall understanding of the two documents.
- At each step, an event is highlighted in a blue box on the left document (aka. target event). Now, your goal is to identify underlined events from the right document that are the same as the target event from the left document.
- Once you select an event from the right document (an annotation), you are presented a few follow-up questions. Make sure you answer these questions to the best of your knowledge.
- If you change your mind while answering the questions, you can click the "Cancel" button to remove your annotation.
- After you have identified all possible same events from the right document (if any), please use the "Next event" button to move to the next target event on the left document.

Table 2.4: Instructions as shown to the annotators on the interface.

Instructions for using the tool (contd.)

FAQs

Q: I made a mistake and incorrectly marked two events as the same. How do I correct this?
If you are still answering the follow-up questions, you can just click on the "Cancel" button. If you have already moved to the next target event, you can use the "Back" button to move back the previously finished target events.

Q: I am not sure how to respond to the follow-up questions. How should I proceed?
The follow-up questions help us understand more about your decision that two events are the same. It is important to note that the response to these questions need not always be "Yes". In fact, in many cases, you may not have enough information to respond with a definite "Yes" or "No", then please feel free to select "Not enough information".

Q: How do I decide if two events are the same or different?
We understand that this decision is not always easy. To help you with this, we compiled a bunch of examples. You can quickly glance through them using the "View Examples" button on the tool.

Q: How do I contact the authors of the task?
For any comments, feedback and/or suggestions, please use this form (XXXX). We strive to make this a great experience for you.

Table 2.5: Instructions as shown to the annotators on the interface. (contd)

Examples

Goal of the Task
You will help us identify the same events from different documents.

What is an event?
People use text to describe what happen(ed) in the world. These are called events in text. We often use verbs, sometimes even (pro)nouns, and adjectives as events. For example:

It rained a lot yesterday.
There was a fire last night.
He got sick.

How do we know that the two events are the same?
In the following examples (1 to 5), two events are the same.

1. When two events refer to the same thing, they should be the same in terms of meaning, or semantically identical.
   - Taken as a whole, the evidence suggests that the plan to bomb the Boston Marathon took shape over three months.
   - Dzhokhar Tsarnaev apologized for suffering caused by the Boston Marathon bombing.

2. When two events are the same, one event may be the synonym for the other.
   - A 16-year-old southern Utah boy was accused of bringing a homemade bomb to his high school.
   - The teen was charged Monday with attempted murder and use of a weapon of mass desctuction, both first-degreen felonies.

3. Sometimes one event may be the pronoun (e.g.,it) or the anaphora (e.g., this, that) of the other, when they are the same.
   - Both drones carried explosives, and no YPF ("People's Defence Units") fighters were injured in the incident.
   - This would not be the first terrorist drone strike.

4. The same events do not have to take place at the same time. In the following example, one event ("go") would happen in the future, while the other ("went") did occur.
   - The couple had been planning to go to Paris for a long time.
   - They finally went there last month.

5. Sometimes the same events are described from different perspectives. The following example refers to the exchange of the gift from two perspectives.
   - John gave a gift to Mary.
   - Mary received a gift from John.

Table 2.6: Examples for coreference and non-coreference, as shown to the annotators on the interface.

Examples (contd.)

In the following examples (6 to 8), two events are not the same.

6. When one event is a part of the other larger event, they are not the same.

- Following the trial of Mahammed Alameh, the first suspect in the bombing, investigators discovered a jumble of chemicals, chemistry implements and detonating materials.
- The explosion killed at least five people. ("bombing" refers to the entire process which starts with making a bomb and ends with destructions, damages and injuries, while "explosion" is a smaller event that occurs in that processes)

7. Two events are not the same even if they are the same semantically. The first example refers to the general bomb-making process, while the second one indicates a particular bomb-making event that took place in the garage.

- They obtained the online manual of bomb-making. (general bomb-making process)
- They made a bomb in the garage. (specific bomb-making event that happened in the specific place)

8. When one event consists of, or is a member of the other event, they are not the same. The first example refers to the specific death of a 44-year-old man, while the second one refers to the deaths of 305 people.

- The government announced that a 44-year-old man died from the COVID. (death of a 44-year-old man)
- There are more than 14,300 confirmed COVID cases, and 305 people have died. (deaths of 305 people)

Table 2.7: Examples for coreference and non-coreference, as shown to the annotators on the interface. (contd)

Consent From

This task is part of a research study conducted by XXX at XXX and is funded by XXX.

Purpose
The goal of this study is to collect datasets of coreference-labeled pairs sampled from public online news articles through the help of crowd workers.

Procedures
You will be directed to a website implemented by the research team to complete the task. You will be asked to read upto 3 pairs of articles. For each pair of articles, you will need to label pieces of text that refer to the same event, and answer additional questions about your labeling. Labeling one pair of articles whose length sums up to 40 sentences is expected to take around 15 minutes.

Participant Requirements
Participation in this study is limited to individuals age 18 and older, and native English speakers.

Risks
The risks and discomfort associated with participation in this study are no greater than those ordinarily encountered in daily life or during other online activities.

Benefits
There may be no personal benefit from your participation in the study but the knowledge received may be of value to humanity.

Compensation & Costs
For this task, you will receive between $2 to $3 for annotating each pair of articles. The exact reward for each pair depends on the length of corresponding articles. You will not be compensated if you provide annotations of poor quality.
There will be no cost to you if you participate in this study.

Future Use of Information and/or Bio-Specimens
In the future, once we have removed all identifiable information from your data (information or bio-specimens), we may use the data for our future research studies, or we may distribute the data to other researchers for their research studies. We would do this without getting additional informed consent from you (or your legally authorized representative). Sharing of data with other researchers will only be done in such a manner that you will not be identified.

Confidentiality
The data captured for the research does not include any personally identifiable information about you except your IP address and Mechanical Turk worker ID.
By participating in this research, you understand and agree that XXX may be required to disclose your consent form, data and other personally identifiable information as required by law, regulation, subpoena or court order. Otherwise, your confidentiality will be maintained in the following manner:

Table 2.8: Consent Form attached to each of our HITs. We anonymize the document for the conference review process.

Consent From (contd.)

Confidentiality (contd.)
Your data and consent form will be kept separate. Your consent form will be stored in a secure location on XXX property and will not be disclosed to third parties. By participating, you understand and agree that the data and information gathered during this study may be used by XXX and published and/or disclosed by XXX to others outside of XXX. However, your name, address, contact information and other direct personal identifiers will not be mentioned in any such publication or dissemination of the research data and/or results by XXX. Note that per regulation all research data must be kept for a minimum of 3 years.
The Federal government offices that oversee the protection of human subjects in research will have access to research records to ensure protection of research subjects.

Right to Ask Questions & Contact Information
If you have any questions about this study, you should feel free to ask them by contacting the Principal Investigator now at XXX, XXX, or by phone at XXX, or via email at XXX. If you have questions later, desire additional information, or wish to withdraw your participation please contact the Principal Investigator by mail, phone or e-mail in accordance with the contact information listed above.
If you have questions pertaining to your rights as a research participant; or to report concerns to this study, you should contact the XXX at XXX. Email: XXX. Phone: XXX or XXX.

Voluntary Participation
Your participation in this research is voluntary. You may discontinue participation at any time during the research activity. You may print a copy of this consent form for your records.

I am age 18 or older. □ Yes □ No

I have read and understand the information above. □ Yes □ No

I want to participate in this research and continue with the task. □ Yes □ No

Table 2.9: Consent Form attached to each of our HITs. We anonymize the document for the conference review process. (contd)

#	Text	Answer	Type
1	A 500lb bomb packed in the Cavalier is detonated with a remote trigger. The explosion tears through Market Street.	yes	Synonym
2	The death toll of the Omagh bomb blast in Northern Ireland has risen to 29 following the death of a man in hospital.	no	Member
3	Ahmed al-Mughassil was arrested in Beirut and transferred to Riyadh, the Saudi capital, according to the Saudi newspaper Asharq Alawsat. The Saudi Interior Ministry and Lebanese authorities had no immediate comment on the capture.	yes	Synonym
4	The blast didn't cause the destruction its planners intended. But it opened up a multi-story crater in the building, injured more than 1,000 people and ultimately killed six.	no	Member
5	March 4, 1998 - Four defendants, Salameh, Ayyad, Abouhalima, and Ajaj, are convicted. They are sentenced to prison terms of 240 years each. In 1998, the sentences were vacated. In 1999, the men were re-sentenced to terms of more than 100 years.	no	Unrelated
6	Perhaps the only early clues to emerge on an early quiet second day of the Boston Marathon bombing investigation - from the ATF and the FBI and the Boston police, from anonymous law enforcement officials and doctors pulling ball bearings out of victims limbs - concern the Boston bombs themselves. A similar scene played out in the Boston suburb of Newton, where a bomb used a robot to investigate a suspicious object that turned out to be a circuit board.	no	Member
7	As of Tuesday morning, jurors began reviewing evidence and witness testimony, which will play a role in helping them divide Dzhokhar Tsarnaev's guilt on each of the 30 charges he faces. A key issue for jurors - both in the guilt phase and later the penalty phase if Tsarnaev is convicted - will be whether the jurors see Tsarnaev as an equal partner with his old brother, Tamerian Tsarnaev, in the Boston Marathon bombing and the violent events that followed.	yes	Synonym
8	Though building the bomb was relatively easy, the experts say, it was not by any means free of danger. The bulkiest part of the bomb, they say, was extremely stable and could only have been touched off with a tremendous kick, like that provided by nitroglycerine. Making the nitroglycerin, blending some of the chemicals, was the trickiest part of the process.	yes	Synonym
9	An ongoing Somali offensive, backed by the U.S. and an African Union peacekeeping force has recaptured territory from al Shabaab in south-central Somalia, but has not eliminated al Shabaab's ability to conduct VBIED attacks. U.S.-backed Somali ground operations along with improved counter-VBIED capabilities among Somali forces may have slightly decreased VBIED attacks between November 2017 and January 2018.	yes	Synonym
10	According to the United Nations, more than 2.3 million Venezuelans have left their country in recent years. Increasingly they are leaving with no money and are traveling on foot across South American countries like Colombia, Ecuador and Peru, in dangerous journeys that can take several weeks.	no	Member

Table 2.10: Examples used with the qualification test on Mechanical Turk. For each paragraph with two highlighted events, we ask the question, “In the above paragraph, are the highlighted events the same?". The crowd worker has to select one of the “Yes" or “No" options.

#	Text	Answer	Type
11	Spain's King Juan Coarlos and Queen Sofia traveled to their summer residence in Majorca Saturday just two days after a bombing blamed on Basuqe separatists killed two policemen on the resort island.	no	Member
12	Yahoo Inc. is preparing to lay off between 600 and 700 workers in the latest shakeup triggered by the Internet company lackluster growth. Employees could be notified of the job cuts as early as Tuesday, according to a person familiar with Yahoo's plans.	yes	Synonym
13	A man shot and killed by police officers during a burglary here early Monday was identified by law enforcement authorities as the suspect in a string of five shooting deaths in South Carolina over the last 10 days. Sheriff Bill Blanton of Cherokee Country, S.C., where the killings took place, confirmed Monday evening that the authorities had been seeking the man killed in the burglary, Patrick T. Burris, a felon with a long record who had served seven years in prison and was paroled in April.	yes	Synonym
14	Staff Sgt. Robert Bales offered a tearful apology Thursday for gunning down 16 unarmed Afghan civilians inside their homes but said he still could not explain why he had carried out one of the worst U.S. war crimes in years. The unsworn statement from Bales, 40, came on the third day of hearing to determine whether he should ever be eligible for parole in the March 2012 Massacre.	yes	Synonym
15	In January two men were hanged after being convicted of involvement in protests, and in May, four Iranian Kurds and another man accused of terrorism were executed.	no	Unrelated
16	The Dow Corning Corporation filed for bankruptcy protection in a Federal court in Bay City, Michigan. Dow Corning said that seeking the protection of the bankruptcy court was the only way it could devise an enforceable plan to deal with the claims against it.	no	Realis
17	The UN report accused both Israel and Palestinian armed groups of commiting war crimes during the three-week war in Gaza that erupted on December 27, killing some 1,400 Palestinians and 13 Israelis.	no	Realis
18	A judge has ordered the surviving children of the Rev. Martin Luther King Jr. and Coretta Scott King to hold a shareholder's meeting to discuss their father's estate. The three siblings are the sole shareholders, directors and officers of a company that manages their father's intellectual property, but they have not met for an annual shareholder's meeting since 2004.	no	Realis
19	The first attack was a failure, but if the report is accurate, then it signals a dangerous new terror threat. The report showed pictures of the remains of a homemade attack drone.	no	Realis
20	A key issue for jurors - both in the guilt phase and later in the penalty phase if Tsarnaev is convicted - will be whether the jurors see Tsarnaev as an equal partner with his older brother, Tamerlan Tsarnaev, in the Boston Marathon bombing and the violent events that followed. Taken as a whole, the evidence suggests that the plan to bomb the Boston Marathon took shape over three months.	yes	Realis

Table 2.11: Examples used with the qualification test on Mechanical Turk. For each paragraph with two highlighted events, we ask the question, “In the above paragraph, are the highlighted events the same?". The crowd worker has to select one of the “Yes" or “No" options. (contd)

Screening Test

In this test, we ask you to identify whether two events (**highlighted** in each paragraph) indicate the same thing or not. Read each paragraph carefully and answer the question by selecting the appropriate option, *Yes* or *No*.

In total, you are presented with 8 questions and the time limit for this test is 20 minutes.

Note: It is important you do this test on your own because our HITs are similar to the questions presented in this test. For your reference, we provide five examples below,

He **died** of injuries from the accident. His friends were all saddened to hear his **death**.
Question: In the above paragraph, are the highlighted events the same?
Answer: Yes (both words, **died** and **death** indicate the person's death)

The suspect was **shot** and killed in the **raid** by the armed officers.
Question: In the above paragraph, are the highlighted events the same?
Answer: No (**shot** happened during the **raid**)

The couple had been planning to **go** to Paris for a long time. They finally **went** there last month.
Question: In the above paragraph, are the highlighted events the same?
Answer: Yes (The two events do not have to take place at the same time. Here, **go** would happen in the future, and **went** did occur.)

John **gave** a gift to Mary. Mary **received** a gift from John.
Question: In the above paragraph, are the highlighted events the same?
Answer: Yes (Same events described from different perspectives.)

Following the trial of Mahammed Alameh, the first suspect in the **bombing**, investigators discovered a jumble of chemicals, chemistry implements and detonating materials. The **explosion** killed at least five people.
Question: In the above paragraph, are the highlighted events the same?
Answer: No (One event is part of the other larger event. **bombing** refers to the entire process which starts with making a bomb and ends with destructions, damages and injuries, while **explosion** is a smaller event that occurs in that processes.)

Q1. ....
☐ Yes ☐ No
Q2. ....
☐ Yes ☐ No
...

Table 2.12: The template used in the qualification test to screen annotators. In addition to instructions and examples, we present eight yes/no questions.

Annotating Event Coreference in News Articles

In this HIT, you will be using our tool to perform the task. For a short tutorial on using our interface, see this 1 minute video: XXX. This HIT contains the following two steps,

- Visit the URL provided below to perform the task.
- At the end of the task, you will be provided a secret code. To submit this HIT, copy the secret code and paste it into the box provided below. Note that the secret code is unique for each task.

Link to the task: XXX

Fill in the secret code

Paste the secret code provided at the end of the task into the text box (*required)

Table 2.13: The template used for each Human Intelligence Task (HIT) on Mechanical Turk.

Place: Do you think the two events happen at the same place?

☐ Exactly the same ☐ The places overlap ☐ Not at all ☐ Cannot determine

Time: Do you think the two events happen at the same time?

☐ Exactly the same ☐ They overlap in time ☐ Not at all ☐ Cannot determine

Participants: Do you think the two events have the same participants?

☐ Exactly the same ☐ They share some participants ☐ Not at all ☐ Cannot determine

Inclusion: Do you think one of the events is part of the other?

☐ Yes, the left event is part of right one ☐ Yes, the right event is part of left one
☐ No, they are exactly the same ☐ Cannot determine

Table 2.14: Follow-up questions used for each annotated coreference link.

Chapter 3

Multilingual Event Linking to Wikidata

This work was published at the Multilingual Information Access workshop at NAACL 2022 (Pratapa et al., 2022).[1]

3.1 Introduction

Language grounding refers to linking concepts (e.g., events/entities) to a context (e.g., a knowledge base) (Chandu et al., 2021). Knowledge base (KB) grounding is a key component of information extraction stack and is well-studied for linking entity references to KBs like Wikipedia (Ji and Grishman, 2011). In this work, we present a new multilingual task that involves linking *event* references to Wikidata KB.[2]

Event linking differs from entity's as it involves taking into account the event participants as well as its temporal and spatial attributes. Nothman et al. (2012) defines event linking as connecting event references from news articles to a news archive consisting of first reports of the events. Similar to entities, event linking is typically restricted to prominent or report-worthy events. In this work, we use a subset of Wikidata as our event KB and link mentions from Wikipedia/Wikinews articles.[3] Figure 3.1 illustrates our event linking methodology.

Event linking is closely related to the more commonly studied task of cross-document event coreference (CDEC). The goal in CDEC is to understand the identity relationship between event mentions. This identity is often complicated by subevent and membership relations among events (Pratapa et al., 2021). Nothman et al. (2012) proposed event linking as an alternative to coreference that helps ground report-worthy events to a KB. They showed linking helps avoid the traditional bottlenecks seen with the event coreference task. We postulate *linking to be a complementary task to coreference*, where the first mention of an event in a document is typically linked or grounded to the KB and its relationship with the rest of the mentions from the document is captured via coreference. Additionally, due to computational constraints, coreference resolution is often restricted to a small batch of documents. Grounding, however, can be performed efficiently using dense retrieval methods (Wu et al., 2020) and is scalable to any large

[1] Code and data are available at: `https://github.com/adithya7/xlel-wd`.

[2] `www.wikidata.org`

[3] We define *mention* as the textual expression that refers to an *event* from the KB.

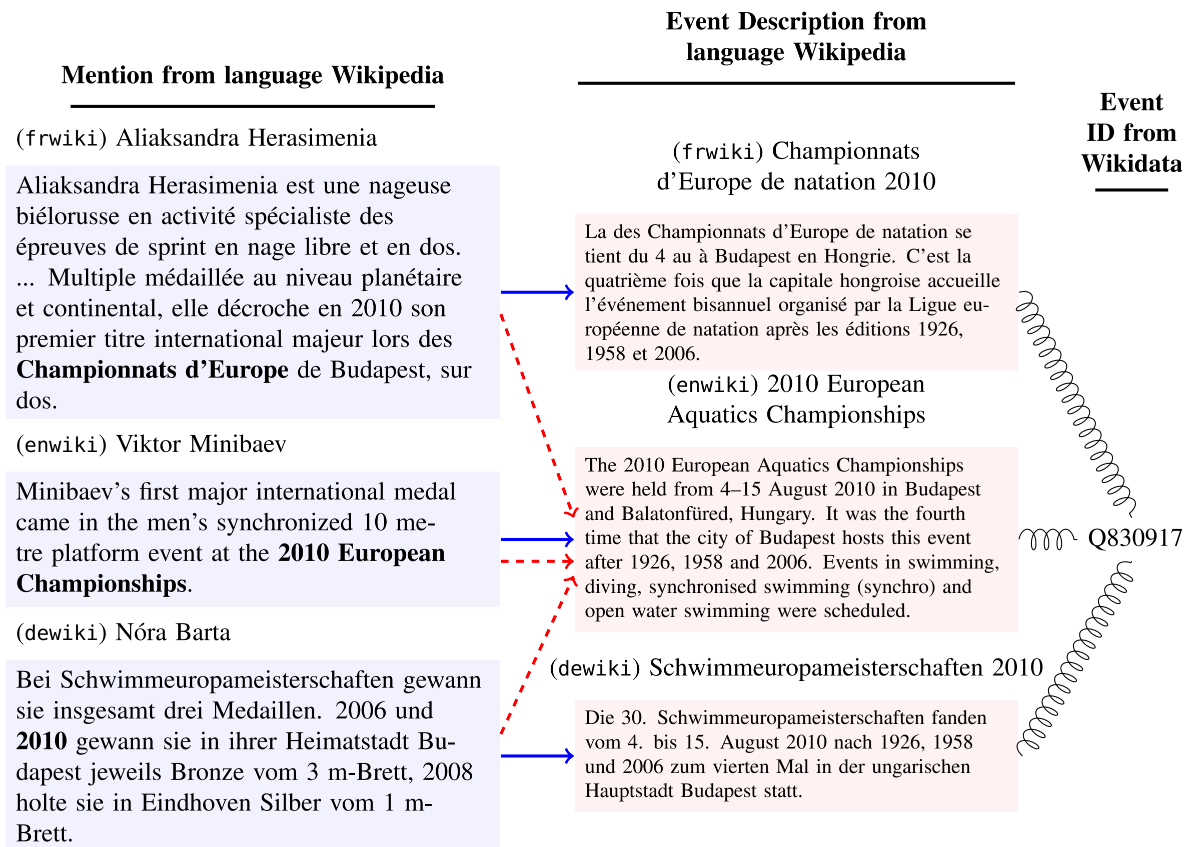


Figure 3.1: An illustration of multilingual event linking with Wikidata as our interlingua. Mentions from French, English and German Wikipedia (column 1) are linked to the same event from Wikidata (column 3). The title and descriptions for the event Q830917 are compiled from the corresponding language Wikipedias (column 2). The solid blue arrows (⟶) presents our multilingual task, to link `lgwiki` mention to event using `lgwiki` description. The dashed red arrows (⇢) showcases the crosslingual task, to link `lgwiki` mention to event using `enwiki` description.

multi-document corpora.

Grounding event references to a KB has many downstream applications. First, event identity encompasses multiple aspects such as spatio-temporal context and participants. These aspects typically spread across many documents, and KB grounding helps construct a shared global account for each event. Second, grounding is a complementary task to coreference. In contrast to coreference, event grounding formulated as the nearest neighbor search leads to efficient scaling.

For the event linking task, we present a new multilingual dataset that grounds mentions from multilingual Wikipedia/Wikinews articles to the corresponding event in Wikidata. Figure 3.1 presents an example from our dataset that links mentions from three languages to the same Wikidata item. To construct this dataset, we make use of the hyperlinks in Wikipedia/Wikinews articles. These links connect anchor texts (like '2010 European Championships' or "Championnats d'Europe") in context to the corresponding event Wikipedia page ('2010 European Aquatics Championships' or "Championnats d'Europe de natation 2010"). We further connect the event

Wikipedia page to its Wikidata item ('Q830917'), facilitating multilingual grounding of mentions to KB events. We use the title and first paragraph from the language Wikipedia pages as our event descriptions (column 2 in Figure 3.1).

Such hyperlinks have previously been explored for named entity disambiguation (Eshel et al., 2017), entity linking (Logan et al., 2019) and cross-document coreference of events (Eirew et al., 2021) and entities (Singh et al., 2012). Our work is closely related to the English CDEC work of Eirew et al. (2021), but we view the task as linking instead of coreference. This is primarily due to the fact that most hyperlinked event mentions are prominent and typically cover a broad range of subevents, conflicting directly with the notion of coreference. Additionally, our dataset is multilingual, covering 44 languages, with Wikidata serving as our *interlingua*. Botha et al. (2020) is a related work from entity linking literature that covers entity references from multilingual Wikinews articles to Wikidata.

We use the proposed dataset to develop multilingual event linking systems. We present two variants to the linking task, multilingual and crosslingual. In the multilingual task, mentions from individual language Wikipedia are linked to the events from Wikidata with descriptions taken from the same language (see solid blue arrows (⟶) in Figure 3.1). The crosslingual task requires systems to use English event description irrespective of the mention language (see dashed red arrows (⇢) in Figure 3.1). In both tasks, the end goal is to identify the Wikidata ID (e.g. Q830917). Following prior work on entity linking (Logeswaran et al., 2019), we adopt a *zero-shot* approach in all of our experiments. We present results using a retrieve+rank approach based on Wu et al. (2020) that utilizes BERT-based biencoder and crossencoder for our multilingual event linking task. We experiment with two multilingual encoders, mBERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020) and we find biencoder and crossencoder significantly outperform a tf-idf-based baseline, BM25+ (Lv and Zhai, 2011a). Our results indicate the crosslingual task is more challenging than the multilingual task, possibly due to differences in typology of source and target languages. Our key contributions are,

- We propose a new multilingual NLP task that involves linking multilingual text mentions to a knowledge base of events.
- We release a large-scale dataset for the zero-shot multilingual event linking task by compiling mentions from Wikipedia and their grounding to Wikidata. Our dataset captures 1.8M mentions across 44 languages refering to over 10K events. To test out-of-domain generalization, we additionally create a small Wikinews-based evaluation set.
- We present two evaluation setups, multilingual and crosslingual event linking. We show competitive results across languages using a retrieve and rank methodology.

3.2 Related Work

Our focus task of multilingual event linking shares resemblance with entity/event linking, entity/event coreference and other multilingual NLP tasks.

3.2.1 Entity Linking

Our work utilizes hyperlinks between Wikipedia pages to identify event references. This idea was previously explored in multiple entity related works, both for dataset creation (Mihalcea and Csomai, 2007; Botha et al., 2020) and data augmentation during training (Bunescu and Paşca, 2006; Nothman et al., 2008). Another related line of work utilized hyperlinks from general web pages to Wikipedia articles for the tasks of cross-document entity coreference (Singh et al., 2012) and named entity disambiguation (Eshel et al., 2017). Sil et al. (2012); Logeswaran et al. (2019) highlighted the need for zero-shot evaluation. We adopt this standard by using a disjoint sets of events for training and evaluation (see subsection 3.3.2).

3.2.2 Event Linking

Event linking is important for downstream tasks like narrative understanding. For instance, consider a prominent event like ‘2020 Summer Olympics’. This event has had a large influx of articles in multiple languages. It is often useful to ground the references to specific prominent subevents in KB. Some examples of such events from Wikidata are “Swimming at the 2020 Summer Olympics – Women’s 100 metre freestyle” (Q64513990) and “Swimming at the 2020 Summer Olympics – Men’s 100 metre backstroke” (Q64514005). Event linking task while important is albeit less explored. Nothman et al. (2012) linked event-referring expressions from news articles to a news archive. These links are made to the first-reported news article regarding the event. In contrast, we focus on prominent events that have a corresponding Wikidata item. Concurrent to our work, Yu et al. (2023) presents a dataset for linking event mentions to Wikipedia. Similar to our work, they utilize hyperlinks within Wikipedia pages but are restricted to only English. They also create a newswire based evaluation set from NYTimes articles. In contrast, our work utilizes events from Wikidata and covers a larger set of languages. While our work also includes a newswire based evaluation set from Wikinews, it does not explicitly target verb mentions.

3.2.3 Event Coreference

Event coreference resolution is closely related to event grounding but assumes a stricter notion of identity between mentions (Nothman et al., 2012). Multiple cross-document coreference resolution works made use of Wikipedia (Eirew et al., 2021) and Wikinews (Minard et al., 2016; Pratapa et al., 2021) for dataset collection. Minard et al. (2016) obtained human translations of English Wikinews articles to create a crosslingual event coreference dataset. In contrast, our dataset uses the original multilingual event descriptions written by language Wikipedia contributors (column 2 in Figure 3.1).

3.2.4 Multilingual Tasks

A majority of the existing NLP datasets (/systems) cater to a fraction of world languages (Joshi et al., 2020). There is a growing effort on creating more multilingual benchmarks for tasks like natural language inference (XNLI; Conneau et al. (2018)), question answering (TyDi-QA; Clark

	Train	Dev	Test	Total
Events	8653	1090	1204	10947
Event Sequences	6758	844	846	8448
Mentions	1.44M	165K	190K	1.8M
Languages	44	44	44	44

Table 3.1: Dataset Summary

et al. (2020), XOR QA; Asai et al. (2021)), linking (Mewsli-9; Botha et al. (2020)) as well as comprehensive evaluations (XTREME-R; Ruder et al. (2021)). To the best of our knowledge, our work presents the first benchmark for multilingual event linking.

3.3 Multilingual Event Linking Dataset

Our data collection methodology is closely related to the zero shot entity linking work of Botha et al. (2020) but we take a top-down approach starting from Wikidata. Eirew et al. (2021) identified event pages from English Wikipedia by processing the infobox elements. However, we found relying on Wikidata for event identification to be more robust. Additionally, Wikidata serves as our *interlingua* that connects mentions from numerous languages.

3.3.1 Dataset Compilation

To compile our dataset, we follow a three-stage pipeline, 1) identify Wikidata items that correspond to events, 2) for each Wikidata event, collect links to language Wikipedia articles and 3) iterate through all the language Wikipedia dumps to collect mention spans that refer to these events.

Wikidata Event Identification: Events are typically associated with time, location and participants, distinguishing them from entities. To identify events from the large pool of Wikidata (WD) items, we make use of the properties listed on WD.[4] Specifically, we consider a WD item to be a candidate event if it contains the following two properties, temporal[5] and spatial[6]. We perform additional postprocessing on this candidate event set to remove non-events like empires (Roman Empire: Q2277), missions (Surveyor 7: Q774594), TV series (Deception: Q30180283) and historic places (French North Africa: Q352061).[7] Each event in our final set has caused a state change and is grounded in a spatio-temporal context. This distinguishes our set of events from the rest of the items from Wikidata. Following the terminology from Weischedel et al. (2013), these KB events can be characterized as *eventive nouns*.

[4] https://www.wikidata.org/wiki/Wikidata:List_of_properties

[5] duration OR point-in-time OR (start-time AND end-time)

[6] location OR coordinate-location

[7] see Table 3.8 in subsection 3.7.2 of Appendix for the full list of exclusion properties.

2016 Summer Olympics (Q8613)

↑ part-of

athletics at the 2016 Summer Olympics (Q18193712)

↑ part-of

athletics at the 2016 Summer Olympics–men's 100 metres (Q25397537)

Figure 3.2: An illustration of event hierarchy in Wikidata.

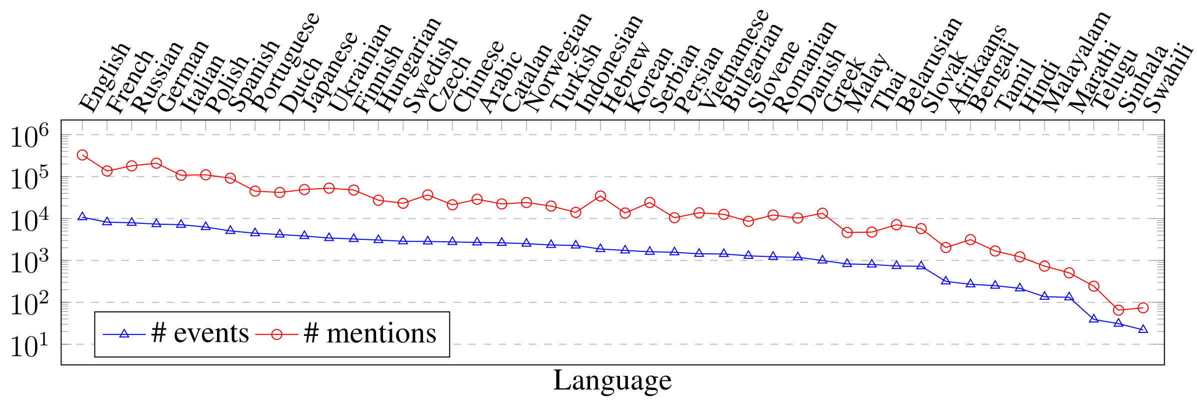


Figure 3.3: Statistics of events and mentions per language in the proposed dataset. The languages are sorted in the decreasing order of # events. The counts on y-axis are presented in log scale.

A Note on WD Hierarchy: WD is a rich structured KB and we observed many instances of hierarchical relationship between our candidate events. See Figure 3.2 for an example. While this hierarchy adds an interesting challenge to the event grounding task, we observed multiple instances of inconsistency in links. Specifically, we observed references to parent item (Q18193712) even though the child item (Q25397537) was the most appropriate link in context. Therefore, in our dataset, we only include *leaf nodes* as our candidate event set (e.g. Q25397537). This allows us to focus on most atomic events from Wikidata. Expanding the label set to include the hierarchy is an interesting direction for future work.

Wikidata ⟿ Wikipedia: WD items have pointers to the corresponding language Wikipedia articles.[8] We make use of these pointers to identify Wikipedia articles describing our candidate WD events. Figure 3.1 illustrates this through the coiled pointers (⟿) for the three languages. We make use of the event's Wikipedia article title and its first paragraph as the description for the WD event. Each language version of a Wikipedia article is typically written by independent contributors, so the event descriptions vary across languages.

[8] https://meta.wikimedia.org/wiki/List_of_Wikipedias

Mention Identification: Wikipedia articles are often connected through hyperlinks. We iterate through each language Wikipedia and collect anchor texts of hyperlinks to the event Wikipedia pages (column 1 in Figure 3.1). We retain both the anchor text and the surrounding paragraph (context). Notably, the anchor text can occasionally be a temporal expression or location relevant to the event. In the German mention from Figure 3.1, the anchor text '2010' links to the event Q830917 (2010 European Aquatics Championships). This event link can be infered by using the context ('Schwimmeuropameisterschaften': European Aquatics Championships). In fact, the neighboring span '2006' refers to a different event from Wikidata (Q612454: 2006 European Aquatics Championships). We use the September 2021 XML dumps of language Wikipedias and the October 2021 dump of Wikidata. We use Wikiextractor tool (Attardi, 2015) to extract text content from the Wikipedia dumps. We retain the hyperlinks in article texts for use in mention identification. Overall, the mentions in our datasets can be categorized into the following types, 1) eventive noun (like the KB event), 2) verbal, 3) location and 4) temporal expression. Such a diversity in the nature of mentions also differentiates the event linking task from the standard named entity linking or disambiguation.

Postprocessing: To link a mention to its event, the context should contain the necessary temporal information. For instance, its important to be able to differentiate between links to '2010 European Aquatics Championships' vs '2012 European Aquatics Championships'. Therefore, we heuristically remove mention (+context) if it completely misses the temporal expressions from the corresponding language Wikipedia title and description. Additionally, we also remove mentions if their contexts are either too short or too long (<100, >2000 characters). We also prune WD events under the following conditions: 1) only contains mentions from a single language, 2) >50% of the mentions match their corresponding language Wikipedia title (i.e., low diversity), 3) very few mentions (<30). Table 3.1 presents the overall statistics of our dataset. The full list of languages with their event and mention counts are presented in Figure 3.3. Each WD event on average has mention references from 9 languages indicating the highly multilingual nature of our dataset. See Table 3.9 in Appendix for details on the geneological information for the chosen languages. We chose our final set of languages by maximizing for the diversity in language typology, language resources (in event-related tasks and general) and the availability of content on Wikipedia. Wikipedia texts and Wikidata KB are available under CC BY-SA 3.0 and CC0 1.0 license respectively. We will release our dataset under CC BY-SA 3.0.

Wikinews ⇢ Wikidata: To test the out-of-domain generalization, we additionally prepare a small evaluation set based on Wikinews articles.[9] Inspired by prior work on multilingual entity linking (Botha et al., 2020), we collect hyperlinks from event mentions in multilingual Wikinews articles to Wikidata. We restrict the set of events to the previously identified 10.9k events from Wikidata (Table 3.1). We again use Wikiextractor tool to collect raw texts from March 2022 dumps of all language Wikinews. We identify hyperlinks to Wikipedia pages or Wikinews categories that describe the events from Wikidata.

Table 3.2 presents the overall statistics of our Wikinews-based evaluation set. This set is much smaller in size compared to Wikipedia-based dataset primarily due to significantly smaller

[9] https://www.wikinews.org

	Cross-domain	Zero-shot
Events	802	149
Mentions	2562	437
Languages	27	21

Table 3.2: Summary of Wikinews-based evaluation set. We present two evaluation settings, cross-domain and zero-shot. Zero-shot evaluation set is a subset of cross-domain set as it only includes events from dev and test splits of Wikipedia-based evaluation set (Table 3.1).

footprint of Wikinews.[10] Following the taxonomy from Logeswaran et al. (2019), we present two evaluation settings, cross-domain and zero-shot. Cross-domain evaluation gauges model generalization to unseen domains (newswire). Zero-shot evaluation tests on unseen domain and unseen events.[11]

Unlike Wikipedia, Wikinews articles contains meta information such as news article title and publication date that help provide broader context for the document. In section 3.5, we perform ablations studies to see the impact of this meta information.

Mention Distribution: Following the categories from Logeswaran et al. (2019), we compute mention distributions in the following four buckets, 1) high overlap: mention span is the same as the event title, 2) multiple categories: event title includes an additional disambiguation phrase, 3) ambiguous substring: mention span is a substring of the event title, and 4) low overlap: all other cases. For the Wikipedia-based dataset, the category distribution is 22%, 6%, 14%, and 58%.[12] For the Wikinews-based dataset, the category distribution is 18%, 4%, 6%, and 72%. We also computed the fraction of mentions that are temporal expressions. We used HeidelTime library (Strötgen and Gertz, 2015) for 25 languages and found 6% of the spans in the dev set are temporal expressions.

3.3.2 Task Definition

Given a mention and a pool of events from a KB, the task is to identify the mention's reference in the KB. For instance, the three mentions from column 1 in Figure 3.1 are to be linked to the Wikidata event, Q830917. Following Logeswaran et al. (2019), we assume an in-KB evaluation approach, therefore, every mention refers to a valid event from the KB (Wikidata). We collect descriptions for the Wikidata events from all the corresponding language Wikipedias. The article title and the first paragraph constitute the event description. This results in multilingual descriptions for each event (column 2 in Figure 3.1). We propose two variants of the event linking task, *multilingual* and *crosslingual*, depending on the source and target languages. We define the input mention and event description as source and target respectively. The event label itself (e.g. Q830917) is language-agnostic.

[10] For comparison, English Wikinews contains 21K articles while English Wikipedia contains 6.5M pages.

[11] we consider dev and test events from Table 3.1 as unseen.

[12] The disambiguation phrase is typically a suffix in the title for English (Logeswaran et al., 2019), but in our multilingual setting, it can be anywhere in the title.

Multilingual Event Linking: Given a mention from language $\mathcal{L}$, the linker searches through the event candidates from the same language $\mathcal{L}$ to identify the correct link. The source and target language are the same in this task. The size of event candidate pool varies across languages (Figure 3.3), thereby varying the task difficulty.

Crosslingual Event Linking: Given a mention from any language $\mathcal{L}$, the linker searches the entire pool of event candidates to identify the link. Here, we restrict the target language to English, requiring the linker to only make use of the English descriptions for candidate events. Note that, all the events in our dataset have English descriptions.

Creating Splits: The train, dev and test distributions are presented in Table 3.1. The two tasks, multilingual and crosslingual share the same splits except for the difference in target language descriptions. Following the standard in entity linking literature, we focus on the zero-shot linking, that requires the evaluation and train events to be completely disjoint. Due to prevalence of event sequences in Wikidata, a simple random split is not sufficient.[13] We add an additional constraint that event sequences are disjoint between splits. Systems need to perform temporal and spatial reasoning to distinguish between events within a sequence, making the task more challenging.

3.4 Modeling

In this section, we present our systems for multilingual and crosslingual event linking to Wikidata. We follow the entity linking system BLINK (Wu et al., 2020) to adapt a retrieve and rank approach. Given a mention, we first use a BERT-based biencoder to retrieve top-k events from the candidate pool. Then, we use a crossencoder to rerank these top-k candidates and identify the best event label. Additionally, following the baselines from entity linking literature, we also experiment with BM25 as a candidate retrieval method.

3.4.1 BM25

BM25 is a commonly used tf-idf based ranking function and a competitive baseline for entity linking. We explore three variants of BM25, BM25Okapi (Robertson et al., 1994), BM25+ (Lv and Zhai, 2011a) and BM25L (Lv and Zhai, 2011b). We use the implementation of Brown (2020) with mention as query and event description as documents.[14] Since BM25 is a bag-of-words method, we only use in the multilingual task. To create the documents, we use the concatenation of title and description of events. For the query, we experiment with increasing context window sizes of 8, 16, 32, 64 and 128 along with a mention-only baseline.

[13] 2008, 2010, 2012 iterations of Aquatics Championships from Figure 3.1

[14] To tokenize text across the 44 languages, we used bert-base-multilingual-uncased tokenizer from Huggingface.

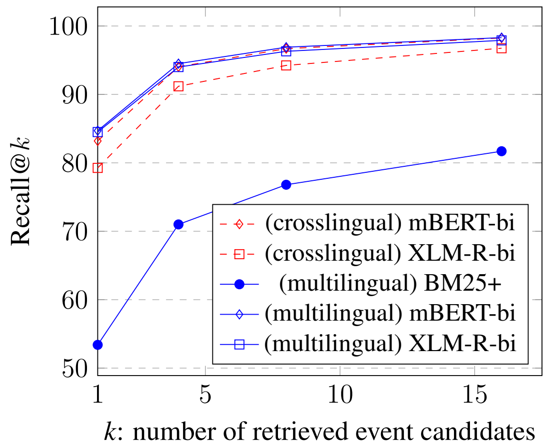


Figure 3.4: Retrieval performance on dev split.

Model	Multilingual Dev	Multilingual Test	Crosslingual Dev	Crosslingual Test
BM25+	53.4	50.1	–	–
mBERT-bi	84.7	84.6	83.2	**83.9**
XLM-R-bi	84.5	84.3	79.3	79.1
mBERT-cross	89.8	**89.3**	81.3	73.9
XLM-R-cross	88.8	87.3	81.0	75.6

Table 3.3: Event Linking Accuracy. For biencoder models, we report Recall@1.

3.4.2 Retrieve+Rank

We adapt the standard entity linking architecture (Wu et al., 2020) to the event linking task. This is a two-stage pipeline, a retriever (biencoder) and a ranker (crossencoder).

Biencoder: Using two multilingual transformers, we independently encode the context and event candidates. The input context is constructed as `[CLS]` left context `[MENTION_START]` mention `[MENTION_END]` right context `[SEP]`. Candidate events use a concatenation of event's title and description, `[CLS]` title `[EVT]` description `[SEP]`. In both cases, we use the final layer `[CLS]` token representation as our embedding. For each context, we score the event candidates by taking a dot product between the two embeddings. We follow prior work (Lerer et al., 2019; Wu et al., 2020) to make use of in-batch random negatives during training. At inference, we run a nearest neighbour search to find the top-k candidates.

Crossencoder: In our crossencoder, the input constitutes a concatenation of the context and a given event candidate.[15] We take the `[CLS]` token embedding from last layer and pass it through a classification layer. We run crossencoder training only on the top-k event candidates retrieved

[15] `[CLS]` left context `[MENTION_START]` mention `[MENTION_END]` right context `[SEP]` title `[EVT]` description `[SEP]`

Model	Multilingual		Crosslingual	
	CD	ZS	CD	ZS
BM25+	53.5	58.6	-	-
mBERT-bi	81.2	76.7	85.4	78.0
XLM-R-bi	82.2	76.7	82.6	76.4
mBERT-cross	90.1	84.4	89.3	76.2
XLM-R-cross	89.7	84.4	88.9	76.0

Table 3.4: Event linking accuracy on Wikinews test set. CD and ZS indicate cross-domain and zero-shot.

Mention Context: At the 2000 Summer Olympics in Sydney, Sitnikov competed only in two swimming events. ... Three days later, in the **100 m freestyle**, Sitnikov placed fifty-third on the morning prelims. ...
Predicted Label: Swimming at the 2008 Summer Olympics – Men's 100 metre freestyle
Gold Label: Swimming at the 2000 Summer Olympics – Men's 100 metre freestyle

Mention Context: ... war er bei der Oscarverleihung 1935 erstmals für einen Oscar für den besten animierten Kurzfilm nominiert. Eine weitere Nominierung in dieser Kategorie erhielt er **1938** für "The Little Match Girl" (1937).
Predicted Label: The 9th Academy Awards were held on March 4, 1937, ...
Gold Label: The 10th Academy Awards were originally scheduled ... but due to ... were held on March 10, 1938, ..

Mention Context: Ivanova won the silver medal at the 1978 World Junior Championships. She made her senior World debut at the **1979 World Championships**, finishing 18th. Ivanova was 16th at the 1980 Winter Olympics.
Predicted Label: FIBT World Championships 1979
Gold Label: 1979 World Figure Skating Championships

Mention Context: ...攝津號與其姐妹艦河號於1914年10月至11月間參與了青島戰役的最後階段...
Predicted Label: Battle of the Yellow Sea
Gold Label (English): Siege of Tsingtao: The siege of Tsingtao (or Tsingtau) was the attack on the German port of Tsingtao (now Qingdao) ...
Gold Label (Chinese): 青島戰役（，）是第一次世界大戰初期日本進攻國膠州灣殖民地及其首府青島的一場戰役，也是唯一的一場戰役。

Table 3.5: Examples of errors by the event linking system.

by the biencoder. During training, we optimize a softmax loss to predict the gold event candidate within the retrieved top-k. For inference, we predict the highest scoring context-candidate tuple from the top-k candidates. We experiment with two multilingual encoders, mBERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020), we refer to the bi- and cross-encoder configurations as mBERT-bi, XLM-RoBERTa-bi and mBERT-cross, XLM-RoBERTa-cross. For crossencoder training and inference, we use the retrieval results from the same BERT-based biencoder.[16]

3.5 Evaluation

We present our results on the development and test splits of the proposed dataset. In our experiments, we use bert-base-multilingual-uncased and xlm-roberta-base from Huggingface trans-

[16]see section 3.7.3 in Appendix for other details.

formers (Wolf et al., 2020). For the multilingual task, even though the candidate set is partly different between languages, we share the model weights across languages. We believe this weight sharing helps in improving the performance on low-resource languages (Arivazhagan et al., 2019). We follow the standard metrics from prior work on entity linking, both for retrieval and reranking. **Recall@*k*** measures fraction of contexts where the gold event is contained in the top-k retrieved candidates. **Accuracy** measures fraction of contexts where the predicted event candidate matches the gold candidate. We use the unnormalized accuracy score from Logeswaran et al. (2019) that evaluates the overall end-to-end performance (retrieve+rank).

3.5.1 Results

Figure 3.4 presents the retrieval results on dev split for both multilingual and crosslingual tasks. The biencoder models significantly outperform the best BM25 configuration, BM25+ (with a context window of 16).[17] The performance is mostly similar for k=8 and k=16 for both biencoder models, therefore, we select k=8 for our crossencoder experiments.[18] Table 3.3 presents the accuracy scores for the crossencoder models and R@1 scores for retrieval methods. On the multilingual task, mBERT crossencoder model performs the best and significantly better than the corresponding biencoder model. However, on the crosslingual task, mBERT biencoder performs the best. As expected, the crosslingual task is more challenging than the multilingual task. Due to the large number of model parameters, all of our reported results were based on a single training run.

We also measure the cross-domain and zero-shot performance of these systems on the proposed Wikinews evaluation set (section 3.3.1). As seen in Table 3.4, we notice good cross-domain but moderate zero-shot transfer. This highlights that unseen events from unseen domains present a considerable challenge. We noticed further gains (4-12%) when the meta information (date and title) is included with the context. Our ablation studies showed that this gain is primarily due to article date.[19]

3.5.2 Analysis

Performance by Language: Multilingual and crosslingual tasks have three major differences: 1) source & target language, 2) language-specific descriptions can be more informative than English descriptions, and 3) candidate pool varies language (see Figure 3.3). While the performance is largely the same across languages, we noticed slightly lower crosslingual performance, especially for medium and low-resource languages.[20]

We also perform qualitative analysis of errors made by our mBERT-based biencoder models on multilingual and crosslingual tasks. We summarize our observations from this analysis below,

Temporal Reasoning: The event linker occasionally performs insufficient temporal reasoning in the context (see example 1 in Table 3.5). Since our dataset contains numerous event sequences,

[17] For a detailed comparison of various configurations of BM25 baseline, refer to Figure 3.5 in Appendix.
[18] see Table 3.6 in Appendix for Recall@8 scores for all the configurations.
[19] see section 3.7.3 in Appendix for full results.
[20] see Figure 3.8 and Figure 3.9 in Appendix

such temporal reasoning is often important.

Temporal and Spatial expressions: In cases where the anchor text is a temporal or spatial expression, we found the system sometimes struggle to link to the event even if the link can be infered given the context information (see example 2 in Table 3.5). We believe these examples will also serve as interesting challenge for future work on our dataset.

Event Descriptions: Crosslingual system occasionally struggles with the English description. In example 4 from Table 3.5, we notice the mention matches exactly with the language Wikipedia title but it struggles with English description. Therefore, depending on the event, we hypothesize that language-specific event descriptions can sometimes be more informative than the English description.

Dataset Errors: We found instances where the context doesn't provide sufficient information needed for grounding (see example 3 in Table 3.5). Albeit uncommon, we found a few cases where the human annotated hyperlinks in Wikipedia can sometimes be incorrect.[21]

3.5.3 Discussion

Retrieve+rank based methods have been effective for entity linking tasks (Wu et al., 2020; Botha et al., 2020). Our results indicate that the same retrieve+rank approach is useful for the task of event linking. However, our zero-shot results on Wikinews hint toward potential challenges in adapting to new domains. Additionally, as described above, event linking presents added challenges in dealing with temporal/spatial expressions and temporal reasoning. For further analysis, it would be interesting to contrast the performance differences between planned (e.g., sports competitions) and unplanned (e.g., wars) events.

3.6 Conclusion & Future Work

We present the task of multilingual event linking to Wikidata. To support this task, we first compile a dictionary of events from Wikidata using temporal and spatial properties. We prepare descriptions for these events from multilingual Wikipedia pages. We then identify a large collection of inlinks from various language Wikipedia. Depending on the language of event description, we present two variants of the task, multilingual (`lg`⟶`lg`) and crosslingual (`lg`⇢`en`). Furthermore, to test cross-domain generalization we create a small evaluation set based on Wikinews articles. Our results using a retrieve+rank approach indicate that the crosslingual task is more challenging than the multilingual.

Event linking task has multiple interesting future directions. First, the Wikidata-based event dictionary can be expanded to include hierarchical event structures (Figure 3.2). Since events are inherently hierarchical, this will present a more realistic challenge for the linking systems. Second, mention coverage of our dataset can be expanded to include more verbal events. Third,

[21] For more detailed examples, refer to Table 3.10, Table 3.12 and Table 3.13 in Appendix.

event linking systems can be improved with better temporal reasoning and improved handling of temporal and spatial expressions. Fourth, the Wikidata-based event dictionary can be expanded to include events that do not contain any English Wikipedia descriptions.

3.7 Appendix

3.7.1 Ethical Considerations

In this work, we presented a new dataset compiled automatically from Wikipedia, Wikinews and Wikidata. After the initial collection process, we perform rigorous post-processing steps to reduce potential errors in our dataset. Our dataset is multilingual with texts from 44 languages. In our main paper, we state these languages as well as their individual representation in our dataset. As we highlight in the paper, the proposed linking systems only work for specific class of events (eventive nouns) due to the nature of our dataset.

3.7.2 Dataset

After identifying potential events from Wikidata, we perform additional post-processing to remove any non-event items. Table 3.8 presents the list of all Wikidata properties used for removing non-event items from our corpus. Table 3.9 lists all languages from our dataset along with their language genealogy and distribution in the dataset.

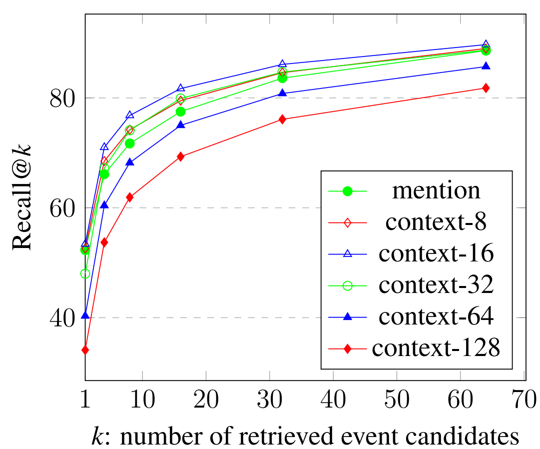


Figure 3.5: Effect of context window size on BM25+ retrieval performance.

3.7.3 Modeling

Experiments: We use the base versions of mBERT and XLM-RoBERTa in all of our experiments. In the biencoder model, we use two multilingual encoders, one each for context and candidate encoding. In crossencoder, we use just one multilingual encoder and a classification layer. In all of our experiments, we optimize all the encoder layers. For biencoder training, we

Retriever	Multilingual		Crosslingual	
	Dev	Test	Dev	Test
BM25+	76.8	70.5	–	–
mBERT-bi	96.9	**97.1**	96.7	**97.2**
XLM-R-bi	96.3	96.7	94.2	95.3

Table 3.6: Event candidate retrieval results, Recall@8.

use AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of 1e-05 and a linear warmup schedule. We restrict the context and candidate lengths to 128 sub-tokens and select the best epoch (of 5) on the development set. For crossencoder training, we also use AdamW optimizer with a learning rate of 2e-05 and a linear warmup schedule. We restrict the overall sequence length to 256 sub-tokens and select the best epoch (of 5) on the development set. We ran our experiments on a mix of GPUs, TITANX, v100, A6000 and a100. Each training and inference runs were run on a single GPU. Both biencoder and crossencoder were run for 5 epochs and we select the best set of hyperparameters based on the dev set performance. On a single a100 GPU, biencoder training takes about 1.5hrs per epoch and the crossencoder takes $\sim$20hrs per epoch (with k=8).

Results: In Figure 3.5, we present results on the development set from all the explored configurations. In Table 3.6, we show the Recall@8 scores from all the retrieval models. Based on the performance on development set, we selected k=8 for our crossencoder training and inference. We also report the test scores for completeness. Figure 3.6 presents the retrieval recall scores. Figure 3.7 presents the retrieval recall scores for BM25+ (context length 16) method. Figure 3.9 presents a detailed comparison of per-language accuracies between multilingual and crosslingual tasks for each configuration.

Wikinews: Each Wikinews article contains meta information such as article title and publication date. Since this meta information provide additional context to the linker, we experimented by including this meta information along with the mention context. The meta information is encoded with the context as "`[CLS]` title `[SEP]` date `[SEP]` left context `[MENTION_START]` mention `[MENTION_END]` right context `[SEP]`". Table 3.7 presents the detailed results on the Wikinews evaluation set.

Examples: We also present full examples of system errors we identified through a qualitative analysis. Table 3.10 presents examples of system errors due to insufficient temporal reasoning in the context. Table 3.11 presents examples of system errors on mentions that are temporal or spatial expressions. Table 3.12 presents examples of system errors on crosslingual task due to issues related with tackling non-English mentions. Table 3.13 presents examples of system errors that were caused due to dataset errors.

Model	Multilingual				Crosslingual			
	Ctxt	Ctxt+date	Ctxt+title	Ctxt+date+title	Ctxt	Ctxt+date	Ctxt+title	Ctxt+date+title
cross-domain								
mBERT-bi	81.2	87.4	83.4	87.7	85.4	90.0	87.4	90.6
XLM-R-bi	82.2	89.4	85.1	90.8	82.6	88.8	85.3	90.0
mBERT-cross	90.1	95.0	91.5	95.6	89.3	93.5	90.8	93.8
XLM-R-cross	89.7	94.0	91.6	94.7	88.9	93.6	90.6	93.7
zero-shot								
mBERT-bi	76.7	86.3	78.0	86.7	78.0	85.6	80.3	87.4
XLM-R-bi	76.7	86.0	80.1	89.0	76.4	85.8	78.7	87.2
mBERT-cross	84.4	92.2	86.5	93.8	76.2	81.7	77.6	81.5
XLM-R-cross	84.4	90.6	84.9	92.2	76.0	84.2	76.4	83.5

Table 3.7: Event linking accuracy on Wikinews test set. For each configuration, we report results using just the mention context (Ctxt), mention context + article publication date (Ctxt+date), mention context + article title (Ctxt+title) and mention context + article date & title (Ctxt+date+title). Most of the gain comes from including the date across all model configurations and tasks.

Property	Property_Label	URI	URI_Label
P31	instance_of	Q48349	empire
P31	instance_of	Q11514315	historical_period
P31	instance_of	Q3024240	historical_country
P31	instance_of	Q11042	culture
P31	instance_of	Q28171280	ancient_civilization
P31	instance_of	Q1620908	historical_region
P31	instance_of	Q3502482	cultural_region
P31	instance_of	Q465299	archaeological_culture
P31	instance_of	Q568683	age
P31	instance_of	Q763288	lander
P31	instance_of	Q4830453	business
P31	instance_of	Q24862	short_film
P31	instance_of	Q1496967	territorial_entity
P31	instance_of	Q68	computer
P31	instance_of	Q486972	human_settlement
P31	instance_of	Q26529	space_probe
P31	instance_of	Q82794	geographic_region
P31	instance_of	Q43229	organization
P31	instance_of	Q15401633	archaeological_period
P31	instance_of	Q5398426	television_series
P31	instance_of	Q24869	feature_film
P31	instance_of	Q11424	film
P31	instance_of	Q718893	theater
P31	instance_of	Q1555508	radio_program
P31	instance_of	Q17343829	unincorporated_community_in_the_United_States
P31	instance_of	Q254832	Internationale_Bauausstellung
P31	instance_of	Q214609	material
P31	instance_of	Q625298	peace_treaty
P31	instance_of	Q131569	treaty
P31	instance_of	Q93288	contract
P31	instance_of	Q15416	television_program
P31	instance_of	Q1201097	detachment
P31	instance_of	Q16887380	group
P31	instance_of	Q57821	fortification
P31	instance_of	Q15383322	cultural prize
P31	instance_of	Q515	city
P31	instance_of	Q537127	road_bridge
P31	instance_of	Q20097897	sea_fort
P31	instance_of	Q1785071	fort
P31	instance_of	Q23413	castle
P31	instance_of	Q1484988	project
P31	instance_of	Q149621	district
P31	instance_of	Q532	village
P31	instance_of	Q2630741	community
P31	instance_of	Q3957	town
P31	instance_of	Q111161	synod
P31	instance_of	Q1530022	religious_organization
P31	instance_of	Q51645	ecumenical_council
P31	instance_of	Q10551516	church_council
P31	instance_of	Q1076486	sports_venue
P31	instance_of	Q17350442	venue
P31	instance_of	Q13226383	facility
P31	instance_of	Q811979	architectural_structure
P31	instance_of	Q23764314	sports_location
P31	instance_of	Q15707521	fictional_battle
P36	capital	*	
P2067	mass	*	
P1082	population	*	
P1376	captial_of	*	
P137	operator	*	
P915	filming_location	*	
P162	producer	*	
P281	postal_code	*	
P176	manufacturer	*	
P2257	event_interval	*	
P527	has_part	*	
P279	subclass_of	*	

Table 3.8: List of properties used for postprocessing Wikidata events. If a candidate event has the property ‘P31’, we prune them depending on the corresponding. For example, we only prune items that are instances of empire, historical period etc., For other properties like P527, P36, we prune items if they contain this property.

Language	**Code**	**Events**	**Mentions**	**Genus**
Afrikaans	af	316	2036	Germanic
Arabic	ar	2691	28801	Semitic
Belarusian	be	737	7091	Slavic
Bulgarian	bg	1426	12570	Slavic
Bengali	bn	270	3136	Indic
Catalan	ca	2631	22296	Romance
Czech	cs	2839	36658	Slavic
Danish	da	1189	10267	Germanic
German	de	7371	209469	Germanic
Greek	el	997	13361	Greek
English	en	10747	328789	Germanic
Spanish	es	5064	91896	Romance
Persian	fa	1566	10449	Iranian
Finnish	fi	3253	47944	Finnic
French	fr	8183	136482	Romance
Hebrew	he	1871	34470	Semitic
Hindi	hi	216	1219	Indic
Hungarian	hu	3067	27333	Ugric
Indonesian	id	2274	14049	Malayo-Sumbawan
Italian	it	7116	108012	Romance
Japanese	ja	3832	49198	Japanese
Korean	ko	1732	13544	Korean
Malayalam	ml	136	730	Southern Dravidian
Marathi	mr	132	507	Indic
Malay	ms	824	4650	Malayo-Sumbawan
Dutch	nl	4151	41973	Germanic
Norwegian	no	2514	24092	Germanic
Polish	pl	6270	110381	Slavic
Portuguese	pt	4466	45125	Romance
Romanian	ro	1224	12117	Romance
Russian	ru	7929	180891	Slavic
Sinhala	si	31	65	Indic
Slovak	sk	726	5748	Slavic
Slovene	sl	1288	8577	Slavic
Serbian	sr	1611	24093	Slavic
Swedish	sv	2865	23152	Germanic
Swahili	sw	22	74	Bantoid
Tamil	ta	250	1682	Southern Dravidian
Telugu	te	39	243	South-Central Dravidian
Thai	th	800	4749	Kam-Tai
Turkish	tr	2342	19846	Turkic
Ukrainian	uk	3428	53098	Slavic
Vietnamese	vi	1439	13744	Viet-Muong
Chinese	zh	2759	21259	Chinese
Total		**10947**	**1805866**	

Table 3.9: Proposed dataset summary (by languages)

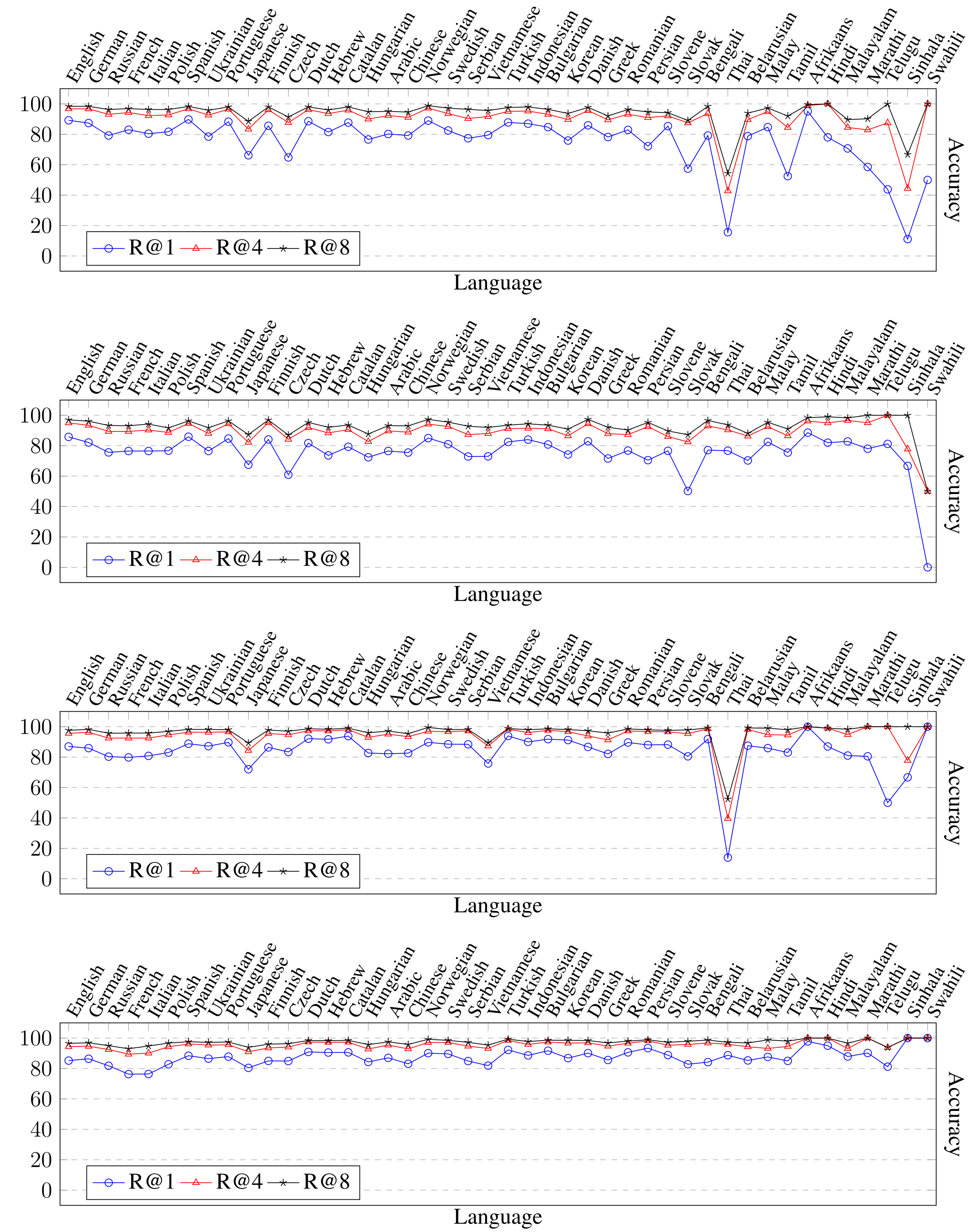


Figure 3.6: Retrieval recall scores on development set for mBERT and XLM-R in multilingual and crosslingual settings.

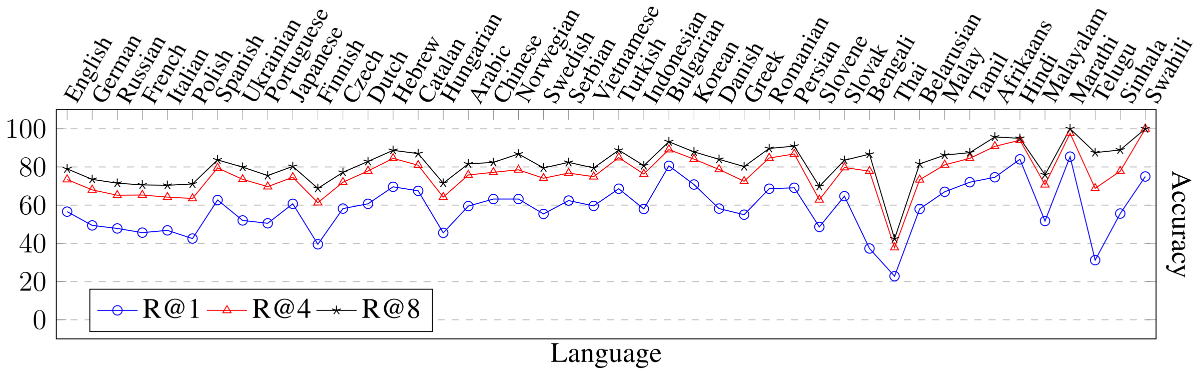


Figure 3.7: Retrieval recall scores on development set for BM25+ in multilingual setting.

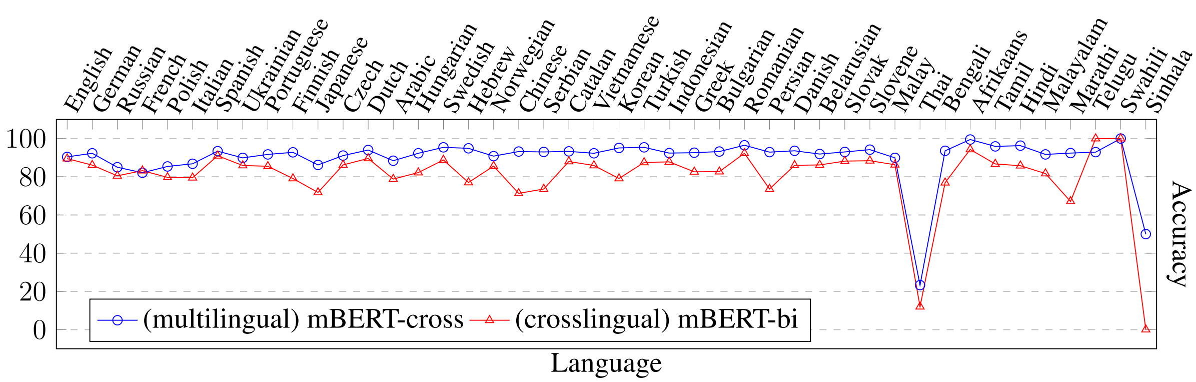


Figure 3.8: Test accuracy of mBERT-bi and mBERT-cross in multilingual and crosslingual tasks. The languages on the x-axis are sorted in the increasing order of mentions.

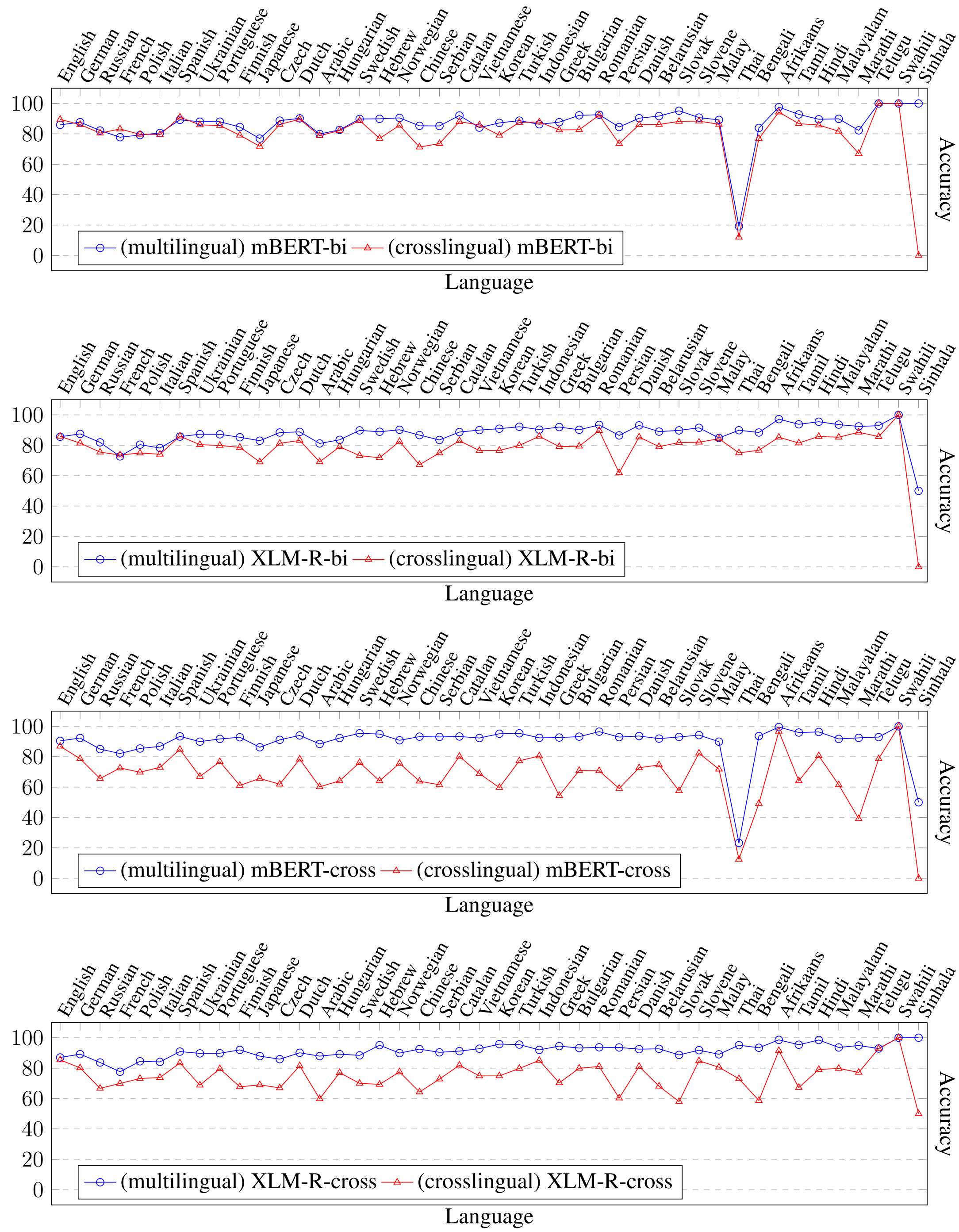


Figure 3.9: Test accuracy of mBERT-bi, XLM-R-bi, mBERT-cross, XLM-R-cross in multilingual and crosslingual tasks. The languages on the x-axis are sorted in the increasing order of mentions.

Mention Context: At the 2000 Summer Olympics in Sydney, Sitnikov competed only in two swimming events. He eclipsed a FINA B-cut of 51.69 (100 m freestyle) from the Kazakhstan Open Championships in Almaty. On the first day of the Games, Sitnikov placed twenty-first for the Kazakhstan team in the 4 $\times$ 100 m freestyle relay. Teaming with Sergey Borisenko, Pavel Sidorov, and Andrey Kvassov in heat three, Sitnikov swam a lead-off leg and recorded a split of 52.56, but the Kazakhs settled only for last place in a final time of 3:28.90. Three days later, in the **100 m freestyle**, Sitnikov placed fifty-third on the morning prelims. Swimming in heat five, he raced to a fifth seed by 0.15 seconds ahead of Chinese Taipei's Wu Nien-pin in 52.57.

Predicted Label: *Swimming at the 2008 Summer Olympics – Men's 100 metre freestyle*: The men's 100 metre freestyle event at the 2008 Olympic Games took place on 12–14 August at the Beijing National Aquatics Center in Beijing, China. There were 64 competitors from 55 nations.

Gold Label: *Swimming at the 2000 Summer Olympics – Men's 100 metre freestyle*: The men's 100 metre freestyle event at the 2000 Summer Olympics took place on 19–20 September at the Sydney International Aquatic Centre in Sydney, Australia. There were 73 competitors from 66 nations. Nations have been limited to two swimmers each since the 1984 Games.

Mention Context: In 2012, WWE reinstated their No Way Out pay-per-view (PPV), which had previously ran annually from 1999 to 2009. The following year, however, No Way Out was canceled and replaced by Payback, which in turn became an annual PPV for the promotion. The first Payback event was held on June 16, 2013 at the Allstate Arena in Rosemont, Illinois. The 2014 event was also held in June at the same arena and was also the first Payback to air on the WWE Network, which had launched earlier that year. In 2015 and 2016, the event was held in May. The 2016 event was also promoted as the first PPV of the New Era for WWE. In July 2016, WWE reintroduced the brand extension, dividing the roster between the Raw and SmackDown brands where wrestlers are exclusively assigned to perform. The **2017 event** was in turn held exclusively for wrestlers from the Raw brand, and was also moved up to late-April.

Predicted Label: *Battleground (2017)*: Battleground was a professional wrestling pay-per-view (PPV) event and WWE Network event produced by WWE for their SmackDown brand division. It took place on July 23, 2017, at the Wells Fargo Center in Philadelphia, Pennsylvania. It was the fifth and final event under the Battleground chronology, as following WrestleMania 34 in April 2018, brand-exclusive PPVs were discontinued, resulting in WWE reducing the amount of yearly PPVs produced.

Gold Label: *Payback (2017)*: Payback was a professional wrestling pay-per-view (PPV) and WWE Network event, produced by WWE for the Raw brand division. It took place on April 30, 2017 at the SAP Center in San Jose, California. It was the fifth event in the Payback chronology. Due to the Superstar Shake-up, the event included two interbrand matches with SmackDown wrestlers. It was the final Payback event until 2020, as following WrestleMania 34 in 2018, WWE discontinued brand-exclusive PPVs, which resulted in the reduction of yearly PPVs produced.

Table 3.10: Examples of errors by the event linking system. (temporal reasoning related)

Mention Context: Paul Wing (August 14, 1892 – May 29, 1957) was an assistant director at Paramount Pictures. He won the **1935** Best Assistant Director Academy Award for "The Lives of a Bengal Lancer" along with Clem Beauchamp. Wing was the assistant director on only two films owing to his service in the United States Army. During his service, Wing was in a prisoner camp that was portrayed in the film "The Great Raid" (2005). **Predicted Label:** *8th Academy Awards*: The 8th Academy Awards were held on March 5, 1936, at the Biltmore Hotel in Los Angeles, California. They were hosted by Frank Capra. This was the first year in which the gold statuettes were called "Oscars". **Gold Label:** *7th Academy Awards*: The 7th Academy Awards, honoring the best in film for 1934, was held on February 27, 1935, at the Biltmore Hotel in Los Angeles, California. They were hosted by Irvin S. Cobb.
Mention Context: Für "Holiday Land" (1934) war er bei der Oscarverleihung 1935 erstmals für einen Oscar für den besten animierten Kurzfilm nominiert. Eine weitere Nominierung in dieser Kategorie erhielt er **1938** für "The Little Match Girl" (1937). **Predicted Label:** *9th Academy Awards*: The 9th Academy Awards were held on March 4, 1937, at the Biltmore Hotel in Los Angeles, California. They were hosted by George Jessel; music was provided by the Victor Young Orchestra, which at the time featured Spike Jones on drums. This ceremony marked the introduction of the Best Supporting Actor and Best Supporting Actress categories, and was the first year that the awards for directing and acting were fixed at five nominees per category. **Gold Label:** *10th Academy Awards*: The 10th Academy Awards were originally scheduled for March 3, 1938, but due to the Los Angeles flood of 1938 were held on March 10, 1938, at the Biltmore Hotel in Los Angeles, California. It was hosted by Bob Burns.

Table 3.11: Examples of errors by the event linking system. (temporal or spatial expression related)

Mention Context: Nel 2018 ha preso parte alle Olimpiadi di Pyeongchang, venendo eliminata nel primo turno della finale e classificandosi diciannovesima nella gara di **gobbe**.

Predicted Label: *Snowboarding at the 2018 Winter Olympics – Women's parallel giant slalom*: The women's parallel giant slalom competition of the 2018 Winter Olympics was held on 24 February 2018 Bogwang Phoenix Park in Pyeongchang, South Korea.

Gold Label: *Freestyle skiing at the 2018 Winter Olympics – Women's moguls*: The Women's moguls event in freestyle skiing at the 2018 Winter Olympics took place at the Bogwang Phoenix Park, Pyeongchang, South Korea from 9 to 11 February 2018. It was won by Perrine Laffont, with Justine Dufour-Lapointe taking silver and Yuliya Galysheva taking bronze. For Laffont and Galysheva these were first Olympic medals. Galysheva also won the first ever medal in Kazakhstan in freestyle skiing.

Mention Context: تقارب إسرائيل واليابان على أساس القيم الديموقراطيةوا لاشتراكية ا لمشتركة، واستطاعت من خ لال عضويتها في ا لاشتراكية الدولية أن تنشئ صلات وثيقة مع الحزب ا لاشتراكي الياباني الذي تبنى مهمة التعريف بإسرائيل ومنجزاتها في اليابان. وإ ّبان **حرب 1956** انضمت اليابان إلى الدول التي طالبت مصر باحترام ا لمعاهدات الدولية الخاصة با لملاحة في قناة السويس. وأصدرت بيان مقتضب .أعلنت فيه أسفها لوصول ا لأمور إلى حد الصدام ا لمسلح

Predicted Label: *Hungarian Revolution of 1956*: The Hungarian Revolution of 1956 (), or the Hungarian Uprising, was a nationwide revolution against the Hungarian People's Republic and its Soviet-imposed policies, lasting from 23 October until 10 November 1956. Leaderless at the beginning, it was the first major threat to Soviet control since the Red Army drove Nazi Germany from its territory at the end of World War II in Europe.

Gold Label: *Suez Crisis*: The Suez Crisis, or the Second Arab–Israeli war, also called the Tripartite Aggression () in the Arab world and the Sinai War in Israel,

---

Mention Context: 攝津號戰艦於1909年4月1日在須賀海軍工廠鋪設龍骨，後於1909年1日18日舉行下水儀式，並於1912年7月1日竣工，總造價為11,010,000日圓。海軍大佐田中盛秀於1912年12月1日出任本艦艦長，並編入第一分遣艦隊。翌年的多數時候，攝津號均巡航於中國外海或是接受戰備操演。當第一次世界大戰於1914年8月間爆發時，本艦正停泊於廣島縣市軍港。攝津號與其姐妹艦河號於1914年10月至11月間參與了青島戰役的最後階段，並於外海以艦砲密集轟炸軍陣地。本艦於1916年12月1日離開第一分遣艦隊，並送往市進行升級作業。升級作業於1917年12月1日完成，該艦隨後編入第二分遣艦隊，直至1918年7月23日重新歸入第一分遣艦隊為止。自此時起，攝津號戰艦上所有的QF 12磅3英吋40倍徑艦砲均移除，並以QF 12磅3英吋40倍徑防空砲取代，另亦移除了兩具魚雷發射管。1918年10月28日，攝津號戰艦成為大正天皇於海上校時所搭乘的旗艦。

Predicted Label: *Battle of the Yellow Sea*: The Battle of the Yellow Sea (;) was a major naval battle of the Russo-Japanese War, fought on 10 August 1904. In the Russian Navy, it was referred to as the Battle of 10 August. The battle foiled an attempt by the Russian fleet at Port Arthur to break out and form up with the Vladivostok squadron, forcing them to return to port. Four days later, the Battle off Ulsan similarly ended the Vladivostok group's sortie, forcing both fleets to remain at anchor.

Gold Label: *Siege of Tsingtao*: The siege of Tsingtao (or Tsingtau) was the attack on the German port of Tsingtao (now Qingdao) in China during World War I by Japan and the United Kingdom. The siege was waged against Imperial Germany between 27 August and 7 November 1914. The siege was the first encounter between Japanese and German forces, the first Anglo-Japanese operation of the war, and the only major land battle in the Asian and Pacific theatre during World War I.

Table 3.12: Examples of errors by the event linking system. (language-related)

Mention Context: He established his own production company, Emirau Productions, named after the **battle in World War II** in which Warren was injured.

Predicted Label: *First Battle of El Alamein*: The First Battle of El Alamein (1–27 July 1942) was a battle of the Western Desert Campaign of the Second World War, fought in Egypt between Axis forces (Germany and Italy) of the Panzer Army Africa () (which included the under Field Marshal () Erwin Rommel) and Allied (British Imperial and Commonwealth) forces (Britain, British India, Australia, South Africa and New Zealand) of the Eighth Army (General Claude Auchinleck).

Gold Label: *Landing on Emirau*: The Landing on Emirau was the last of the series of operations that made up Operation Cartwheel, General Douglas MacArthur's strategy for the encirclement of the major Japanese base at Rabaul. A force of nearly 4,000 United States Marines landed on the island of Emirau on 20 March 1944. The island was not occupied by the Japanese and there was no fighting. It was developed into an airbase which formed the final link in the chain of bases surrounding Rabaul. The isolation of Rabaul permitted MacArthur to turn his attention westward and commence his drive along the north coast of New Guinea toward the Philippines.

Mention Context: Ivanova won the silver medal at the 1978 World Junior Championships. She made her senior World debut at the **1979 World Championships**, finishing 18th. Ivanova was 16th at the 1980 Winter Olympics.

Predicted Label: *FIBT World Championships 1979*: The FIBT World Championships 1979 took place in Königssee, West Germany. It was the first championships that took place on an artificially refrigerated track. The track also hosted the luge world championships that same year, the first time that had ever happened in both bobsleigh and luge in a non-Winter Olympic year (Igls hosted both events for the 1976 games in neighboring Innsbruck.).

Gold Label: *1979 World Figure Skating Championships*: The 1979 World Figure Skating Championships were held in Vienna, Austria from March 13 to 18. At the event, sanctioned by the International Skating Union, medals were awarded in men's singles, ladies' singles, pair skating, and ice dance.

Mention Context: Изначально открытие башни должно было состояться в декабре 2011 года, но после землетрясения строительство замедлилось из-за нехватки средств.

Predicted Label: *2011 Christchurch earthquake*: A major earthquake occurred in Christchurch, New Zealand, on Tuesday 22 February 2011 at 12:51 p.m. local time (23:51 UTC, 21 February). The () earthquake struck the Canterbury region in the South Island, centred south-east of the centre of Christchurch, the country's second-most populous city. It caused widespread damage across Christchurch, killing 185 people, in the nation's fifth-deadliest disaster.

Gold Label: *2011 Tōhoku earthquake and tsunami*: The occurred at 14:46 JST (05:46 UTC) on 11 March. The magnitude 9.0–9.1 (Mw) undersea megathrust earthquake had an epicenter in the Pacific Ocean, east of the Oshika Peninsula of the Tōhoku region, and lasted approximately six minutes, causing a tsunami. It is sometimes known in Japan as the , among other names. The disaster is often referred to in both Japanese and English as simply 3.11 (read s̈an ten ichi-ichiïn Japanese).

Mention Context: ポワント・デュ・オック (Pointe du Hoc) から向かったアメリカ軍のレンジャー部隊の8個中隊と共に、アメリカ第29歩兵師団は海岸の西側の側面を攻撃した。アメリカ第1歩兵師団は東側からのアプローチを行った。これは、この戦争において、北アフリカ、シチリア島に続く3回目の強襲上陸であった。オマハビーチの上陸部隊の主目標は、サン＝ロー (Saint-Lô) の南に進出する前にポール＝アン＝ベッサン (Port-en-Bessin) とヴィル川 (Vire River) 間の橋頭堡を守ることであった。

Predicted Label: *Tunisian campaign*: The Tunisian campaign (also known as the Battle of Tunisia) was a series of battles that took place in Tunisia during the North African campaign of the Second World War, between Axis and Allied forces. The Allies consisted of British Imperial Forces, including a Greek contingent, with American and French corps. The battle opened with initial success by the German and Italian forces but the massive supply interdiction efforts led to the decisive defeat of the Axis. Over 250,000 German and Italian troops were taken as prisoners of war, including most of the Afrika Korps.

Gold Label: *Operation Torch*: Operation Torch (8 November 1942 – 16 November 1942) was an Allied invasion of French North Africa during the Second World War. While the French colonies formally aligned with Germany via Vichy France, the loyalties of the population were mixed. Reports indicated that they might support the Allies. American General Dwight D. Eisenhower, supreme commander of the Allied forces in Mediterranean Theater of Operations, planned a three-pronged attack on Casablanca (Western), Oran (Center) and Algiers (Eastern), then a rapid move on Tunis to catch Axis forces in North Africa from the west in conjunction with Allied advance from east.

Table 3.13: Examples of errors by the event linking system. (also errors in the dataset)

Chapter 4

Background Summarization of Event Timelines

This work was published at EMNLP 2023 (Pratapa et al., 2023).[1]

4.1 Introduction

should Events such as natural disasters, political conflicts, and elections are extensively covered by news agencies and followed by readers throughout the world. Generating concise summaries of these events is a challenging NLP task (Chen et al., 2021). For popular news stories, journalists curate retrospective timelines that highlight key sub-events on a timeline. However, for a new observer of a specific major event, catching up on the historical context needed to understand the significance of the sub-event update can be a challenging ordeal. To this end, we present the task of background news summarization that complements each update in a timeline with a background summary.

Timeline summarization is the task of automatically extracting event timelines from a collection of news articles (Chieu and Lee, 2004). Update summarization (Dang and Owczarzak, 2009) involves summarizing a set of recent articles, assuming the reader is already familiar with a set of background articles. It has since been studied in shared tasks that track events in newswire (Aslam et al., 2015) and Twitter feeds (Sequiera et al., 2018). Our novel proposed task of background summarization presents an orthogonal use case for the update summarization task. A background summary provides *sufficient historical context* to the reader to help them understand the latest news update. It summarizes what has happened previously, in order to explain the background of the current news update.

Background summaries allow the reader to quickly grasp the historical context of an event without having to read through potentially hundreds of news articles or long timelines regarding a specific event. One application would be to contextualize short-text content (e.g., Tweets) with background information from news articles about the events. In addition to providing much-needed context, this can be useful for verifying the factuality of the events described in the tweet (e.g., Twitter Community Notes). In a news-centric conversational AI setting, a background

[1]Code and data are available at: `https://github.com/amazon-science/background-summaries`.

summary may be generated to answer a user request to "tell me what I need to know to understand this event".

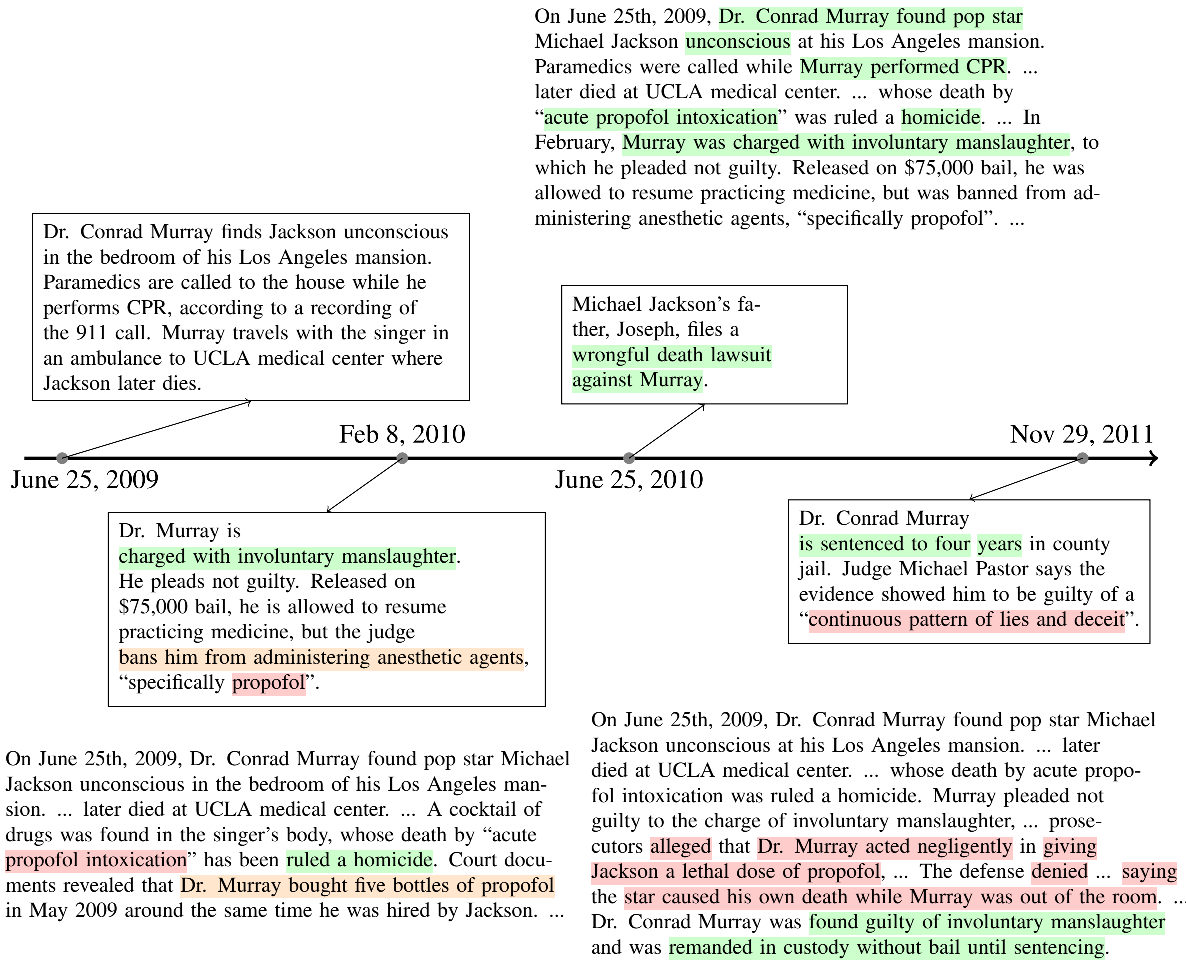


Figure 4.1: An illustration of the background summarization task. This is a snapshot from the timeline of the *Michael Jackson's Death* event. The timeline above shows four news updates between June 25, 2009, and November 29, 2011. Each $\boxed{\text{update}}$ is complemented with a background summary that provides sufficient historical context to the events and entities described in the update. We highlight phrases from the background that provide context to specific phrases in the update text.

To construct a dataset for background summarization, we identify existing timeline datasets centered around major news events. Specifically, we select the widely-used Timeline17 (Binh Tran et al., 2013), Crisis (Tran et al., 2015), and Social Timeline (Wang et al., 2015) datasets – identifying 14 major news events from these datasets and prepare a single timeline of events for each major event. The original datasets included multiple timelines for major events, collected from different news agencies. For simplifying our annotation process, we merged all the timelines for a given major event to create a single timeline. We present these timelines to expert annotators and collect background summaries at each timestep for each news event. Figure 4.1 provides an example from the timeline of Michael Jackson's death. The timeline starts on June 25th, 2009 with a news update. The following updates on Feb 8, June 25, and Nov 29 are substantiated using background summaries that provide historical context regarding the event.

On the resulting annotated dataset, we experiment with a suite of summarization systems based on Flan-T5 (Chung et al., 2022), LongT5 (Guo et al., 2022), and GPT-3.5 (OpenAI, 2022). We propose to formulate the background summarization task as two different variants: (1) In a *generic* summarization setting, we generate a background summary for the current update at time step t by using a concatenation of the past updates at time steps 1 through $t-1$. (2) In a *query-focused* setting (Dang, 2005; Xu and Lapata, 2022), we utilize the update at time step t as a query. While the background summary does not include content from the current update, the current update can still be used for conditioning the summarization of past updates. This could potentially improve the utility of the backgrounds. For the query-focused summarization, we explore multiple variants of the query and present a comparison of Flan-T5 and GPT-3.5.

Our experiments indicate that the fine-tuned Flan-T5 system outperforms GPT-3.5 on the standard ROUGE metric while underperforming on factuality metrics. This illustrates the challenges associated with automatic metrics for summarization. Goyal et al. (2022a) and Zhang et al. (2023) highlight the difficulties in using automatic metrics for comparing fine-tuned system summaries against zero-shot summaries from high-quality large language models (LLMs). Accordingly, we also present a novel question-answering-based evaluation of background summaries in this work that measures the utility of a given background summary to an update. First, we prompt GPT-3.5 to generate questions about the background of events and entities mentioned in the update. Second, we extract answers from the background summaries to measure their effectiveness in providing sufficient historical context to the update. Our proposed Background Utility Score (BUS) measures the percentage of questions about current news updates that are answerable by a background summary. Our human evaluation results show the effectiveness of BUS. Our key contributions are as follows:

1. We present the new task of background summarization to help readers follow day-to-day updates on complex real-world events.
2. We describe an expert-annotated dataset covering 14 major news events from 2005–2014 with over 1,100 background summaries.
3. We conduct benchmark experiments using state-of-the-art zero-shot and fine-tuned summarization systems. We also explore query-focused summarization that generates the background by using the current update text as a query.
4. We propose an effective QA-based evaluation metric, Background Utility Score (BUS), to measure the utility of a background summary with respect to contextualizing an update.

4.2 Related Work

Events in the real world are often complex, consisting of numerous threads (Liu et al., 2017), and are reported by a large number of news agencies across the world. Tracking these events and providing important and useful updates to users has been the focus of many works in natural language processing and information retrieval communities (Chen et al., 2021). We highlight

	Major event	Sources (# timelines)	Time period	# U	$len(U)$	$len(B)$
train	Swine flu	T17 (3)	2009	21	52	45
	Financial crisis	T17 (1)	2008	65	115	147
	Iraq war	T17 (1)	2005	155	41	162
dev	Haitian earthquake	T17 (1)	2010	11	100	61
	Michael Jackson death	T17 (1)	2009–2011	37	36	164
	BP oil spill	T17 (5)	2010–2012	118	56	219
test	NSA leak	SocialTimeline (1)	2014	29	45	50
	Gaza conflict	SocialTimeline (1)	2014	38	183	263
	MH370 flight disappearance	SocialTimeline (1)	2014	39	39	127
	Yemen crisis	Crisis (6)	2011–2012	81	30	125
	Russian-Ukraine conflict	SocialTimeline (3)	2014	86	112	236
	Libyan crisis	T17 (2); Crisis (7)	2011	118	38	177
	Egyptian crisis	T17 (1); Crisis (4)	2011–2013	129	34	187
	Syrian crisis	T17 (4); Crisis (5)	2011–2013	164	30	162

Table 4.1: An overview of the news events used in our background summarization dataset. The events are grouped into train, validation, and test splits. We list the source dataset and the number of source timelines for each event. The time period provides the overall span of the event timeline. The length of the timeline, the average word count of the (rewritten) updates, and newly annotated backgrounds are specified in the final columns.

two specific variants of the event summarization task below,

Timeline summarization: Given a corpus of documents and a query, the task is to retrospectively extract important events from the documents and place them along a timeline (Chieu and Lee, 2004). A typical query consists of major events such as the Haitian earthquake or the BP oil spill. Datasets rely on timelines compiled by news journalists from agencies such as BBC, Reuters, and The New York Times, amongst others. Notable datasets for this task include Timeline17 (Binh Tran et al., 2013), Crisis (Tran et al., 2015), Social Timeline (Wang et al., 2015), entities dataset (Gholipour Ghalandari and Ifrim, 2020), and TLS-Newsroom (Born et al., 2020).

Update summarization: Dang and Owczarzak (2009) first proposed the task of update summarization. Given two sets of documents A & B, the task is to generate a query-focused update summary of the document set B assuming the user of the summary has already read the documents from set A. This task has since been studied on documents from newswire (Aslam et al., 2015) and Twitter feeds (Sequiera et al., 2018). In contrast to the timeline summarization task, systems do not have access to the documents from the future. Updating users about critical news events in real-time is very important to news and government agencies (Guo et al., 2013). However, comprehending these updates can be challenging for new readers. Our proposed task of background summarization serves a complimentary purpose to updates.

Background summarization: Hayashi et al. (2020) proposed the task of *disentangled paper summarization*, in which two separate summaries are generated for an academic paper: one summary describing the paper contribution, and another summarizing paper context. A related setting for news events can be a disentangled summarization of updates and backgrounds of events. Chen et al. (2022) presented a dataset of TV series transcripts and human-written recaps (SummScreen). Similar to a background in our task setup, recaps can help viewers understand the current episode. A key distinction is that a recap typically provides information from the most recent episode(s) only, but it does not provide general context to the story. In contrast, background summaries often include information from the very first update to put the current event in context.

Some prior works have studied the impact of background knowledge in the standard summarization task setup (Louis, 2014; Peyrard and West, 2020). A typical summarization setup requires a system to generate a summary of a collection of documents while ignoring any background knowledge already known to the receiver (or reader). Peyrard and West (2020) used a broader definition of background. In their setup, the background constitutes a document collection that the user is already familiar with, similar to the document set A from the above definition of update summarization. Our definition of background constitutes a summary of previous updates in a given event that are directly relevant to the current, most recent update.

Long-form summarization: Our proposed task often contains long timelines of events, requiring systems to perform long-form summarization. There is a growing effort in the community to improve long-range summarization systems. This includes works on book summarization (Wu et al., 2021), meeting summarization (Zhang et al., 2022), TV script summarization (Chen et al., 2022) and evaluation of long-form summarization systems (Krishna et al., 2023).

4.3 Background Summarization

Event timelines help readers keep track of updates regarding major news events. They provide a concise overview of the event's progress over time, without the need to read through hundreds or thousands of news articles written about the event. However, for long-lasting events, keeping track of all the sub-event threads can pose a major challenge for the user (Liu et al., 2017). We postulate that complementing each update with a short background summary regarding the event's past can assist the user in understanding the update. Our approach is inspired by the standard inverted pyramid structure of news articles (Pöttker, 2003). Typically, news articles consist of new newsworthy information at the top, followed by further details about the story, and end with necessary background information. This background information helps the reader gain a full perspective of the news story. In this work, we extend this to news timelines.

4.3.1 Task

Given an event timeline consisting of a time series of updates $\langle U_1, \ldots, U_T \rangle$, the task is to generate background summaries $\langle B_2, \ldots, B_T \rangle$ for all updates after U_1. For each timestep $t > 1$, we wish to find the background summary B_t that maximizes $p(B_t \mid U_1, \ldots, U_{t-1}; q)$ where q is a query. In the generic baseline setting, q is empty; in the query-focused setting, q is set to the current update U_t. In the latter case, we do not aim to summarize U_t, but we use it to direct the summarization of the previous updates toward content that can help explain the current update U_t.

Note that each background summary B_t is generated directly from the previous updates, independently from the previous background summary B_{t-1}. This enables us to include details relevant to U_t from particular previous updates that may not be found in B_{t-1}.

4.3.2 Dataset Construction

To the best of our knowledge, there are no existing datasets that provide background summaries. Accordingly, we compile a new, expert-annotated dataset for this, building upon three popular news timeline summarization datasets, Timeline17 (Binh Tran et al., 2013), Crisis (Tran et al., 2015), and Social Timeline (Wang et al., 2015).

Timeline17: compiled from an ensemble of news websites, this dataset provides 17 timelines spanning 9 major events from 2005–2013.

Crisis: a follow-up to the Timeline17 dataset, this covers 25 timelines spanning 4 major events. While it mostly covers a subset of events from Timeline17, it adds a new event (the Yemen crisis).

Social Timeline: compiled 6 timelines covering 4 major events from 2014. The timelines were collected from Wikipedia, NYTimes, and BBC.

Table 4.1 provides an overview of the 14 major news events compiled from the three datasets. Since the timelines were collected from various news websites (CNN, BBC, NYTimes, etc.), many events have more than one timeline. As each timeline covers the same underlying event, we merge them using timestamps to create a single timeline per event. During this merging process, we often end up with more than one update text per timestamp with possibly duplicate content. We ask the annotators to first rewrite the input updates to remove any duplicate content. Our annotation process for each news event contains the following three steps:

1. Read the input timeline to get a high-level understanding of the event.
2. For each timestep, read the provided 'rough' update summary. Rewrite the update into a short paragraph, removing any duplicate or previously reported subevents.
3. Go through the timeline in a sequential manner and write a background summary for each timestep.

Based on this process, we hired three professional annotators. For each timeline, we collect three independent (rewritten) update and (new) background pairs. Our full annotator guidelines are provided in Table 4.5 in the Appendix. Due to minor differences in the rewritten updates in the timelines, we do not merge the annotator timelines. Table 4.1 provides average lengths of rewritten updates and newly annotated background summaries for each major event. In our final dataset, each timestep in the timeline has three pairs of rewritten updates and background summaries.

	ROUGE-1	ROUGE-2	ROUGE-L
Rewritten updates			
Annotator 1	80.9	64.4	74.9
Annotator 2	72.9	54.2	66.2
Annotator 3	80.1	63.2	73.3
Background summaries			
Annotator 1	47.9	21.3	43.3
Annotator 2	44.9	16.6	39.5
Annotator 3	48.0	21.1	43.4

Table 4.2: IAA across 14 major events.

4.3.3 Dataset Splits

For our experiments, we split the 14 major events into a train (3 events), validation (3 events), and test set (8 events). Table 4.1 lists the events in each split. We include a mixture of short and long timelines across the splits and the test set is mostly temporally separated from the train/dev splits.

Considering the strong few-shot summarization capabilities of large language models (Goyal et al., 2022a; Zhang et al., 2023), we decided to budget only a small fraction of expert-annotated data for training and development and leave most events to the test set. This allows sufficient data for further fine-tuning instruction-based models (Flan, GPT-3+) to our new task while maintaining sufficient diversity in the test set.

4.3.4 Inter-annotator Agreement

To measure the inter-annotator agreement (IAA), we compute ROUGE scores[2] with one annotator's summary as the hypothesis and the remaining two annotators' summaries as references.[3] Table 4.2 presents the IAA scores for both the re-written updates and the newly annotated backgrounds. As expected, we see high ROUGE scores on the rewritten updates. The scores are lower for background summaries, indicating the inherent variance in background summaries.

4.3.5 Background Utility Score (BUS)

Automatic metrics such as ROUGE are found to correlate poorly with human judgments of summaries (Louis and Nenkova, 2013; Peyrard, 2019). Recent studies highlighted the ineffectiveness of standard metrics when comparing fine-tuned and zero-shot summaries (Goyal et al., 2022a; Zhang et al., 2023). To account for these limitations and the need to evaluate the quality of backgrounds, we propose a QA-based metric for the background summarization task. Our metric, Background Utility Score (BUS), measures the utility of a background B_t to the corresponding update U_t.

[2]For ROUGE-L, we use the Lsum variant in the HuggingFace `evaluate` package.

[3]Multi-reference ROUGE returns a maximum score among references.

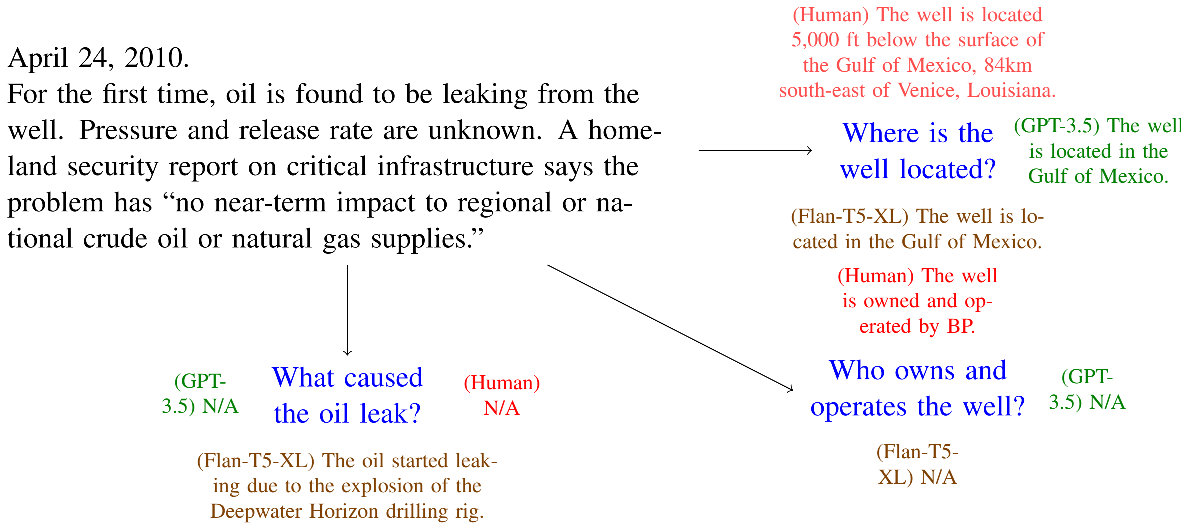


Figure 4.2: Examples of question-answer pairs for BUS (§4.3.5) generated by prompting GPT-3.5. This example shows an update text from the BP oil spill event. Questions are generated from the current update U_t, and the answers are generated based on three different background summaries B_t (Human, GPT-3.5, Flan-T5-XL); N/A means the background summary did not provide an answer. The BUS score is calculated per system as the percentage of questions answered by its background summaries. See Table 4.8 in the Appendix for the full background summaries.

To measure the utility, we first prompt a GPT-based model to generate (background) questions from the update text (U_t). We then re-prompt the model to extract answers from the background text (B_t). BUS measures the percentage of questions answerable by the background. Figure 4.2 presents examples of generated QA pairs. The background summary should be able to answer any questions the reader may have upon observing an update. While these questions are latent, we sample them by prompting a GPT-based model.

BUS is inspired by QuestEval (Scialom et al., 2021), an interpretable QA-based factuality metric for summarization. QuestEval measures the recall score by extracting questions from the source and computing the exact match F_1 between answer spans from the source and summary (vice-versa for precision). BUS is also tangentially related to recent LLM-based evaluation systems such as in Vicuna (Chiang et al., 2023) that explored the use of chatbots for evaluating chatbots.

4.4 Experiments

For our background summarization task, we experiment with three summarization systems, Flan-T5 (Chung et al., 2022), LongT5 (Guo et al., 2022), and GPT-3.5 (OpenAI, 2022).

Flan-T5: an instruction fine-tuned version of T5 (Raffel et al., 2020). We use Flan-T5-XL with a maximum source length of 512 tokens.[4]

[4] https://hf.co/google/flan-t5-xl

	ROUGE-1	ROUGE-2	ROUGE-L	QuestEval	BERTScore P	BUS–GPT-3.5
generic						
Flan-T5-XL	43.5 | 41.4	20.4 | 17.4	39.9 | 37.6	31.2 | 25.0	86.3 | 85.6	46.0 | 42.2
GPT-3.5	40.5 | 37.7	15.5 | 11.7	36.6 | 33.0	37.2 | 30.5	88.2 | 87.2	59.1 | 54.3
query-focused						
Flan-T5-XL	43.0 | 41.3	20.6 | 17.4	39.5 | 37.6	30.8 | 24.9	86.2 | 85.6	46.6 | 43.6
GPT-3.5	40.2 | 40.5	15.4 | 12.9	36.1 | 35.9	36.9 | 31.7	87.9 | 87.5	49.9 | 47.5

Table 4.3: System performance (dev | test) on the background summarization task.

LongT5: a sparse attention variant of T5 that utilizes two efficient attentions, local and transient-global. Source length can be significantly longer than the standard 512 token limits of a T5-based system (§4.3.2). We use the Long-T5-TGlobal-XL with a maximum source length of 4096.[5]

GPT-3.5: a variant of the InstructGPT model (Ouyang et al., 2022) optimized for dialogue using reinforcement learning with human feedback. We use this model in a zero-shot setting. We set a maximum source length of 3696.[6]

We explore both generic and query-focused summarization settings (§4.3.1). In the query-focused setting, we use the current update (U_t) as an additional input to the summarization system.

Generic: We use a task prefix 'summarize:' to instruct T5-based systems. For GPT-3.5, we use a task suffix, 'Provide a short summary of the above article.'

Query-focused: The input for the T5-based systems follows the template, 'Generate a short query-focused summary of the background. Query: <query>, Background: <past updates>.' For GPT-3.5, we use a task suffix, 'Generate a short query-focused summary of the background.' We use 512 and 128 limits for source and query. We consider two variants for queries. First, we use the full update (U_t) as the query. Second, we first extract named entities from and use those keywords as the query. The named entity-based approach removes any potential noise from the update and focuses solely on extracting background information about important persons or locations specified in the update. We use SpaCy English NER model to extract named entities from the query.

Across all our systems, when necessary we truncate the oldest updates from the input.[7] We train both Flan-T5 and LongT5 using DeepSpeed's ZeRO Stage 3 (Rasley et al., 2020). We set a maximum target length of 400 tokens.

BUS: We use GPT-3.5 as our question and answer generation system (ref. BUS–GPT-3.5). We generate five questions per update and use heuristic patterns on GPT answers to identify unanswerable questions.[8] Following recent work that showed better human alignment with GPT-

[5] https://hf.co/google/long-t5-tglobal-xl

[6] At the time of our experiments, this corresponds to the `gpt-3.5-turbo-0301` version. https://platform.openai.com/docs/models/gpt-3-5

[7] Other viable options are truncating middle updates or ranking updates based on their relevance to the current update.

[8] see Table 4.6 in Appendix for the instruction templates.

Human	Flan-T5-XL	GPT-3.5
0.2430	-0.0750	-0.1680

Table 4.4: Results of the human evaluation on AMT using best-worst scaling (BWS). Values range from -1 (worst) to $+1$ (best).

4 (Liu et al., 2023a), we also experiment with BUS–GPT-4.[9]

4.5 Results

4.5.1 Automatic Evaluation

Table 4.3 presents the results on validation and test sets for Flan-T5 and GPT-3.5 in both generic and the NER-based query-focused setups. We report scores on the standard summarization metric ROUGE (Lin, 2004), two factuality metrics QuestEval (Scialom et al., 2021), and BERTScore Precision (Zhang* et al., 2020; Pagnoni et al., 2021) and our proposed utility metric BUS (§4.3.5).

On the generic summarization setup, we observe that fine-tuned Flan-T5 outperforms zero-shot GPT-3.5 on ROUGE. However, the zero-shot GPT-3.5 model fares much better on factuality metrics and BUS. These trends are also valid in the NER-based query-focused formulation. Interestingly, we find the query-focused formulation generally underperforms the generic task.[10] In our experiments, we found Long-T5 underperforms Flan-T5 on the dev set (Table 4.9 in Appendix). We leave further evaluation of Long-T5-based systems for future work.

4.5.2 Human Evaluation

We conduct a human evaluation to determine the relative quality of the human-written backgrounds (subsection 4.3.2) and those generated by Flan-T5-XL and GPT-3.5 (generic; top-half of Table 4.3). We chose to evaluate the generic systems instead of query-focused systems due to their superior performance on ROUGE and factuality metrics on the development set.

Setup: We use the Amazon Mechanical Turk (AMT) platform. We sample 1,000 news updates from the test set and pair each one with the three background summaries, displayed in random order. We collect judgments from three annotators about which of the three displayed summaries is the best (i.e., most helpful) and which one is the worst (i.e., least helpful). We use majority voting to pick the best and worst summaries. Detailed instructions are shown Figure 4.5 in Appendix. Since annotators on the AMT platform are non-experts, we use multiple methods to obtain high-quality judgments, including a qualification test and time controls; details including fair compensation of the annotators are described in Appendix 4.8.2.

[9]We use `gpt-3.5-turbo-0301` and `gpt-4-0613`.

[10]We present further analysis in 4.8.5 in the Appendix.

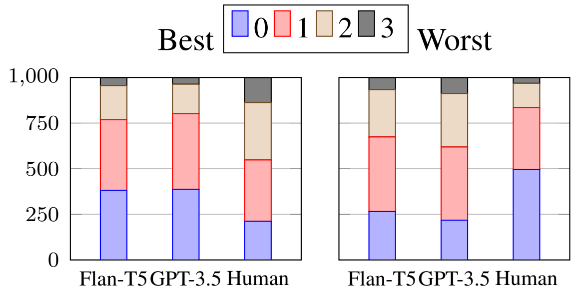


Figure 4.3: Vote distribution for best and worst systems from our human evaluation.

Results: We use best-worst-scaling (BWS; Kiritchenko and Mohammad (2017)); Table 4.4 shows the results. The values are computed as the percentage of times a summary type is chosen as best minus the percentage of times it is selected as worst. Values of 1.0 or -1.0 indicate that the system has been unanimously picked as 'best' and 'worst' respectively. We observe that the human-written summaries are substantially preferred over both Flan-T5-XL and GPT-3.5 summaries.

Agreement: Figure 4.3 presents the vote distribution for the best and worst summaries across the 1k examples. Human-written backgrounds are rated the best by at least two annotators in 45% of the examples. They were rated the worst in less than 17% of the examples. Flan-T5 and GPT-3.5 have very similar best-vote distributions (23% and 20%). We see unanimous agreement on the best or worst system in less than 15% of the examples.

Justifications: Annotators tend to prefer human backgrounds over GPT-3.5's due to their comprehensiveness. In the justifications we collected as a part of our AMT evaluation, human backgrounds were described as 'most comprehensive', and providing 'complete context'. On the other hand, GPT-3.5 backgrounds were described as 'too short', 'just a timeline', and providing 'least information'.

4.6 BUS Analysis

Our human evaluation results showed variance amongst Turkers (Figure 4.3). This is in line with the observations made by prior work on standard summarization datasets (Goyal et al., 2022a; Zhang et al., 2023). While human evaluation can be very useful, past work highlighted the difficulties in choosing evaluation dimensions and task design (Khashabi et al., 2022). Goyal et al. (2022a) recommends using an evaluation setup based on how users utilize the system in practice. To this end, we analyze the effectiveness of BUS (§4.3.5) in measuring the real-world utility of background summaries.

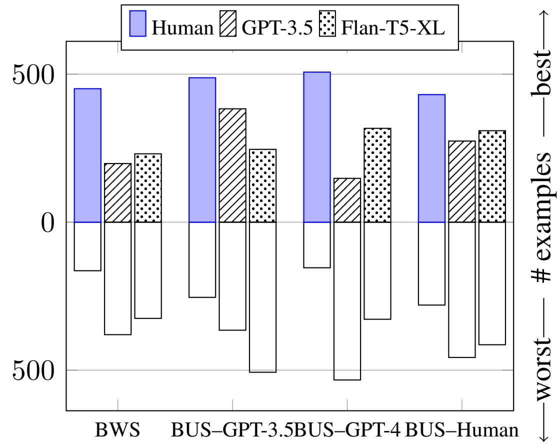


Figure 4.4: Aggregated best-worst votes for human-written, Flan-T5-XL, and GPT-3.5 backgrounds on the test set. The top and bottom halves report voted-best and voted-worst system counts respectively.

4.6.1 BUS–GPT

Setup: Following our human evaluation setup (§4.5.2), we compare human-written, Flan-T5-XL, and GPT-3.5 backgrounds. We compute the percentage of answerable questions using BUS (§4.3.5) and use this score to identify the best and worst systems for each update.

Results: Figure 4.4 provides the best-worst vote counts on the same 1,000 updates from test set using GPT-3.5-based BUS (ref. BUS–GPT-3.5) and GPT-4-based BUS (ref. BUS–GPT-4).[11] For comparison, we also include the vote counts from our human evaluation (§4.5.2; ref. BWS).[12]

With BUS–GPT-3.5, we observe that human-written backgrounds slightly outperform GPT-3.5. Flan-T5 significantly underperforms. BUS–GPT-4 is more closely aligned with our best-worst scaling human evaluation (BWS). This is in line with similar observations from prior work on GPT-4-based evaluation (Liu et al., 2023a).

Overall, BUS–GPT-3.5 and BUS–GPT-4 exhibit different trends for human-written and GPT-3.5 backgrounds. To analyze this discrepancy, we present a BUS–human evaluation that uses question-and-answer pairs compiled by humans.

4.6.2 BUS–Human

Instead of relying on GPT-3.5 (or 4), we use Mechanical Turk to generate question-answer pairs. We first ask annotators to generate five background questions for each of the 1,000 news updates. For each of these tuples of update and questions, we pair it with one of the associated background summaries and ask annotators to attempt to answer all five questions using only information in one of the background summaries (or write *none* if the summary does not contain the answer).

[11] For each example, we use BUS to designate one or more systems as best (or worst).
[12] See subsection 4.8.5 in Appendix for event-level results.

We then calculate BUS–Human as the percentage of answered questions per background summary type.[13] Results are presented in Figure 4.4 (ref. BUS–Human). BUS–Human shows clear alignment with our human evaluation results (BWS) and BUS–GPT-4, illustrating the effectiveness of our proposed BUS metric. However, this also highlights a potential drawback of using an automatic system such as GPT-3.5 for generating question-answer pairs.

4.6.3 Comparison of BUS methods

Questions: We analyze the questions generated by the three variants, BUS–GPT-3.5, BUS–GPT-4, and BUS–Human. In the Appendix, we provide questions generated for example updates from three test events, MH370 flight disappearance (Table 4.10), Yemen crisis (Table 4.11), and Libyan crisis (Table 4.12). Overall, both humans and GPT generate questions that specifically target background knowledge. Turkers' questions are specific and short, while GPT questions are more detailed and often contain two or more sub-questions. Questions target aspects such as named entities (Table 4.10) and past events (Table 4.11, Table 4.12).

However, we also see questions that do not target background information. Some questions from humans and GPT ask for additional details about events described in the update. See Q3 from Turker 2 and Q4 from GPT-3.5 in Yemen crisis (Table 4.11), and Q5 from GPT-4 in Libyan war (Table 4.12). A few questions ask about the consequences of the events described in the update. See Q5 from GPT-4 in MH370 disappearance (Table 4.10), Q4 & Q5 from Turker 2 in Yemen crisis (Table 4.11).

BUS–GPT-3.5 vs BUS–GPT-4: We notice BUS–GPT-3.5 suffers from answer hallucination, i.e., responds with an answer even if its not mentioned in the background text. On the other hand, GPT-4 is better at declining unanswerable questions (Table 4.13 in the Appendix). This explains our observation of better human alignment with BUS–GPT-4.

BUS–Human vs BUS–GPT: Our analysis indicates human evaluation remains the gold standard for our proposed background summarization task (BWS §4.5.2; BUS–Human §4.6.2). GPT-4 presents promising results and could serve as a fast, cost-effective alternative to human evaluation.

Applications: We believe BUS can be extended to related summarization tasks such as TV recaps (Chen et al., 2022) and disentangled summarization of scientific articles (Hayashi et al., 2020). A BUS-like metric can measure the relevancy of the recap to the current TV episode and the paper context to its contributions.

4.7 Conclusion & Future Work

To help readers follow long and complex event timelines, we propose the task of news background summarization. We compliment each update in the timeline with a background summary

[13] Appendix 4.8.3 contains more details about our setup, annotation guidelines, and compensation.

that provides sufficient context to the readers. We present an expert-annotated dataset for this task with over 1,100 background summaries from three annotators. On this dataset, we benchmark a suite of state-of-the-art summarization systems (Flan-T5, LongT5, and GPT-3.5). Our results show the zero-shot GPT-3.5 system outperforms the fine-tuned systems on the factuality metrics while underperforming on ROUGE. Given the lack of a metric that accurately captures the utility of a background summary to the news reader, we propose a novel QA-based metric, BUS, which measures the percentage of questions about the updates that are answerable from the respective background summaries.

For future work, we plan to explore background summarization directly from news articles instead of past updates. Sub-events previously considered unimportant but directly consequential to the latest news update can be captured in this setup. We are also interested in benchmarking aspect-based summarization systems for our task.

4.8 Appendix

4.8.1 Annotation Guidelines for Writing Background Summaries

Table 4.5 presents the guidelines we presented to the annotators who wrote the summaries for our dataset (from subsection 4.3.2). We conducted multiple rounds of training with the annotators, where we reviewed annotator's work and provided feedback on the quality of background summaries.

4.8.2 Details on the MTurk BWS Evaluation

We provide additional details on our Amazon Mechanical Turk setup (from subsection 4.5.2). We give detailed instructions to the annotators, see Figure 4.5. Workers who complete the tasks too quickly are automatically removed from our worker pool; their answers are replaced with new answers. We also use a bonus incentive structure. Every worker who passes the automatic time check receives a bonus at the end. In addition, we only use workers from our pool of about 300 trusted workers from previous studies. These were selected in two stages: (1) We only considered workers from countries whose main language is English and who have completed 100 or more HITs so far with an acceptance rate of 95% or higher. (2) In addition, workers must have passed an initial custom qualification test for a related text classification task we have conducted. Moreover, the resulting pool of workers has been used in more than 50 previous experiments, and we have over time removed any workers who have provided low-quality output in those previous experiments. On our batch of 1,000 HITs for the present human evaluation, we allowed any worker to complete a maximum of 333 HITs so that no worker can dominate the results. We used three annotators per HIT.

Payment: We paid $0.70 per HIT with a bonus of $0.05 for all workers who passed automatic quality checks. 39 workers worked on our HITs overall, spending a median time of 169.0 seconds per HIT. This amounts to an average pay of $14.91 per hour per worker.

4.8.3 Details on the MTurk BUS Evaluation

In order to calculate the BUS metrics based on human-written questions and answers (from subsection 4.6.2), we conducted two separate MTurk evaluations: (1) we obtained questions about news events and (2) we obtained answers to these questions given the different background summaries (human-written, or generated from GPT-3.5 or from Flan-T5). For both evaluations, we used the same general setup and annotator qualifications as described in §4.8.2.

To obtain five background questions for each of 1k news updates, we submitted 1,000 HITs. We paid \$0.75 per HIT with a bonus of \$0.05 for all workers who passed automatic quality checks. 46 workers worked on our HITs overall, spending a median time of 179.0 seconds per HIT. This amounts to an average pay of \$15.08 per hour per worker. The annotation guidelines and an example annotation are shown in Figure 4.6. We allowed any worker to complete a maximum of 333 HITs so that no worker can dominate the results.

To obtain answers to the five questions per news update with respect to the three different background summaries, we submitted 3,000 HITs. We paid \$0.70 per HIT with a bonus of \$0.05 for all workers who passed automatic quality checks. 38 workers worked on our HITs overall, spending a median time of 144.2 seconds per HIT. This amounts to an average pay of \$17.47 per hour per worker. The annotation guidelines and an example annotation are shown in Figure 4.7. We allowed any worker to complete a maximum of 500 HITs.

4.8.4 Experiment Setup

T5-based systems: We perform training using DeepSpeed ZeRO stage 3 on two A6000 GPUs. We train the models for 10 epochs and pick the best model using the ROUGE-L score on the dev set. We use a per-device batch size of 8 and a learning rate of 1e-5. We use the Seq2SeqTrainer from Hugging Face in all of our experiments. At inference time, we use a beam size of 4, a length penalty of 2.0, and a no-repeat ngram size of 3.

GPT-based systems: We use the OpenAI python API for all of our GPT-based systems.

Instructions for BUS–GPT: Table 4.6 presents our instruction templates for question and answer generation using GPT models.

Metrics: ROUGE, BERTScore and QuestEval.[14,15,16]

4.8.5 Additional Results

Event-level BUS: Similar to the results in Figure 4.4, we report the best-worst vote counts per event in the test set. For each event, we report counts for BWS (Figure 4.8), BUS–GPT-3.5 (Figure 4.9), BUS–GPT-4 (Figure 4.10) and BUS–Human (Figure 4.11).

[14] https://hf.co/spaces/evaluate-metric/rouge
[15] https://hf.co/spaces/evaluate-metric/bertscore
[16] https://github.com/ThomasScialom/QuestEval

Terminology

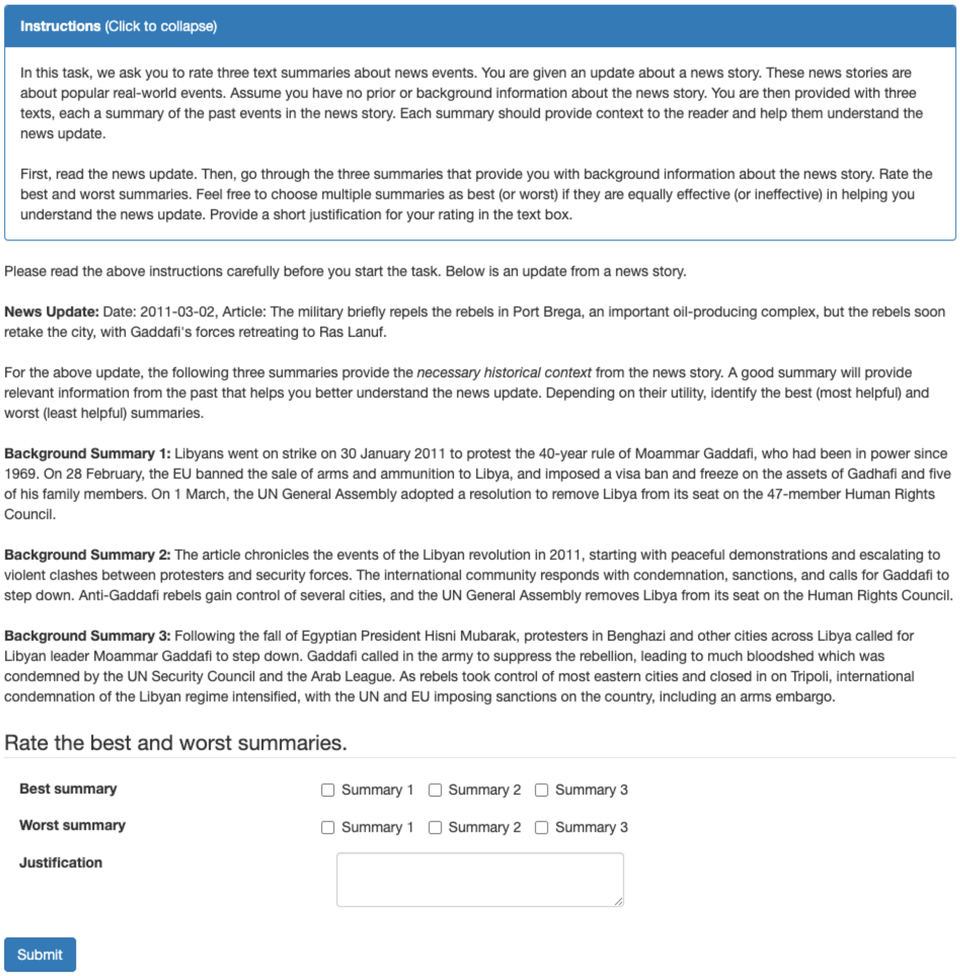
Update: a short text summary of *what's new* in the news story. This text summarizes the latest events, specifically ones that are important to the overall story.
Background: a short text summary that provides *sufficient historical context* for the current update. Background aims to provide the reader a quick history of the news story, without them having to read all the previous updates. Background should cover past events that help in understanding the current events described in the update.
Timestep: day of the event (YYYY-MM-DD).
Timeline: a series of timesteps. Each timestep in a timeline is associated with an update and a background summary.
Super event: the key news story or major event for which we are constructing a timeline. For instance, 'Egyptian Crisis', 'BP oil spill', 'MH 370 disappearance' are some of the super events from our dataset.

Annotation Steps

We follow a three-stage annotation process,
Stage-0: Read the input timeline to get a high-level understanding of the super-event.
Stage-1: For each timestep, read the provided 'rough' update summary. Rewrite the update into a short paragraph, removing any duplicate or previously reported subevents.
Stage-2: Go through the timeline in sequential manner and write background summaries for each timestep.

Table 4.5: Annotation guidelines for the background summarization task.

Query-focused Summarization: In Table 4.3, our query-focused summarization setup did not provide gains. To analyze this behavior, we further experiment with an alternate query format where we use the full update text (U_t) as the query. Table 4.7 presents the results on the Flan-T5 system using ROUGE-L, QuestEval and BUS. We notice only a slight improvement in the performance when using full update text as the query.



Instructions (Click to collapse)

In this task, we ask you to rate three text summaries about news events. You are given an update about a news story. These news stories are about popular real-world events. Assume you have no prior or background information about the news story. You are then provided with three texts, each a summary of the past events in the news story. Each summary should provide context to the reader and help them understand the news update.

First, read the news update. Then, go through the three summaries that provide you with background information about the news story. Rate the best and worst summaries. Feel free to choose multiple summaries as best (or worst) if they are equally effective (or ineffective) in helping you understand the news update. Provide a short justification for your rating in the text box.

Please read the above instructions carefully before you start the task. Below is an update from a news story.

News Update: Date: 2011-03-02, Article: The military briefly repels the rebels in Port Brega, an important oil-producing complex, but the rebels soon retake the city, with Gaddafi's forces retreating to Ras Lanuf.

For the above update, the following three summaries provide the *necessary historical context* from the news story. A good summary will provide relevant information from the past that helps you better understand the news update. Depending on their utility, identify the best (most helpful) and worst (least helpful) summaries.

Background Summary 1: Libyans went on strike on 30 January 2011 to protest the 40-year rule of Moammar Gaddafi, who had been in power since 1969. On 28 February, the EU banned the sale of arms and ammunition to Libya, and imposed a visa ban and freeze on the assets of Gadhafi and five of his family members. On 1 March, the UN General Assembly adopted a resolution to remove Libya from its seat on the 47-member Human Rights Council.

Background Summary 2: The article chronicles the events of the Libyan revolution in 2011, starting with peaceful demonstrations and escalating to violent clashes between protesters and security forces. The international community responds with condemnation, sanctions, and calls for Gaddafi to step down. Anti-Gaddafi rebels gain control of several cities, and the UN General Assembly removes Libya from its seat on the Human Rights Council.

Background Summary 3: Following the fall of Egyptian President Hisni Mubarak, protesters in Benghazi and other cities across Libya called for Libyan leader Moammar Gaddafi to step down. Gaddafi called in the army to suppress the rebellion, leading to much bloodshed which was condemned by the UN Security Council and the Arab League. As rebels took control of most eastern cities and closed in on Tripoli, international condemnation of the Libyan regime intensified, with the UN and EU imposing sanctions on the country, including an arms embargo.

Rate the best and worst summaries.

Best summary	☐ Summary 1	☐ Summary 2	☐ Summary 3
Worst summary	☐ Summary 1	☐ Summary 2	☐ Summary 3
Justification			

Submit

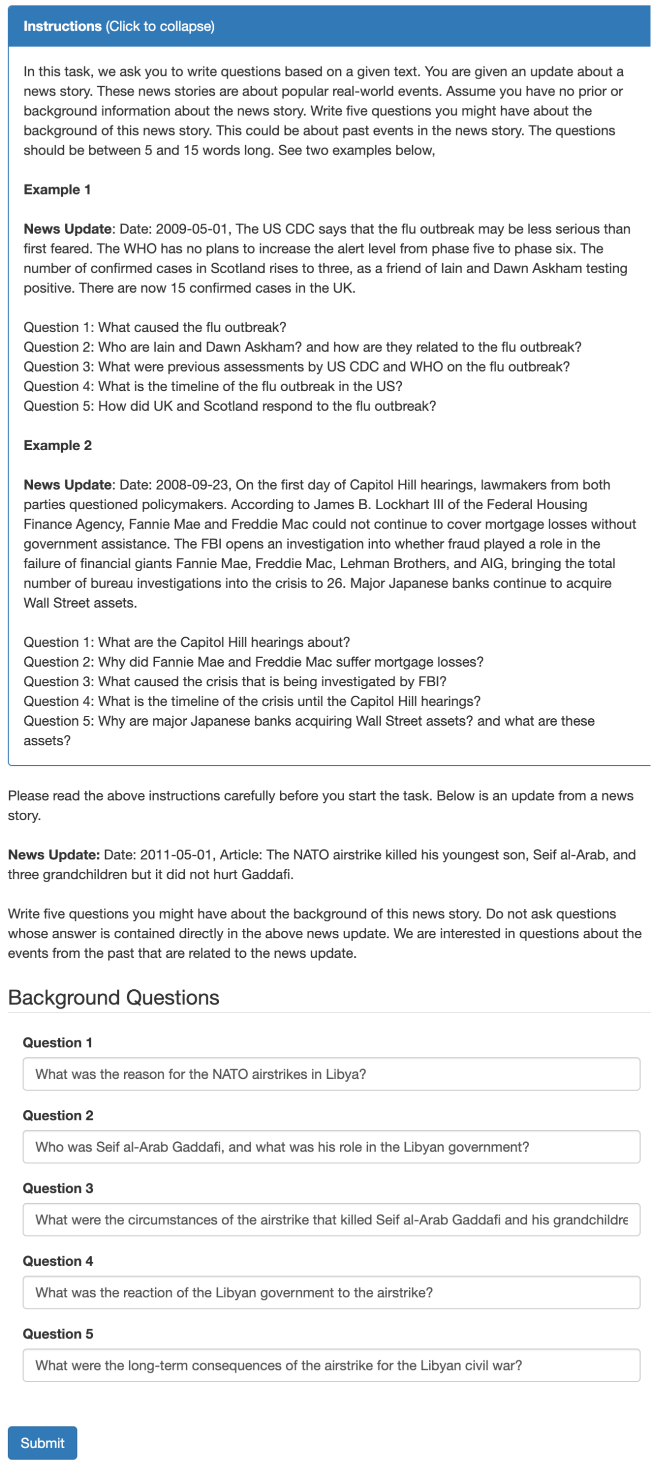
Figure 4.5: This screenshot shows the human annotation interface to determine the best and the worst background summary for for best-worst scaling. In this example, the random order of displayed summaries is Flan-T5-XL, GPT-3.5, followed by the human-written summary. Here, both annotators marked the human-written summary as the best and the GPT-3.5 summary as the worst.

BUS question generation
Update: {update} Imagine you read the above update about a news story. You have no prior information about the story. Generate five background questions you might have about the story.
BUS answer extraction
Background: {background} Questions: {questions} For each question, list answers from the background text when available. Say "unanswerable" if the question is not answered in the background text.

Table 4.6: Instruction templates for GPT-based question-answer generation.

Flan-T5-XL	ROUGE-L	QuestEval	BUS
query: U_t	39.5	30.9	46.9
query: NE(U_t)	39.5	30.8	46.6

Table 4.7: Ablation studies on different queries for background summarization task (dev set).



Instructions (Click to collapse)

In this task, we ask you to write questions based on a given text. You are given an update about a news story. These news stories are about popular real-world events. Assume you have no prior or background information about the news story. Write five questions you might have about the background of this news story. This could be about past events in the news story. The questions should be between 5 and 15 words long. See two examples below,

Example 1

News Update: Date: 2009-05-01, The US CDC says that the flu outbreak may be less serious than first feared. The WHO has no plans to increase the alert level from phase five to phase six. The number of confirmed cases in Scotland rises to three, as a friend of Iain and Dawn Askham testing positive. There are now 15 confirmed cases in the UK.

Question 1: What caused the flu outbreak?
Question 2: Who are Iain and Dawn Askham? and how are they related to the flu outbreak?
Question 3: What were previous assessments by US CDC and WHO on the flu outbreak?
Question 4: What is the timeline of the flu outbreak in the US?
Question 5: How did UK and Scotland respond to the flu outbreak?

Example 2

News Update: Date: 2008-09-23, On the first day of Capitol Hill hearings, lawmakers from both parties questioned policymakers. According to James B. Lockhart III of the Federal Housing Finance Agency, Fannie Mae and Freddie Mac could not continue to cover mortgage losses without government assistance. The FBI opens an investigation into whether fraud played a role in the failure of financial giants Fannie Mae, Freddie Mac, Lehman Brothers, and AIG, bringing the total number of bureau investigations into the crisis to 26. Major Japanese banks continue to acquire Wall Street assets.

Question 1: What are the Capitol Hill hearings about?
Question 2: Why did Fannie Mae and Freddie Mac suffer mortgage losses?
Question 3: What caused the crisis that is being investigated by FBI?
Question 4: What is the timeline of the crisis until the Capitol Hill hearings?
Question 5: Why are major Japanese banks acquiring Wall Street assets? and what are these assets?

Please read the above instructions carefully before you start the task. Below is an update from a news story.

News Update: Date: 2011-05-01, Article: The NATO airstrike killed his youngest son, Seif al-Arab, and three grandchildren but it did not hurt Gaddafi.

Write five questions you might have about the background of this news story. Do not ask questions whose answer is contained directly in the above news update. We are interested in questions about the events from the past that are related to the news update.

Background Questions

Question 1

What was the reason for the NATO airstrikes in Libya?

Question 2

Who was Seif al-Arab Gaddafi, and what was his role in the Libyan government?

Question 3

What were the circumstances of the airstrike that killed Seif al-Arab Gaddafi and his grandchildre

Question 4

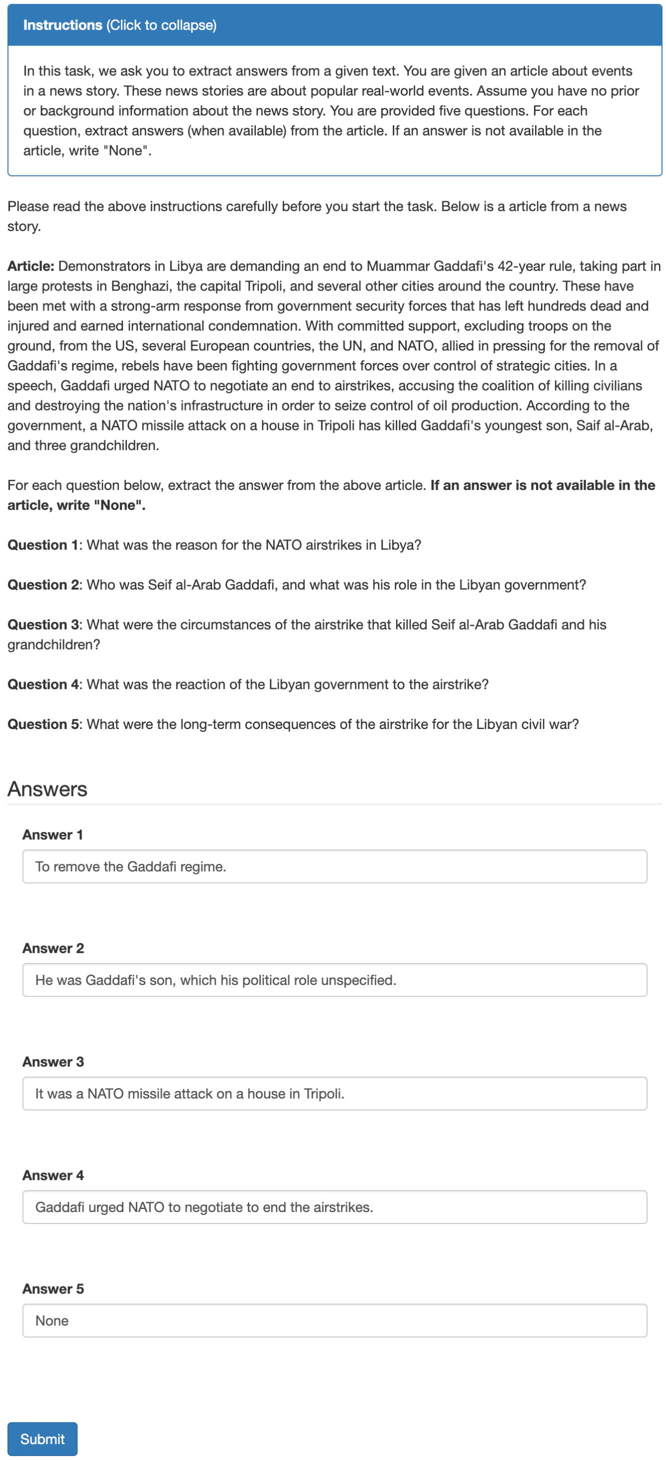
What was the reaction of the Libyan government to the airstrike?

Question 5

What were the long-term consequences of the airstrike for the Libyan civil war?

Submit

Figure 4.6: This screenshot shows the annotation interface for MTurk annotators to write five questions about a news event. The questions from an annotator are shown in the text fields as an example.



Instructions (Click to collapse)

In this task, we ask you to extract answers from a given text. You are given an article about events in a news story. These news stories are about popular real-world events. Assume you have no prior or background information about the news story. You are provided five questions. For each question, extract answers (when available) from the article. If an answer is not available in the article, write "None".

Please read the above instructions carefully before you start the task. Below is a article from a news story.

Article: Demonstrators in Libya are demanding an end to Muammar Gaddafi's 42-year rule, taking part in large protests in Benghazi, the capital Tripoli, and several other cities around the country. These have been met with a strong-arm response from government security forces that has left hundreds dead and injured and earned international condemnation. With committed support, excluding troops on the ground, from the US, several European countries, the UN, and NATO, allied in pressing for the removal of Gaddafi's regime, rebels have been fighting government forces over control of strategic cities. In a speech, Gaddafi urged NATO to negotiate an end to airstrikes, accusing the coalition of killing civilians and destroying the nation's infrastructure in order to seize control of oil production. According to the government, a NATO missile attack on a house in Tripoli has killed Gaddafi's youngest son, Saif al-Arab, and three grandchildren.

For each question below, extract the answer from the above article. **If an answer is not available in the article, write "None".**

Question 1: What was the reason for the NATO airstrikes in Libya?

Question 2: Who was Seif al-Arab Gaddafi, and what was his role in the Libyan government?

Question 3: What were the circumstances of the airstrike that killed Seif al-Arab Gaddafi and his grandchildren?

Question 4: What was the reaction of the Libyan government to the airstrike?

Question 5: What were the long-term consequences of the airstrike for the Libyan civil war?

Answers

Answer 1

To remove the Gaddafi regime.

Answer 2

He was Gaddafi's son, which his political role unspecified.

Answer 3

It was a NATO missile attack on a house in Tripoli.

Answer 4

Gaddafi urged NATO to negotiate to end the airstrikes.

Answer 5

None

Submit

Figure 4.7: This screenshot shows the annotation interface for MTurk annotators to answer five questions about a news updates, given one of the background summaries. The answers from an annotator are shown in the text fields as an example.

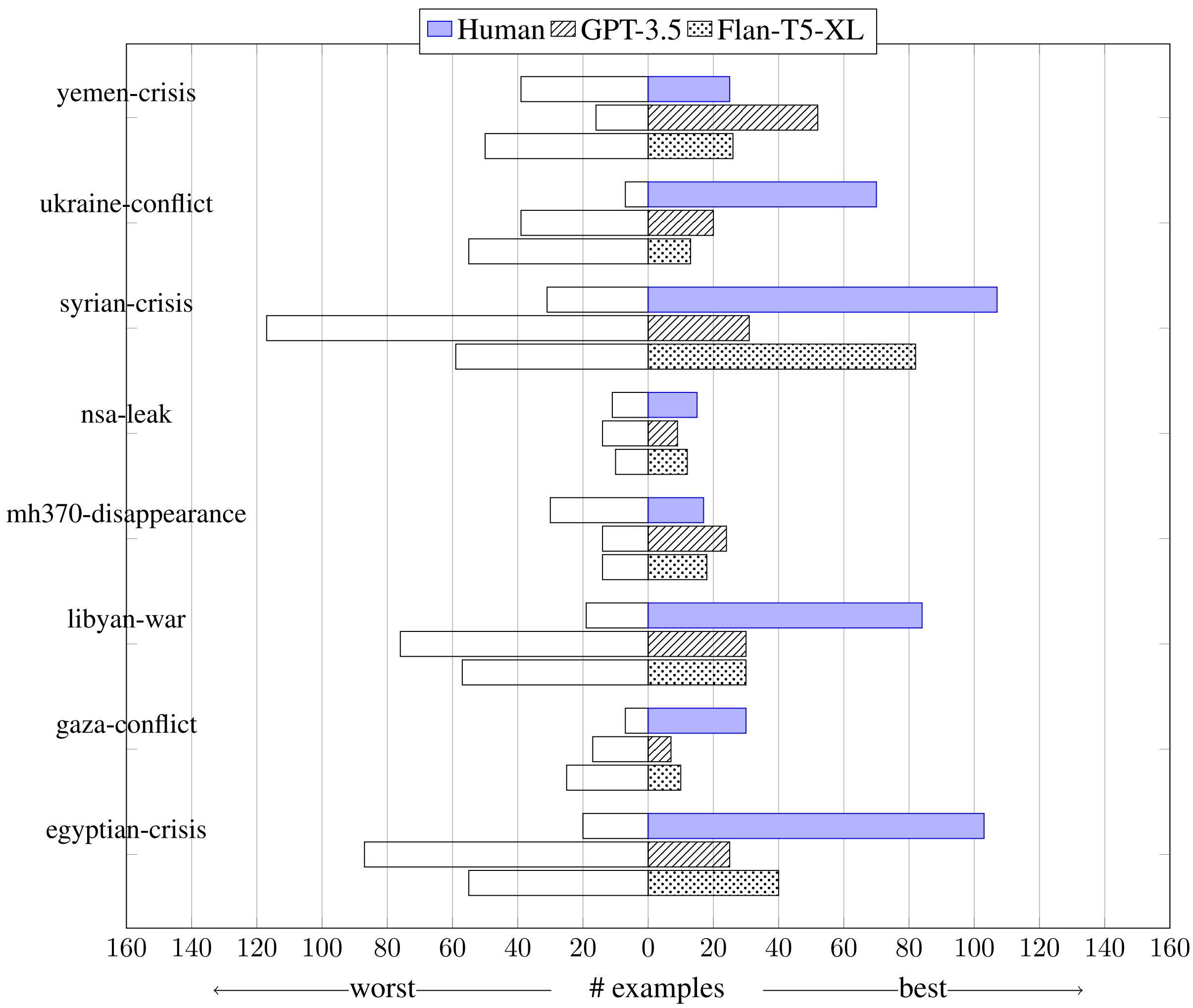


Figure 4.8: Aggregated best-worst votes for human-written, Flan-T5, GPT-3.5 backgrounds (BWS). The left and right halves report voted-worst and voted-best system counts respectively.

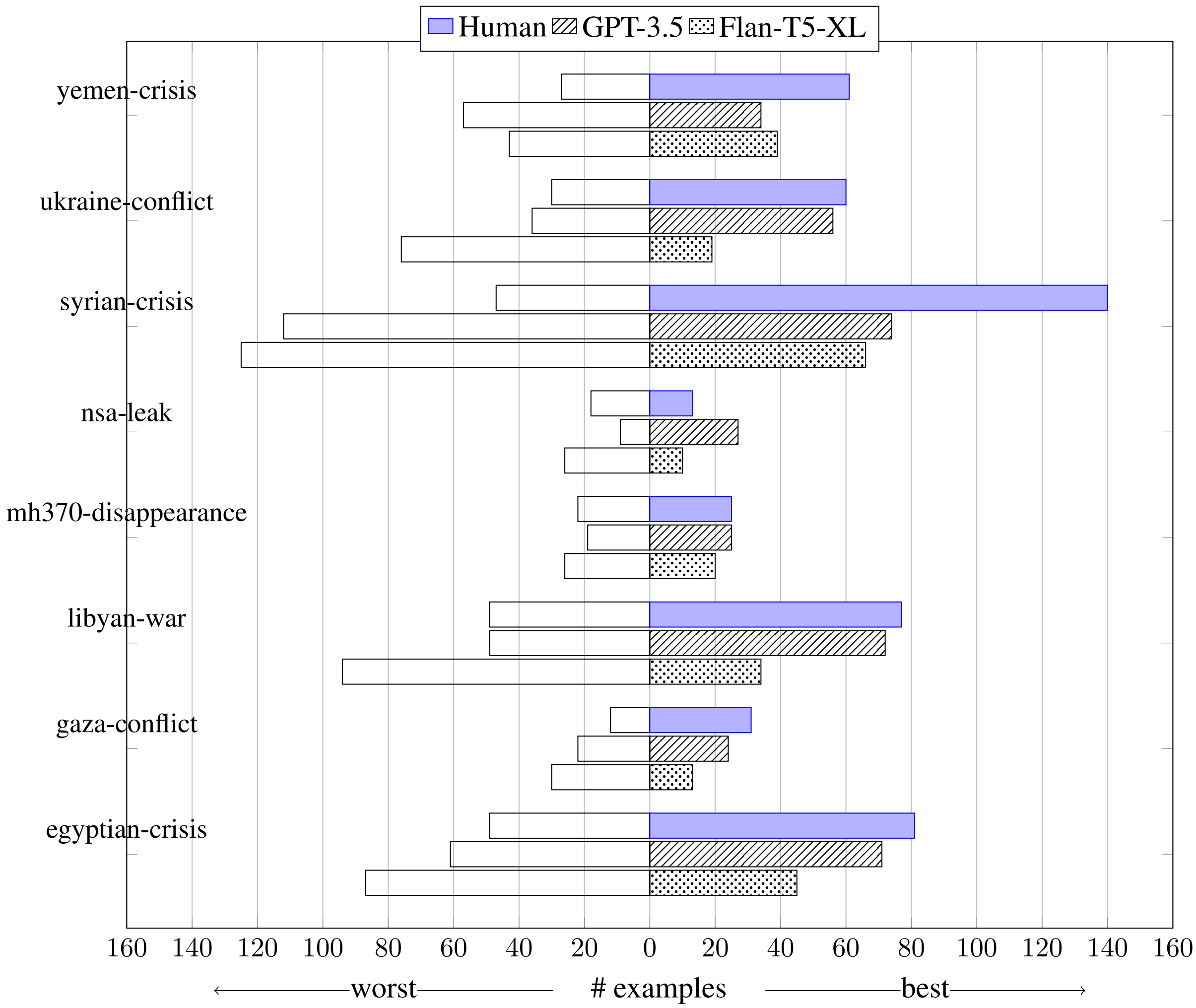


Figure 4.9: Aggregated best-worst votes for human-written, Flan-T5, GPT-3.5 backgrounds (BUS–GPT-3.5). The left and right halves report voted-worst and voted-best system counts respectively.

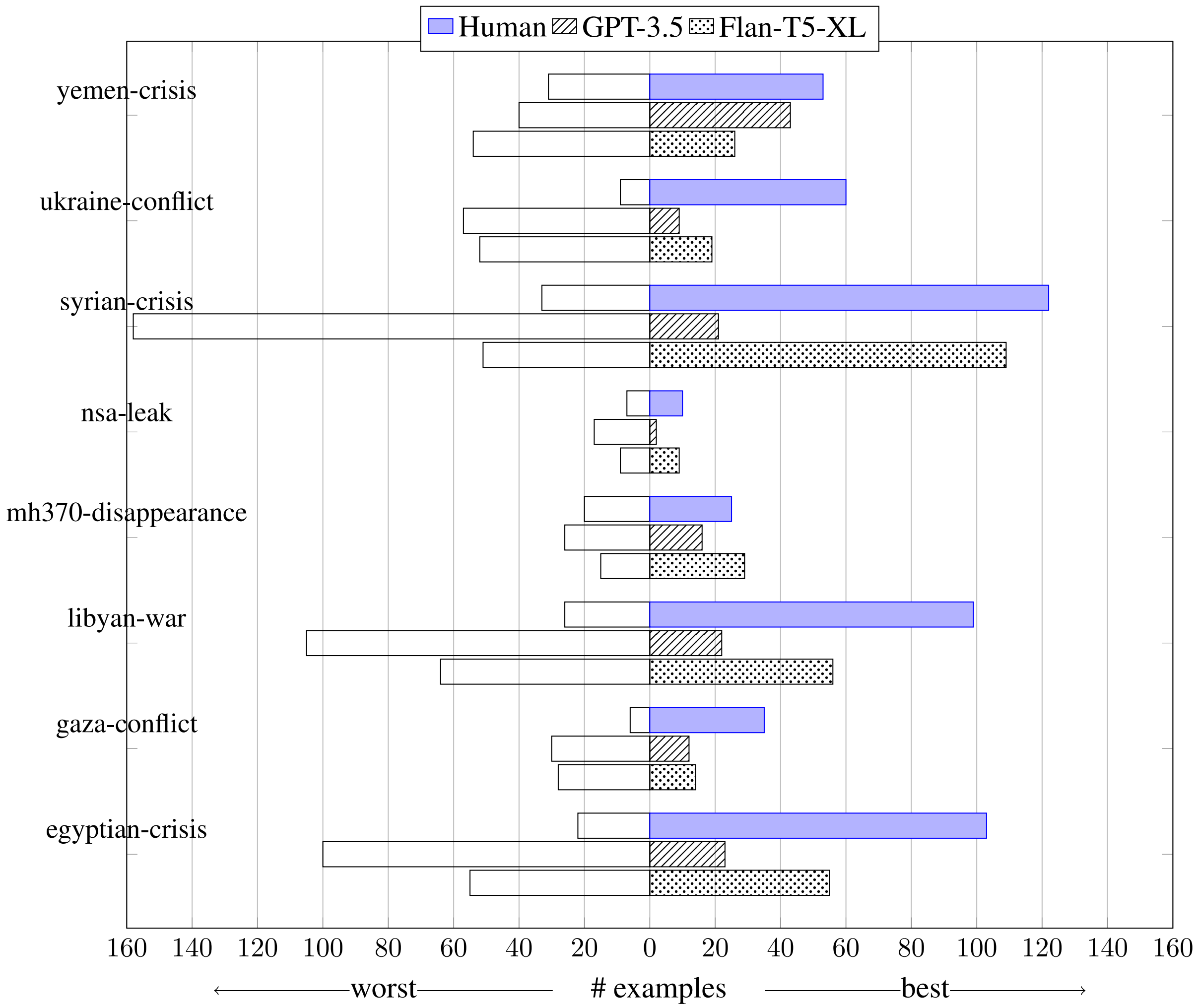


Figure 4.10: Aggregated best-worst votes for human-written, Flan-T5, GPT-3.5 backgrounds (BUS–GPT-4). The left and right halves report voted-worst and voted-best system counts respectively.

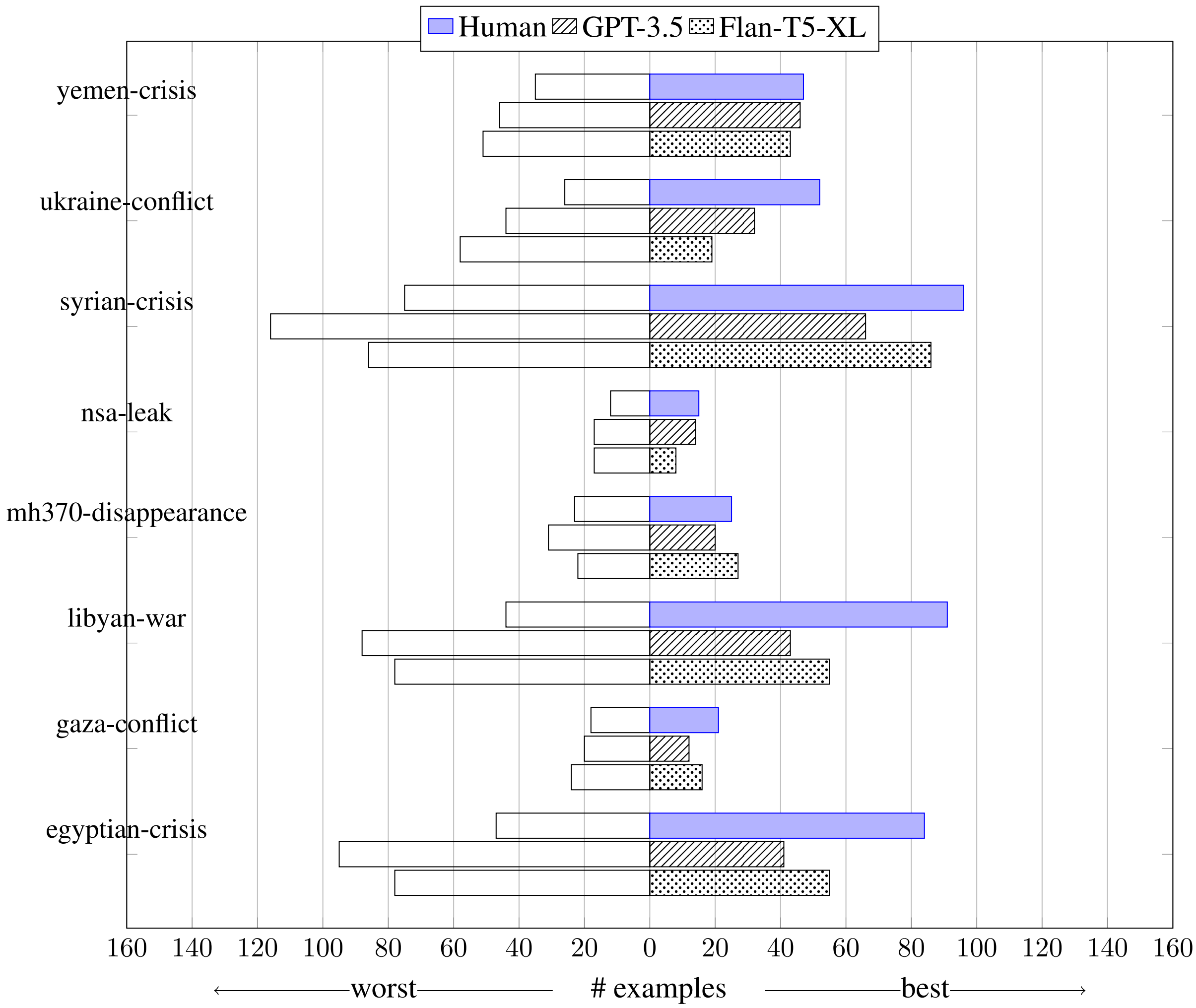


Figure 4.11: Aggregated best-worst votes for human-written, Flan-T5, GPT-3.5 backgrounds (BUS–Human). The left and right halves report voted-worst and voted-best system counts respectively.

Date: April 24, 2010. *Update*: For the first time, oil is found to be leaking from the well. Pressure and release rate are unknown. A homeland security report on critical infrastructure says the problem has "no near-term impact to regional or national crude oil or natural gas supplies."
Background: (Human) On April 20th, BP-operated drilling rig Deepwater Horizon exploded in the Gulf of Mexico, 84km south-east of Venice, Louisiana, when a blowout preventer failed to activate, leaving 11 dead and 17 injured. The rig has been found upside down about a quarter-mile from the blowout preventer. Multiple federal departments and agencies are involved in 28 search-and-rescue operations covering about 5,300 square miles. Oil sheen has been reported, but the Coast Guard says it's had no indication that oil is leaking from the well 5,000 ft below the surface of the Gulf. nA homeland security department risk analysis has said the incident "poses a negligible risk to regional oil supply markets". White House press secretary Robert Gibbs played down any possible effects on President Obama's plans to expand offshore drilling, saying "I don't honestly think it opens up a whole new series of questions because, you know, in all honesty I doubt this is the first accident that has happened and I doubt it will be the last".
Background: (GPT-3.5) On April 20, 2010, the Deepwater Horizon drilling rig exploded in the Gulf of Mexico, leaving 11 workers missing and presumed dead, and injuring approximately 17. The blowout preventer failed to activate, and the rig sank to the bottom of the Gulf, potentially allowing oil to spew out from the well. The National Response Team, involving 16 federal departments and agencies, was activated, and chemical dispersants were readied to break up the oil slick. The US coast guard suspended the search for the missing workers and reported no indication of oil leaking from the well. The incident was deemed to pose a negligible risk to regional oil supply markets and was not expected to cause significant national economic impacts.
Background: (Flan-T5-XL) The Deepwater Horizon drilling rig exploded on April 20, 2010 while under construction in the Gulf of Mexico, killing 11 workers. The rig was found upside down about a quarter-mile from the blowout preventer. The force of the explosion broke off the rig's drillpipe, potentially allowing oil to spew out into the gulf. A National Response Team, involving 16 federal departments and agencies including the White House, Coast Guard, the departments of Defense, Homeland Security and Interior, and Environmental Protection Agency, among others, was activated. A Coast Guard official said the Macondo well, which the rrig had been drilling, could be releasing up to 8,000 barrels of oil per day. A remotely operated vehicle with camera went underwater and confirmed no visible oil flow from the well, but there were reports of an oil slick five miles long. Chemical dispersants were readied to break it up. A homeland security department risk analysis said the incident "poses a negligible risk to regional oil supply markets and will not cause significant national economic impacts". White House press secretary Robert Gibbs was asked whether the incident might affect Obama's plans to expand offshore drilling. He said, "I don't honestly think it opens up a whole new series of questions because, you know, in all honesty I doubt this is the first accident that has happened and I doubt it will be the last"

Table 4.8: This table provides the background summaries generated by humans, GPT-3.5, and Flan-T5 for the given update. See Figure 4.2 for the QA pairs generated from these update-background pairs by prompting `gpt-3.5-turbo`.

	ROUGE-1	ROUGE-2	ROUGE-L	QuestEval	BERTScore P	BUS–GPT-3.5
Flan-T5-XL	43.5	20.4	39.9	31.2	86.3	46.0
Long-T5-TGlobal-XL	40.1	16.7	36.4	33.4	86.9	46.2

Table 4.9: A comparison of Flan-T5 and Long-T5 systems on the dev set.

Date: June 4, 2014.
Update: Australian researchers release a recording of an underwater sound that could have been MH370 hitting the water.

BUS–human (Turker 1)
Q1: What is MH370?
Q2: What was the name of the researchers?
Q3: Where was the crash?
Q4: What else could the sound have been?
Q5: How did they record the sound?

BUS–human (Turker 2)
Q1: What was the flight path of MH370?
Q2: What were the last known communications from MH370?
Q3: What are the search parameters being used by the Australian researchers?
Q4: What are the other possible explanations for the underwater sound?
Q5: What are the implications of the underwater sound for the search for MH370?

BUS–GPT-3.5
Q1: What is MH370 and why is it significant?
Q2: How did the Australian researchers obtain the recording of the underwater sound?
Q3: What other evidence has been found regarding the disappearance of MH370?
Q4: What is the current status of the investigation into the disappearance of MH370?
Q5: What impact could this new evidence have on the families of the passengers and crew on board MH370?

BUS–GPT-4
Q1: What is MH370 and what happened to it?
Q2: Who are the Australian researchers involved in this investigation?
Q3: How were the researchers able to capture this underwater sound?
Q4: What evidence suggests that this sound could be MH370 hitting the water?
Q5: Has this new evidence brought any significant progress in the investigation of MH370?

Table 4.10: For an update from the ‘MH370 flight disappearance’ event, this table provides BUS questions generated by humans (MTurk), GPT–3.5 and GPT-4.

Date: November 23, 2011.
Update: President Saleh signs a deal brokered by the Gulf Cooperation Council which will see him hand over power to his vice president in exchange for immunity from prosecution for him and his family. He will remain honorary president for 90 days.

BUS–human (Turker 1)
Q1: What is Saleh president of?
Q2: What is the Gulf Cooperation Council?
Q3: What crimes is Saleh immune to prosecution for?
Q4: Who is Saleh's vice president?
Q5: What else was in the deal that was signed?

BUS–human (Turker 2)
Q1: What were the reasons for the protests against President Saleh?
Q2: What was the role of the Gulf Cooperation Council in mediating the deal?
Q3: What were the terms of the deal that President Saleh signed?
Q4: What was the reaction of the Yemeni people to the deal?
Q5: What are the long-term implications of the deal for Yemen?

BUS–GPT-3.5
Q1: What led to President Saleh's decision to hand over power to his vice president?
Q2: Who is President Saleh's vice president and what is his political background?
Q3: What is the Gulf Cooperation Council and how did they broker this deal?
Q4: What kind of immunity will President Saleh and his family receive and what crimes are they being protected from?
Q5: What will happen after the 90-day period when President Saleh's honorary presidency ends?

BUS–GPT-4
Q1: Who is President Saleh and what country is he the president of?
Q2: What led to the need for President Saleh to hand over power to his vice president?
Q3: What is the Gulf Cooperation Council and what role did they play in brokering this deal?
Q4: Why is immunity from prosecution a part of the deal for President Saleh and his family?
Q5: Who is the vice president that will be taking over power and what is their political background?

Table 4.11: For an update from the 'Yemen crisis' event, this table provides BUS questions generated by humans (MTurk), GPT–3.5 and GPT-4.

Date: November 22, 2011. *Update*: Libya's interim prime minister Abdel Rahim al-Keeb announces a new cabinet.
BUS–human (Turker 1) Q1: What led to the need for a new cabinet announcement in Libya? Q2: Who was the previous prime minister ? Q3: What challenges did the interim government face in forming the new cabinet? Q4: What are the key responsibilities and goals of the new cabinet ? Q5: How was the new interim prime minister selected? *BUS–human* (Turker 2) Q1: What happened to Libya's previous prime minister? Q2: Can an interim prime minister create a new cabinet? Q3: Do the people of Libya like Abdel Rahim al-Keeb? Q4: When will there be a new permanent prime minister? Q5: What happened to Libya's previous cabinet?
BUS–GPT-3.5 Q1: Who is Abdel Rahim al-Keeb and how did he become Libya's interim prime minister? Q2: What were the reasons for the formation of a new cabinet in Libya? Q3: Who are the members of the new cabinet and what are their qualifications and experience? Q4: What are the main challenges facing the new cabinet in Libya and how do they plan to address them? Q5: What is the current political situation in Libya and how does the new cabinet fit into the broader political landscape?
BUS–GPT-4 Q1: Who was the previous prime minister of Libya and why did they leave the position? Q2: What led to Abdel Rahim al-Keeb becoming the interim prime minister? Q3: What is the political situation in Libya that necessitated a new cabinet? Q4: What are the main challenges facing the new cabinet in Libya? Q5: Who are the members of the new cabinet and what are their political backgrounds?

Table 4.12: For an update from the 'Libyan crisis' event, this table provides BUS questions generated by humans (MTurk), GPT–3.5 and GPT-4.

Date: June 16, 2012. *Update*: Escalating violence leads to the suspension of the UN observer mission in Syria. *Background (GPT-3.5)*: The article provides a timeline of events during the Syrian uprising from May 2011 to May 2012. It includes the imposition of sanctions on Assad and senior members of the government by the EU, violent clashes and protests, defections, and the Arab League peace plan. The article also covers the UN-brokered cease-fire and the presence of UN observers, as well as the continued violence and massacres that occurred during this period.
BUS–GPT-3.5 Q1: What was the purpose of the UN observer mission in Syria? Q2: What is the current level of violence in Syria and who is responsible for it? Q3: How long has the UN observer mission been in Syria and what progress have they made? Q4: What actions have been taken by the international community to address the violence in Syria? Q5: What are the potential consequences of suspending the UN observer mission in Syria? A1: The purpose of the UN observer mission in Syria was to monitor the implementation of a UN-brokered cease-fire. A2: The article does not provide information on the current level of violence in Syria or who is responsible for it. A3: The UN observer mission has been in Syria since April 2012, and the article does not provide information on the progress they have made. A4: The international community has imposed sanctions on Assad and senior members of the government by the EU, and the Arab League has proposed a peace plan. The UN has also brokered a cease-fire and sent observers to monitor it. A5: The potential consequences of suspending the UN observer mission in Syria are not discussed in the background text.
BUS–GPT-4 Q1: What is the UN observer mission in Syria and what is its purpose? Q2: What specific events led to the escalation of violence in Syria? Q3: How long has the UN observer mission been in Syria? Q4: What are the potential consequences of suspending the UN observer mission in Syria? Q5: Who are the main parties involved in the violence in Syria? A1: Unanswerable A2: Unanswerable A3: Unanswerable A4: Unanswerable A5: Unanswerable

Table 4.13: For an update from the ‘Syrian Crisis’ event, this table provides BUS questions and answers extracted GPT–3.5 and GPT-4. BUS–GPT-3.5 suffers from answer hallucination, i.e., responds with an answer even if its not mentioned in the background text. See A1, A3 in BUS–GPT-3.5.

Chapter 5

Scaling Multi-Document Event Summarization: Evaluating Compression vs. Full-Text Approaches

This work was published in NAACL 2025 (Pratapa and Mitamura, 2025b).[1]

5.1 Introduction

Summarizing events described in document collections has long interested the NLP community with shared tasks for event tracking (Allan et al., 1998) and summarization (Chieu and Lee, 2004; Dang and Owczarzak, 2009; Aslam et al., 2015). Given an input collection of hundreds of text documents, systems have to extract and summarize salient information about the event. The length and diversity of the input presents a challenge to recent large language models (LLMs). In this work, we contrast two classes of systems for large-scale multi-document summarization (MDS), compression-based, and full-text systems.[2]

Full-text systems promise a lossless approach by providing the summarizer access to the entire input. They are based on the long-context reasoning abilities of LMs, having already shown strong retrieval performance on long inputs (Hsieh et al., 2024). However, their capabilities on large-scale MDS are not as well understood. In a recent work, Laban et al. (2024b) introduced a synthetic MDS benchmark that resembles the Needle in a Haystack evaluation (Kamradt, 2023). In addition to this dataset, we evaluate on two large-scale event summarization datasets: Background (Pratapa et al., 2023) and WCEP (Gholipour Ghalandari et al., 2020). We contrast the end-to-end full-context method[3] with three compression-based methods: retrieval, hierarchical, and incremental. Each method *compresses* the input in a multistage pipeline (§5.2.2). We evaluated the content selection aspects of the summary using the Atomic Content Unit (A3CU) metric (Liu et al., 2023c).

[1] Code and data are available at: `https://github.com/adithya7/scaling-mds`.

[2] We use the term *scale* to refer to the large number of documents associated with each summary.

[3] We use full-text and full-context interchangeably.

Our experiments show that full-context and retrieval perform best in most settings (§5.3). To better understand the performance of compression-based methods, we measure A3CU recall to track the salient information retention in their intermediate outputs (§5.3.4). Across all settings, we find that compression-based methods show high recall in intermediate stages but suffer information loss in their multistage pipeline. In particular, the intermediate recall is often much higher than the full-context system recall. We highlight two key takeaways: First, while iterative methods (hierarchical & incremental) were previously found effective for book summarization and small-scale MDS, they underperform on large-scale MDS. Second, full-context systems are suboptimal on large-scale MDS datasets. We advocate for hybrid methods that combine input compression and long-context models. Such hybrid approaches are also scalable to even larger MDS tasks that go far beyond the context window limits of current LLMs.

5.2 Experimental Setup

5.2.1 Datasets

Our three datasets provide different flavors of the multi-document summarization task (Table 5.1).

SummHay: A query-focused dataset that covers the news and conversation domains (Laban et al., 2024b). Synthetically generated using GPT-3.5 and GPT-4o, each summary constitutes a set of insights. To keep our evaluation setup consistent across datasets, we concatenate these insights into a free-form summary. Following the original work, we include an oracle setting that only retains documents containing the reference insights.

Background: This dataset provides summaries of complex news events (Pratapa et al., 2023). The task is based on an event timeline. For a given day, the goal is to generate a background summary by summarizing past new articles related to the event. We expand the original dataset to use news articles instead of just news updates. The dataset includes three human-written background summaries.

WCEP: A newswire dataset collected from Wikipedia Current Events Portal (Gholipour Ghalandari et al., 2020). The summaries come from the portal and the documents include a combination of cited source articles and a retrieved collection of related articles from the Common Crawl archive.

Our choice of datasets collectively represents the real-world use-cases of multi-document summarization systems. Previous work has shown the effectiveness of full-context methods in retrieval tasks. To this end, we include the query-focused SummHay dataset. On the other hand, Background and WCEP provide different variants of the task. Background task requires accumulation of salient content units over the entire input. WCEP has high information redundancy, with many articles providing support for the salient units.

5.2.2 Methods

We now describe our long-context methods and transformers. The key difference between our methods is the length of the input passed to the summarization system (transformer) at any stage.

Dataset	# Ex.	# Docs/Ex.	Avg. length Doc.	Summ.
SummHay	92	100	884	185
Background	658	186	1033	174
WCEP	1020	76	468	34

Table 5.1: An overview of our multi-document summarization datasets. We report the number of examples in the test set, and average statistics for # documents per example, document and summary lengths (words).

	Llama-3.1-8B	Llama-3.1-70B	Command-R	Jamba-1.5-Mini
SummHay	-53% -32% +4% 33.9	-44% -37% +17% 31.1	-63% -63% 0% 30.4	-54% -32% +6% 32.9
SummHay (oracle)	-27% -17% +1% 37.1	-35% -29% -6% 41.8	-33% -47% -3% 32.6	-16% -18% +3% 35.1
Background	-16% -36% +8% 15.6	-9% -31% -3% 16.1	-10% -6% +15% 10.3	-17% -15% -11% 12.3
WCEP	-15% -22% -3% 30.7	-13% -22% -2% 31.1	-11% -13% -1% 28.9	-2% -13% -1% 29.3

Table 5.2: Performance of hierarchical, incremental and retrieval methods relative to the full-context baseline.

Full-context: The transformer has access to the full input and relies on its long context reasoning abilities to generate the summary.

Iterative: Multi-stage summarization where we iteratively pass chunks of the input to the transformer. We explore two methods, hierarchical and incremental. The hierarchical method summarizes each document and iteratively merges these to compile the final summary. The incremental method processes documents in order while maintaining a running summary of the input. Previous work explored these methods for book summarization (Chang et al., 2024) and small-scale multi-document summarization (Ravaut et al., 2024).

Retrieval: We rank the input documents according to their relevance to the query.[4] We then select the top-ranked documents (up to 32k tokens) and pass their concatenation to the transformer. We use SFR Embedding-2 (Meng* et al., 2024) for the retrieval task and order-preserving RAG following the recommendation from Yu et al. (2024a). We set 32k as the limit because all of our transformers are effective at this context length (Hsieh et al., 2024).

[4] If a query is unavailable, we default to using 'Generate a summary of the document' as the query.

5.2.3 Transformers

For our summarization systems, we experiment with three transformer-based models, Llama-3.1, Command-R, and Jamba-1.5. Each model supports a context window of at least 128k tokens. They rely on a different long-context methodologies, and represent the broad class of open-weight LLMs. All the three models show competitive performance on the RULER benchmark for long-context LMs (Hsieh et al., 2024).

Llama-3.1: Pretrained on 15T+ tokens, it supports long context by using a large base frequency of 500,000 and non-uniform scaling of RoPE dimensions (Meta, 2024). We use both 8B and 70B variants to test the effect of model scaling.

Command-R: A transformer-based model that uses NTK-aware interpolation with a very large RoPE base frequency of 4M (Cohere For AI, 2024). We use the 32B variant.

Jamba-1.5: A hybrid architecture with interleaved Transformer and Mamba layers (Team et al., 2024a). It involves both mid-training on long texts and post-training on (synthetic) long-context tasks. We use the 52B Jamba-1.5-Mini mixture-of-experts model with 12B active parameters.

For a fair comparison of above methods and transformers, we set the maximum input length to 128k across all settings. If the input is longer than 128k tokens, we first truncate the longest documents. In the case of Background, we also ensure equal representation from the past events by budgeting the token limit to each past timestamp. We also set a minimum document length (128 tokens) and drop documents if this cannot be achieved. To ensure that all methods see the same input, we adopt the same truncation strategy across full-text and compression-based methods. Theoretically, compression-based methods could work with even longer input (>128k), but we limit all settings to 128k tokens for a fair comparison.

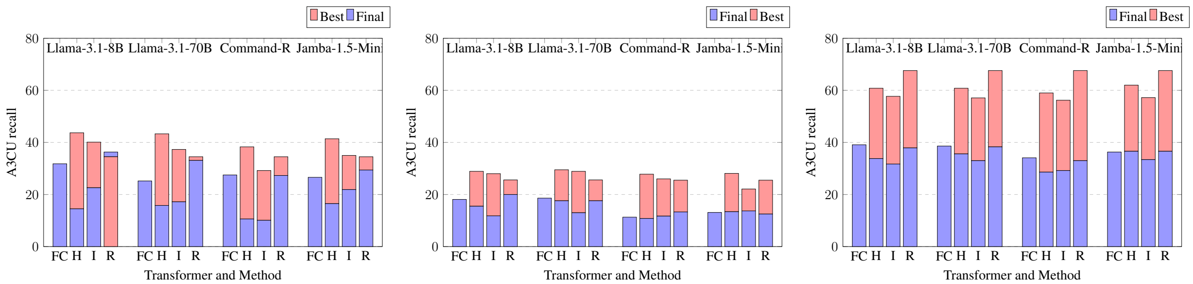
See §5.6.2 in the Appendix for additional details about our experimental setup including our summarization prompt (Table 5.4). We sample summaries with a temperature of 0.5. We note that the summaries could be slightly different across different seeds. Vig et al. (2022) compared end-to-end and RAG for query-focused summarization, but limited to the short input setting.

5.3 Results

5.3.1 Metrics

We focus our analysis on the *content selection* aspect of summarization. Nenkova and Passonneau (2004) first studied the content selection evaluation using the pyramid method on summarization of content units. Follow-up efforts have automated various parts of this method (Shapira et al., 2019; Liu et al., 2023c). In this work, we use the reference-based Atomic Content Unit (A3CU) metric (Liu et al., 2023c) that is based on the definition of atomic content units of Liu et al. (2023b). This metric is trained to predict a score that measures the overlap of atomic content units between the reference and predicted summaries.

Recent works also studied faithfulness (Kim et al., 2024a), coherence (Chang et al., 2024), and position bias (Huang et al., 2024; Ravaut et al., 2024; Laban et al., 2024b). Although these evaluations are important, content selection remains a core issue for large-scale MDS.



(a) SummHay (b) Background (c) WCEP

Figure 5.1: Salient information retention in the intermediate and final summaries (A3CU *recall*). For each compression method, we report the best recall from the intermediate outputs and the recall of the final summary. (H: hierarchical, I: incremental, R: retrieval, FC: full-context)

5.3.2 Overall Results

Table 5.2 reports the A3CU F1 scores for compression-based methods relative to the full-context baseline.[5] Full-context and retrieval perform the best, being particularly effective on the query-focused SummHay dataset. The two iterative methods perform poorly in most settings. We also find that the performance of transformers and methods varies considerably across the datasets and even within examples in each dataset.[6] Below, we break down these results and analyze the effect of transformer and compression methods.

Due to the high costs of running API-based models on long texts, we mostly limit our evaluation to open-weight LLMs. We report preliminary results using Gemini-1.5 on SummHay in Table 5.10 in the Appendix. We noticed trends similar to those of open-weight LLMs.

5.3.3 Analysis: Full-context & Transformer

In the full-context setting, we see mixed results across transformers, with none performing the best across all datasets. Interestingly, Llama-3.1-8B outperforms 70B on SummHay. This surprising result aligns with their relative performance on the RULER benchmark at 128k context length. The 70B model fares better in the oracle setting and shows similar performance on non-retrieval-style datasets. We believe that the 70B model needs additional post-training to improve its long-context retrieval performance.

Command-R underperforms the much smaller Llama-3.1-8B. This could be attributed to its use of RoPE (Su et al., 2021). Command-R increases the base frequency while Llama-3.1 additionally scales RoPE dimensions non-uniformly, likely leading to better long-context capabilities (Ding et al., 2024). However, without specific details on the mid- and post-training with long texts, it would be difficult to identify the exact cause. We direct the reader to Peng et al. (2023) and Lu et al. (2024) for a discussion on long-context methods.

[5] We report ROUGE and A3CU precision, recall in §5.6.3.

[6] See Figure 5.3 in the Appendix for example-level trends.

5.3.4 Analysis: Full-context vs. Compression

With the exception of retrieval on query-focused SummHay dataset, compression-based methods generally underperform full-context (Table 5.2). To analyze this, we use A3CU *recall* to track the retention of salient information in intermediate outputs. These intermediate outputs correspond to the retrieved documents (retrieval) and intermediate summaries (hierarchical, incremental). Figure 5.1 reports the recall scores for the final summary and the best intermediate output (excl. final). For comparison, we also report the recall score for the full-context summary. Across datasets, the best intermediate recall is significantly higher than the final summary recall, even outperforming full-context.[7]

We highlight two key observations. First, iterative methods suffer catastrophic information loss in their multistage pipeline. Second, the best intermediate recall scores from compression methods show areas of improvement for full-context systems. As a control setting, we evaluated on SummHay-oracle and found full-context to be comparable to the best intermediate recall from compression methods (Figure 5.2 in the Appendix).

Retrieval: Relative performance of full-context and retrieval varies widely across examples and transformers. Karpinska et al. (2024a) observed similar behavior for claim verification on books. In particular, for Llama-3.1-8B on SummHay, we find the final summary to be better than the best intermediate output (Figure 5.1). This is the optimal scenario, illustrating the system's effectiveness in aggregating information from the retrieved documents. We do not see this behavior in other settings.

Iterative: We qualitatively analyze the outputs from iterative methods. The hierarchical method tends to generate increasingly abstract summaries at higher levels. It often skips details such as entities and numerals in the summaries. We observe this behavior across all transformers. With the incremental method, we attribute poor performance to the large number of intermediate steps (# documents). Even though the system retrieves salient information at an intermediate stage, the model often gets distracted by non-salient information seen in documents thereafter. We provide examples in Table 5.15 and Table 5.16 in the Appendix.

In the Appendix (§5.6.5), we also experiment with short-context transformers such as Llama-3 (Table 5.11), varying chunk sizes for the hierarchical method, an alternative embedding method for retrieval (Table 5.13), and grounded generation templates for Jamba and Command-R.

5.3.5 Human Evaluation

To complement our automatic evaluation, we perform a reference-based human evaluation. We randomly sample 62 examples from the SummHay dataset (≈67%) and ask a human expert[8] to rate the system summaries. We follow recommendations from prior work (Kiritchenko and Mohammad, 2017; Goyal et al., 2022a; Pratapa et al., 2023) to use the best-worst rating scale. For each example, the human evaluator picks the best and worst summaries (multiple allowed) among

[7]Since recall is impacted by the summary length, we report average length of summaries for each system in Table 5.9 in the Appendix. We do not find any noticeable correlation.

[8]This task was done by the first author.

Transformer	Method	Best	Worst
Llama-3.1-8B	Full-Context	28	10
Llama-3.1-8B	Hierarchical	13	44
Llama-3.1-8B	Incremental	18	21
Llama-3.1-8B	Retrieval	45	4

Table 5.3: Best-worst ratings from human evaluation on a random sample of 62 examples from SummHay. We report the counts for number of times a system was rated the best or worst amongst the four summaries. We compare each system summary against the reference.

the four methods, full context, hierarchical, incremental, and retrieval (Llama-3.1-8B). They use reference summaries to perform content selection evaluation. We shuffle the presentation order of the system summaries in each example, and system labels are completely hidden from the human evaluator. The results of our human evaluation are presented in Table 5.3. Retrieval-based summaries are rated the best, followed by full-context, incremental, and hierarchical. These results strongly correlate with our automatic evaluation (Table 5.2).

5.3.6 Recommendations for Future Work

Based on our analysis, we make two recommendations for future work on large-scale MDS. First, hybrid systems that combine input compression methods with long-context LLMs. Second, a reference-free content selection evaluation that facilitates further scaling of MDS.

Hybrid Methods: Our analysis using A3CU recall shows the scope for improvement of full-context systems (Figure 5.1). Recent studies have shown that long-context models are not as effective as claimed for retrieval tasks (Hsieh et al., 2024; Karpinska et al., 2024a), and our results support this for large-scale MDS. Iterative methods were previously used for book summarization (Chang et al., 2024) and small-scale MDS (Ravaut et al., 2024). In large-scale MDS, they show a significant loss of salient information. Based on these observations, we advocate for a hybrid approach that utilizes selective input compression methods (Sarthi et al., 2024; Xu et al., 2024; Jiang et al., 2024) in conjunction with a long-context LLM. A hybrid approach could provide optimal performance while improving the runtime over full-context. It also allows for scaling to a very large-scale MDS that goes far beyond the model context window.

Reference-free evaluation: In our analysis, we used a reference-based A3CU metric. As we scale the MDS task to include hundreds or thousands of documents, obtaining high-quality human-written reference summaries will be infeasible. Therefore, reference-free content selection evaluation metrics are needed. Synthetic tasks such as SummHay present a promising alternative.

5.4 Conclusion

In this work, we contrast the full-context method against three compression-based methods for large-scale MDS. We evaluated on three datasets, SummHay, Background, and WCEP using the A3CU content selection evaluation metric. We find that the full-context and retrieval-based methods perform the best. Iterative methods suffer from significant information loss. Our analysis shows that full-context methods provide suboptimal performance, and we recommend future work to explore hybrid methods that combine the strengths of input compression methods with advances in long-context LLMs.

5.5 Limitations

In this work, we rely on high-quality reference summaries to measure the content selection aspects of system-generated summaries. We acknowledge that human evaluation is the gold standard for text summarization. However, for large-scale multi-document summarization ($\approx$100 docs per example), it is prohibitively expensive to perform human evaluation. Karpinska et al. (2024a) reported that a human takes about 8-10 hours to read an average book (of similar length to our setting). We leave the extension of human evaluation of full-context and compression-based systems to future work. We also limit our evaluation to models with publicly available weights. We report preliminary results on SummHay using Gemini-1.5 (Table 5.10 in Appendix). Due to the high API costs of running Gemini on long inputs, we couldn't run them for other datasets. We did not conduct an extensive search for optimal prompts for the summarization task. So, it is possible that the performance of some system configurations could be improved with additional prompt tuning.

5.6 Appendix

We use GitHub copilot and Claude-3.5 Sonnet for assistance with coding and editing.

5.6.1 Datasets

For background summarization, we use the news articles from the original timeline summarization datasets, Timeline17 (Binh Tran et al., 2013), Crisis (Tran et al., 2015) and Social Timeline (Wang et al., 2015). To constrain the input length, we use a maximum of five news articles from any given day. We also experimented with prefiltering the articles using the news update of the given day, but this did not show improvements in summary quality.

{document} Question: {question} Answer the question based on the provided document. Be concise and directly address only the specific question asked. Limit your response to a maximum of {num_words} words.

Table 5.4: Prompt for our summarization task. We pass the input documents concatenated together by a \n character. The number of words in the summary are determined by the dataset (Table 5.1).

5.6.2 Experimental Setup

Transformers: We use weights from Huggingface for Llama-3.1-8B,[9] Llama-3.1-70B,[10] Command-R,[11] and Jamba-1.5-Mini.[12]

Compute: We run inference using vLLM on four 48G GPUs (Kwon et al., 2023b). Given its large size, we load Llama-3.1-70B with fp8 precision. For the smaller Llama-3.1-8B, we use a single 48G GPU. Our setup includes a mix of Nvidia's A6000, L40, and 6000 Ada GPUs.

Iterative methods: For both iterative methods, we set the maximum chunk size to 4096 tokens. For the hierarchical method, we first generate summaries for each input document. Then, we pack consecutive document summaries into the maximum chunk size for the next summarization step. We stop the process when we only have one summary. For the incremental method, we start by generating the summary of the first document. Then, we concatenate this summary with the following document for the next summarization step. We iterate through every document in the input, in the order provided by the dataset. The document order is relevant for Background (event timelines), but might not be as relevant for SummHay and WCEP.

Retrieval: We limit each document to 1024 tokens and the post-retrieval input to 32k tokens.

Summary length: To set the maximum summary words for each dataset, we first tokenize the summaries in the validation split using NLTK. We use the 80th percentile as the maximum summary words for the systems. To account for the differences in tokenizers for Llama-3.1, Command-R, and Jamba-1.5, we set the maximum number of summary *tokens* by multiplying the maximum summary words with model-specific word-to-token ratios. The word-to-token ratios for Llama-3.1, Command-R, and Jamba-1.5-Mini are 1.145, 1.167, and 1.219 respectively. For iterative methods, we use the same maximum summary token limit at each intermediate step. In Table 5.9, we report the average length of system-generated summaries.

Prompt: Table 5.4 provides our prompt for the text summarization task. We use the same prompt for all transformers and methods. We follow the recommendations from model providers and use the model-specific chat templates from Huggingface tokenizers when prompting the instruction-fine-tuned models.

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

[10]https://hf.co/meta-llama/Llama-3.1-70B-Instruct

[11]https://hf.co/CohereForAI/c4ai-command-r-08-2024

[12]https://hf.co/ai21labs/AI21-Jamba-1.5-Mini

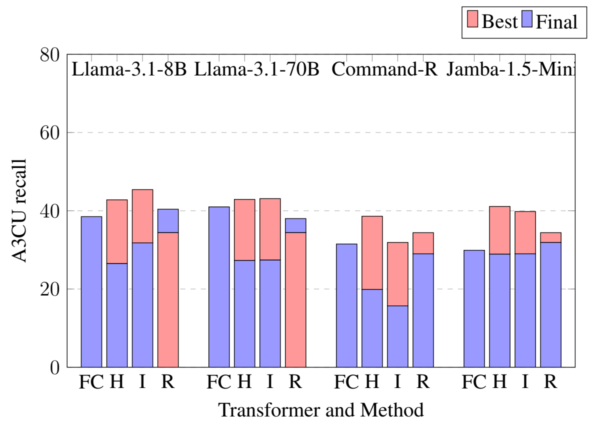


Figure 5.2: Salient information retention in the intermediate and final summaries (A3CU *recall*) for SummHay (oracle). For each compression method, we report the best recall from the intermediate outputs and the recall of the final summary. (H: hierarchical, I: incremental, R: retrieval, FC: full-context)

5.6.3 Full Metrics

We report the precision, recall, and F1 scores for A3CU and ROUGE scores (Lin, 2004) for each dataset: SummHay (Table 5.5), SummHay oracle (Table 5.6), Background (Table 5.7), and WCEP (Table 5.8). We use Huggingface evaluate for ROUGE and the original repo for A3CU.[13]

5.6.4 Example-level Trends

Figure 5.3 shows the distribution of A3CU F1 scores across examples. We notice a significant variance in system performance across all datasets.

5.6.5 Ablations

We perform ablation studies to further study our choice of models and hyperparameters. Given its small size, we used SummHay for our ablation experiments.

Gemini-1.5: We run some preliminary experiments with Gemini-1.5 Flash and Pro (Table 5.10). Across methods, we consistently found that Gemini-1.5 models generate short summaries and underperform open source models. It is possible that we could improve their summaries using a different prompt, but we leave this extension to future work. Due to the high costs associated with Gemini API, we did not run experiments with our larger Background and WCEP datasets.

Llama-3: Our iterative methods do not require a long-context transformer, so we experiment with short-context transformers to see if they are better suited for this task. We run inference with Llama-3 8B and 70B (8k context window) in the SummHay and SummHay oracle settings (Table 5.11). We found that both models are either comparable or underperform their Llama-3.1 counterparts. It is likely that the Llama-3.1 models are better at short-text summarization.

[13] https://github.com/Yale-LILY/AutoACU

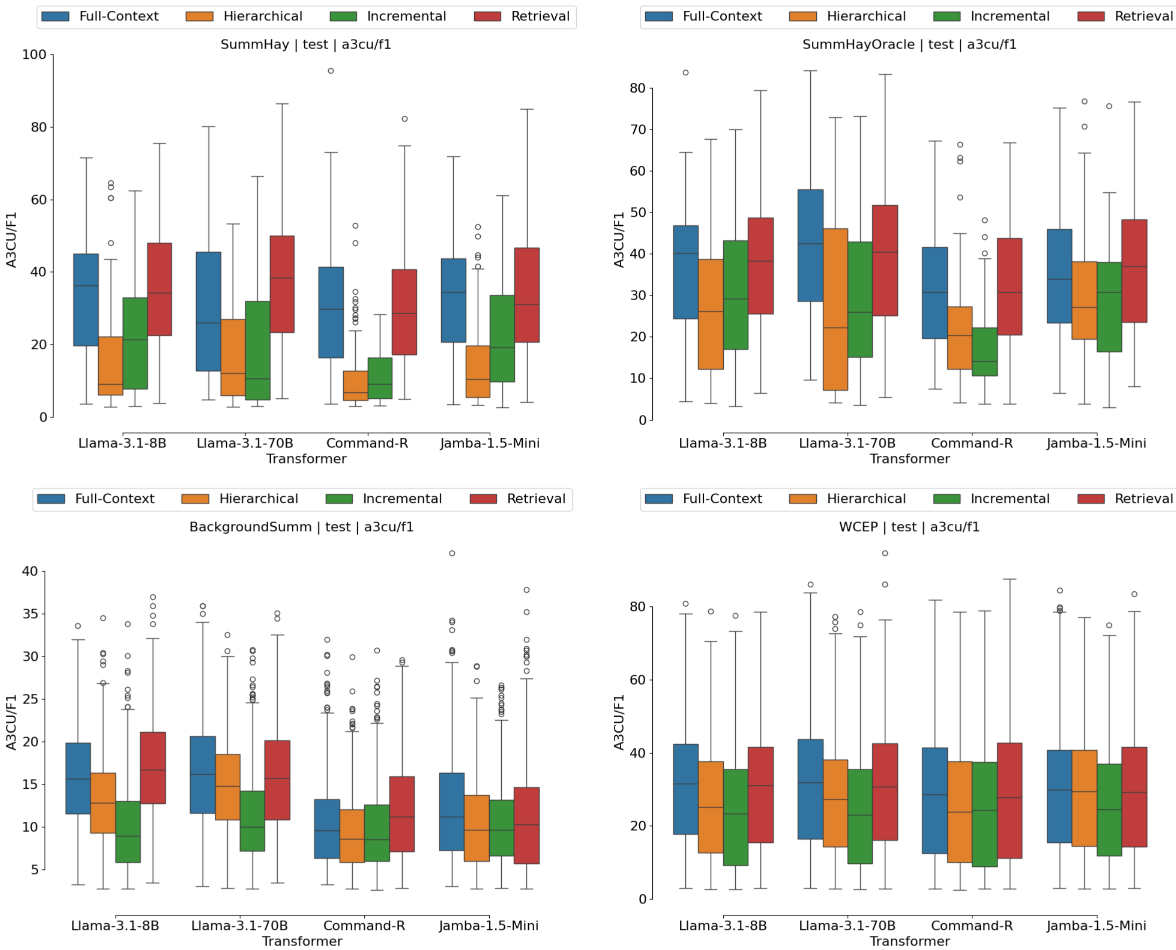


Figure 5.3: A3CU F1 score distribution across examples.

Chunk size: As we have highlighted earlier, the hierarchical method exhibits a significant degradation in summary recall. We experiment with larger chunk sizes that allow for packing more intermediate summaries into the transformer. Our results using 8k, 16k and 32k chunk sizes show minimal improvements over our default 4k chunk size.

Retriever: Following the setup of SummHay (Laban et al., 2024b), we experiment with the E5-RoPE embedding for retrieval.[14] We report results in Table 5.13. E5-RoPE performs slightly worse than the SFR-Embedding-2 results from Table 5.5.

Grounded generation: Jamba provides a grounded generation option in which the documents are passed as a separate object in the chat template. We experiment with this chat template to see if it provides any gains over our default setting of concatenating documents in the message. We report results in Table 5.14. Interestingly, this template helps improve the performance of hierarchical and incremental methods and hurts performance in full-context and retrieval settings. This needs further investigation. Command-R also includes a grounded generation template, but it is recommended for documents (or chunks) that contain 100-400 words. We couldn't make it

[14] https://huggingface.co/dwzhu/e5rope-base

Transformer	Method	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum	A3CU Recall	Precision	F1
Llama-3.1-8B	Full-Context	49.4	25.4	28.5	46.4	31.8	39.5	33.9
Llama-3.1-8B	Hierarchical	29.4	10.8	16.4	27.1	14.5	23.3	16.0
Llama-3.1-8B	Incremental	41.5	16.4	22.5	38.0	22.6	27.5	23.2
Llama-3.1-8B	Retrieval	51.8	27.0	29.3	48.9	36.3	36.7	35.3
Llama-3.1-70B	Full-Context	43.7	23.8	25.9	41.3	25.2	46.3	31.1
Llama-3.1-70B	Hierarchical	30.0	11.0	16.4	27.2	15.8	23.6	17.3
Llama-3.1-70B	Incremental	33.1	13.6	19.3	30.5	17.2	27.5	19.7
Llama-3.1-70B	Retrieval	50.2	26.7	29.3	47.1	33.1	43.8	36.3
Command-R	Full-Context	45.0	19.0	24.4	41.2	27.5	38.1	30.4
Command-R	Hierarchical	35.4	8.0	18.4	32.0	10.6	13.9	11.4
Command-R	Incremental	33.0	7.7	17.8	29.7	10.1	15.9	11.4
Command-R	Retrieval	45.0	19.6	24.9	41.8	27.3	38.3	30.4
Jamba-1.5-Mini	Full-Context	44.2	22.0	27.0	41.2	26.6	47.7	32.9
Jamba-1.5-Mini	Hierarchical	38.1	11.6	19.2	35.0	16.5	15.9	15.1
Jamba-1.5-Mini	Incremental	40.7	15.9	21.8	37.1	21.9	27.8	22.5
Jamba-1.5-Mini	Retrieval	46.4	22.8	27.6	42.8	29.4	46.4	34.7

Table 5.5: Results on SummHay.

work with full documents from our datasets.

Filtered Background: Our results showed that Background is the most challenging of the three datasets. To simplify the task, we pre-filter the documents using the update summary from the event timeline. We use the E5RoPE model (Zhu et al., 2024) to prefilter up to 128k tokens in the input for each example. However, we did not observe any significant improvements with this filtered dataset.

Transformer	Method	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum	A3CU Recall	A3CU Precision	A3CU F1
Llama-3.1-8B	Full-Context	53.4	29.0	29.7	50.1	38.5	37.9	37.1
Llama-3.1-8B	Hierarchical	40.7	18.2	21.4	38.0	26.5	31.9	27.0
Llama-3.1-8B	Incremental	48.0	21.8	25.2	44.6	31.8	32.9	30.9
Llama-3.1-8B	Retrieval	53.7	28.8	29.8	50.5	40.4	37.2	37.5
Llama-3.1-70B	Full-Context	54.1	30.1	30.7	51.0	41.0	45.8	41.8
Llama-3.1-70B	Hierarchical	37.6	18.3	21.1	34.9	27.3	32.3	27.2
Llama-3.1-70B	Incremental	41.8	20.2	23.5	38.7	27.4	37.8	29.5
Llama-3.1-70B	Retrieval	53.3	28.7	30.1	50.3	38.0	44.0	39.3
Command-R	Full-Context	48.3	20.2	25.4	44.2	31.5	38.0	32.6
Command-R	Hierarchical	41.7	12.5	21.3	38.1	19.9	26.8	21.7
Command-R	Incremental	37.1	11.0	19.8	33.3	15.7	22.6	17.2
Command-R	Retrieval	46.5	19.9	25.1	42.7	29.0	38.6	31.8
Jamba-1.5-Mini	Full-Context	47.6	24.3	28.2	44.4	29.9	47.8	35.1
Jamba-1.5-Mini	Hierarchical	46.7	20.3	25.6	43.5	28.9	33.5	29.6
Jamba-1.5-Mini	Incremental	46.2	20.5	24.4	42.9	29.0	32.5	28.9
Jamba-1.5-Mini	Retrieval	48.5	24.7	28.0	45.2	31.9	46.2	36.3

Table 5.6: Results on SummHay (oracle).

Transformer	Method	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum	A3CU Recall	A3CU Precision	A3CU F1
Llama-3.1-8B	Full-Context	36.5	8.4	18.3	33.2	18.1	15.4	15.6
Llama-3.1-8B	Hierarchical	35.2	7.2	17.5	32.0	15.5	12.8	13.1
Llama-3.1-8B	Incremental	34.4	6.6	16.4	31.1	11.8	10.5	10.0
Llama-3.1-8B	Retrieval	37.7	8.7	19.0	34.2	20.0	16.2	16.9
Llama-3.1-70B	Full-Context	36.6	8.7	18.4	33.4	18.6	15.8	16.1
Llama-3.1-70B	Hierarchical	34.5	7.5	17.4	31.4	17.6	14.2	14.7
Llama-3.1-70B	Incremental	35.2	7.2	16.5	31.9	13.0	11.6	11.1
Llama-3.1-70B	Retrieval	35.7	8.0	18.6	32.2	17.6	16.0	15.7
Command-R	Full-Context	31.9	6.1	17.5	28.6	11.3	11.4	10.3
Command-R	Hierarchical	31.5	5.8	16.7	28.7	10.8	9.5	9.3
Command-R	Incremental	34.6	6.7	16.3	31.3	11.7	9.9	9.7
Command-R	Retrieval	33.2	6.4	17.2	29.9	13.3	12.0	11.8
Jamba-1.5-Mini	Full-Context	33.6	6.8	17.7	30.1	13.1	14.2	12.3
Jamba-1.5-Mini	Hierarchical	33.5	6.0	16.1	30.4	13.4	9.2	10.2
Jamba-1.5-Mini	Incremental	35.5	6.7	16.2	32.1	13.7	9.8	10.4
Jamba-1.5-Mini	Retrieval	33.0	6.1	16.8	29.5	12.5	11.8	11.0

Table 5.7: Results on Background.

Transformer	Method	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum	Recall	A3CU Precision	F1
Llama-3.1-8B	Full-Context	37.5	14.2	26.4	29.6	39.1	29.2	30.7
Llama-3.1-8B	Hierarchical	33.9	11.3	23.8	26.1	33.8	25.3	26.2
Llama-3.1-8B	Incremental	32.7	10.5	22.8	25.6	31.7	22.9	24.0
Llama-3.1-8B	Retrieval	36.8	13.7	26.1	29.0	37.9	28.4	29.7
Llama-3.1-70B	Full-Context	37.5	14.1	26.7	30.0	38.6	30.7	31.1
Llama-3.1-70B	Hierarchical	34.3	11.4	23.8	26.6	35.6	25.7	27.1
Llama-3.1-70B	Incremental	32.5	10.4	22.6	25.5	33.0	22.7	24.2
Llama-3.1-70B	Retrieval	37.5	14.2	26.6	30.0	38.3	29.8	30.5
Command-R	Full-Context	36.6	13.7	26.1	29.9	34.1	30.2	28.9
Command-R	Hierarchical	34.1	11.1	23.9	26.4	28.6	28.4	25.6
Command-R	Incremental	34.3	11.7	24.2	27.4	29.2	27.0	25.1
Command-R	Retrieval	36.7	13.7	26.0	29.7	33.0	29.8	28.5
Jamba-1.5-Mini	Full-Context	36.8	13.8	25.8	29.8	36.3	28.6	29.3
Jamba-1.5-Mini	Hierarchical	35.8	12.8	25.1	28.8	36.6	27.9	28.7
Jamba-1.5-Mini	Incremental	34.3	11.7	23.6	27.7	33.4	24.2	25.4
Jamba-1.5-Mini	Retrieval	36.7	13.7	25.6	29.4	36.6	28.3	29.1

Table 5.8: Results on WCEP.

	Full Context	Retrieval	Hierarchical		Incremental	
			Best	Final	Best	Final
SummHay (Reference: 185)						
Llama-3.1-8B	162	195	172	106	171	141
Llama-3.1-70B	106	148	161	113	150	93
Command-R	135	134	165	151	161	115
Jamba-1.5-Mini	110	120	163	211	177	145
Background (Reference: 174)						
Llama-3.1-8B	228	232	214	222	212	206
Llama-3.1-70B	232	219	208	210	210	205
Command-R	190	215	226	227	236	232
Jamba-1.5-Mini	162	183	213	237	230	233
WCEP (Reference: 35)						
Llama-3.1-8B	44	44	43	41	43	43
Llama-3.1-70B	42	42	43	42	44	43
Command-R	42	41	42	39	42	41
Jamba-1.5-Mini	45	45	45	44	45	44

Table 5.9: Summary length statistics, using NLTK word tokenizer.

Transformer	Method	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum	A3CU Recall	A3CU Precision	A3CU F1
Gemini-1.5-Flash	Full-Context	32.3	15.1	19.7	29.8	19.2	40.6	24.6
Gemini-1.5-Flash	Hierarchical	12.5	4.5	7.2	11.2	8.0	17.2	10.2
Gemini-1.5-Flash	Incremental	37.2	15.5	21.7	34.2	19.6	34.8	23.8
Gemini-1.5-Flash	Retrieval	37.5	18.7	23.3	34.8	22.4	47.4	28.3
Gemini-1.5-Pro	Full-Context	41.8	18.3	23.9	38.8	26.2	36.8	29.2
Gemini-1.5-Pro	Hierarchical	10.9	3.1	6.5	9.7	6.9	17.0	9.2
Gemini-1.5-Pro	Incremental	22.7	6.4	13.4	20.4	10.3	21.8	12.9
Gemini-1.5-Pro	Retrieval	42.5	19.8	24.0	39.3	27.4	41.0	31.6

Table 5.10: Results on SummHay using Gemini 1.5 Flash and Pro.

Transformer	Method	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum	A3CU Recall	A3CU Precision	A3CU F1
SummHay								
Llama-3-8B	Hierarchical	22.0	8.3	13.0	20.3	10.8	23.2	13.6
Llama-3-8B	Incremental	32.6	15.0	20.0	30.0	18.3	36.2	23.2
Llama-3-70B	Hierarchical	17.6	5.0	11.0	16.0	7.4	14.3	9.2
Llama-3-70B	Incremental	34.6	13.8	19.8	31.5	16.7	30.5	20.3
SummHay (oracle)								
Llama-3-8B	Hierarchical	34.0	16.3	19.4	31.4	21.0	35.5	24.6
Llama-3-8B	Incremental	39.2	19.7	23.5	36.3	25.2	45.5	29.9
Llama-3-70B	Hierarchical	30.0	13.3	17.0	27.8	17.0	29.0	19.9
Llama-3-70B	Incremental	39.9	19.0	23.5	36.7	24.1	42.7	29.3

Table 5.11: Results on SummHay using the short context Llama-3 models.

Transformer	Method	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum	A3CU Recall	A3CU Precision	A3CU F1
Llama-3.1-8B	Hierarchical-8K	27.3	10.1	15.3	25.1	14.0	22.9	15.6
Llama-3.1-8B	Hierarchical-16K	30.8	12.6	17.6	28.4	16.7	27.9	18.9
Llama-3.1-8B	Hierarchical-32K	28.9	11.4	16.4	26.8	15.8	26.0	17.5
Jamba-1.5-Mini	Hierarchical-8K	38.2	11.8	19.5	35.2	14.5	18.4	15.2
Jamba-1.5-Mini	Hierarchical-16K	37.7	12.0	20.4	34.5	14.7	19.9	16.0
Jamba-1.5-Mini	Hierarchical-32K	37.0	12.3	19.7	33.6	14.8	21.6	16.3

Table 5.12: Results on SummHay using different chunk sizes for the hierarchical method.

Transformer	Method	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum	A3CU Recall	A3CU Precision	A3CU F1
Llama-3.1-8B	Retrieval-E5	50.1	25.1	28.6	47.3	33.9	35.1	33.2
Llama-3.1-70B	Retrieval-E5	49.8	25.7	28.7	46.8	32.2	41.1	34.6
Command-R	Retrieval-E5	44.8	19.3	24.5	41.5	27.2	36.7	29.5
Jamba-1.5-Mini	Retrieval-E5	44.1	20.8	25.5	40.7	26.9	42.0	31.5

Table 5.13: Results on SummHay using the E5 RoPE retriever instead of SFR-Embedding-2.

Transformer	Method	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum	A3CU		
						Recall	Precision	F1
Jamba-1.5-Mini-Grounded	Full-Context	45.4	22.0	26.7	42.6	26.0	43.5	31.3
Jamba-1.5-Mini-Grounded	Hierarchical	40.8	14.3	20.9	37.7	18.6	20.2	18.2
Jamba-1.5-Mini-Grounded	Incremental	41.8	17.6	22.7	38.5	23.9	32.6	25.9
Jamba-1.5-Mini-Grounded	Retrieval	44.4	21.8	26.3	41.2	26.5	43.2	31.1

Table 5.14: Results on SummHay using grounded generation chat template with Jamba-1.5-Mini.

Gold: China announces its intent to increase tariffs from 10% to 25% on $60 billion worth of imports from the United States effective June 1, 2019. **Prediction (best)**: US markets plummeted after China announced retaliatory tariffs on $60 billion of US imports, including batteries, spinach, and coffee, in response to the Trump administration's hike on Chinese imports from 10% to 25%. **Prediction (final)**: The US and China engage in an escalating trade war, imposing tariffs on each other's goods, causing market losses, economic uncertainty, and affecting global economic growth, trade, and markets, with potential impacts on US consumers and businesses.
Gold: Venezuela's Ambassador to Italy, Isaías Rodríguez, resigns, citing the government's lack of finances amid international sanctions on the Nicolás Maduro government. Rodríguez says, due to the sanctions, he has been unable to pay his staff or the rent for his office in Rome. **Prediction (best)**: The Venezuelan ambassador to Italy, Isaías Rodríguez, has resigned due to the country's financial difficulties caused by US sanctions. He cites inability to pay employees and rent, and will dedicate himself to being a grandfather, stating he "cannot turn water into wine" **Prediction (final)**: The document discusses Venezuela's humanitarian crisis, disputed re-election, and diplomatic efforts to resolve the crisis, including US sanctions, opposition negotiations, and international support for a democratic transition and potential regime change.
Gold: One colleague mentions that their department uses specific productivity metrics like the number of tasks completed per week to measure the impact of remote work on employee productivity. A colleague argues that remote work results in fewer distractions compared to an open office environment, citing a 2019 study that found remote workers are 35% less likely to be distracted by co-workers. Another colleague points out that remote employees often work more overtime, referencing a 2022 company internal report showing a 20% increase in overtime hours since the shift to remote work. A colleague highlights that remote work allows for extended deep work periods, mentioning that they've personally experienced 3-hour uninterrupted work blocks at home, which were rarely possible in the office. One colleague mentions that remote work can align better with individual employees' peak productivity times, citing a 2020 survey where 60% of employees reported higher productivity outside the traditional 9-to-5 work hours. **Prediction (best)**: Yes, productivity level change is discussed in the conversation. \n\n Colleague 3 mentions a 2019 study that found remote workers are 35% less likely to be distracted by co-workers compared to those in an open office, suggesting that remote work can be a productivity booster. Colleague 1 also agrees that the solitude of remote work can lead to deep focus and increased productivity. \n\n Additionally, colleague 2 compares the office environment to a marching band parading through the office, emphasizing how the constant interruptions can break one's flow and hinder productivity. This highlights the potential benefits of remote work in terms of increased productivity due to fewer distractions. \n\n Overall, the conversation suggests that remote work can have a positive impact on productivity, with fewer distractions and interruptions allowing employees to focus more effectively. **Prediction (final)**: Yes, productivity level change is discussed.
Gold: Suspected Baloch nationalist gunmen storm a coal mine in Islamabad, Pakistan, killing at least five people. **Prediction (best)**: At least five people were killed in a gun and bomb attack on a coal mine in southwestern Pakistan by ethnic Baloch separatists, with two miners, two security personnel, and a driver among the dead, amid escalating violence in the region. **Prediction (final)**: The document reports various news articles from 2019, including attacks on hotels in Gwadar, Pakistan, by the Balochistan Liberation Army, as well as human trafficking and security concerns during the Muslim fasting month of Ramadan.

Table 5.15: Examples illustrating loss of salient information in Hierarchical method.

Gold: One student shares that they do 10 minutes of deep breathing exercises each night before going to bed to help reduce stress. A student mentions that taking a 30-minute walk each afternoon helps clear their mind and manage stress levels during exam preparation. One of the students suggests taking a 5-minute break after every 25 minutes of studying, and mentions the Pomodoro technique as a helpful method. A student recommends using a specific meditation app called 'Calm' that they use for 15 minutes each morning to manage stress. Another student shares their habit of writing down 3 positive affirmations about their capabilities each day, which they find significantly helps reduce their stress. **Prediction (best)**: Yes, the students indirectly shared exam stress tips through their conversation. They discussed various techniques and strategies that helped them manage stress and stay focused during exam preparation, including:\n\n* Maintaining a consistent sleep schedule and taking breaks\n* Using the Pomodoro Technique to avoid burnout\n* Incorporating digital detox into their daily routine\n* Practicing deep breathing exercises\n* Taking short walks to clear their minds\n* Trying relaxation techniques such as meditation\n* Using spaced repetition with the Anki app\n* Creating mind maps for visual learners\n* Using a meditation app like 'Calm'\n\n They also shared specific experiences and recommendations, such as Alejandro's 30-minute afternoon walk, which helped him manage stress and clear his mind. They also discussed the potential benefits of physical activity, meditation, and relaxation techniques in reducing anxiety and improving focus. The conversation was supportive and encouraging, with the students offering to help and motivate each other throughout the exam preparation period. **Prediction (final)**: No, they did not directly share exam stress tips. Their conversations centered on sharing personal experiences and strategies related to studying, time management, prioritization, and productivity, which indirectly contributed to managing exam stress. They discussed various approaches to studying, adapting their study plan, and organizing information, but did not explicitly mention stress management techniques like deep breathing exercises or the Pomodoro technique. Their focus was on developing a study plan, sharing notes, and reviewing material together to prepare for the exam, rather than directly addressing stress management.
Gold: Prince Harry and Meghan, the Duke and Duchess of Sussex, name their son Archie Harrison Mountbatten-Windsor. **Prediction (best)**: Prince Harry and Meghan, Duchess of Sussex, have named their baby boy Archie Harrison Mountbatten-Windsor. The palace announced the name after the couple posed with the two-day-old for the media, meeting Queen Elizabeth II for the first time. **Prediction (final)**: Prince Harry and Meghan Markle introduced their son Archie to the world at Windsor Castle, presenting a more casual and intimate image, contrasting with traditional royal protocol. They shared parenting duties and responsibilities, showcasing a new generation of feminist royals.
Gold: The United States lifts tariffs on steel and aluminum from Canada and Mexico. **Prediction (best)**: The US will lift tariffs on steel and aluminum from Canada and Mexico, resolving a yearlong deadlock and providing economic relief to Detroit automakers, who paid billions in increased duties, and addressing a key obstacle to ratifying the new North American trade pact. **Prediction (final)**: The number of asylum seekers in Canada tripled between 2015 and 2018, with a peak of 8,801 claims in August 2017. Most claims came from Haitians, Nigerians, and US citizens, with many fleeing the US due

Table 5.16: Examples illustrating loss of salient information in Incremental method.

Chapter 6

Estimating Optimal Context Length for Hybrid Retrieval-augmented Multi-document Summarization

This work was published in COLM 2025 (Pratapa and Mitamura, 2025a).[1]

6.1 Introduction

Language models increasingly support longer context windows, leading to useful applications in large-scale multi-document summarization. Recent work has shown that these models are not very effective at their claimed context windows (Hsieh et al., 2024; Yen et al., 2025). An alternative to the full context setting is retrieval-augmented generation (RAG), and previous work has illustrated its effectiveness for long input processing (Asai et al., 2024; Li et al., 2024b). RAG systems facilitate better use of the LM context windows by passing only the most relevant information to the model. However, the choice of retrieval length that provides peak RAG performance is often unclear and sensitive to the choice of retriever, language model, and downstream task (Jin et al., 2025). In this work, we present a methodology for estimating this optimal retrieval length as a function of the retriever, summarizer, and dataset. In addition to providing gains over the full context setting, our method also outperforms the context-length estimates identified by standard long-context evaluation benchmarks. Figure 6.1 provides a schematic overview of our method.

Previous efforts to combine RAG and long-context LMs focused on query-based routing (Li et al., 2024b), or iterative RAG (Yue et al., 2025). While these methods are effective, they rely on the model's ability to accurately determine the scope of information need and self-evaluate its own output. This might not always be a feasible option, especially for smaller LMs. In this work, we take a complementary approach to combine RAG and long-context and show its effectiveness for models ranging from 0.5B to 72B parameters. We evaluate on a challenging large-scale multi-document summarization dataset (Laban et al., 2024a).

[1]Code and data are available at `https://github.com/adithya7/hybrid-rag`.

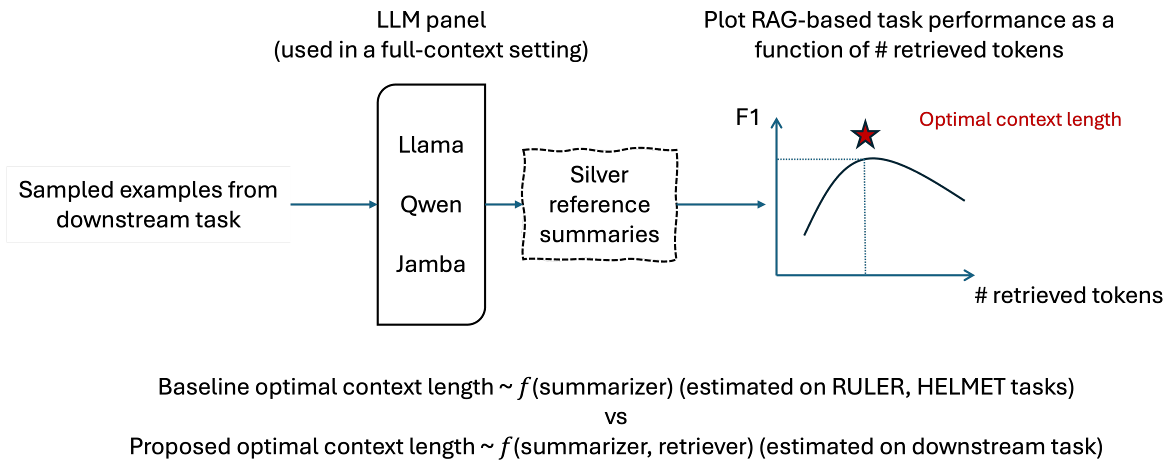


Figure 6.1: A schematic overview of our proposed method. Unlike traditional benchmarks, we estimate the optimal context length for RAG as a function of retriever and summarizer on a given downstream task. We first sample a fraction of examples from downstream task. On this sample, we run a panel of LMs in a full-context setup to create silver candidates. We then identify the top silver candidates using Minimum Bayes Risk decoding. With the help of these silver candidates, we estimate the optimal retrieval length for the given experiment config.

In a recent work, Jin et al. (2025) compared the RAG performance of varying model sizes on the question-answering task and found that the optimal retrieval length varies considerably across model sizes and families. They also found that this length is sensitive to the choice of retriever. Similarly, Yu et al. (2024b) noted the sensitivity of optimal retrieval length to the downstream task. Based on these observations from previous work, we hypothesize that the retrieval length that provides peak performance should be modeled as a function of the three main components of the RAG pipeline: retriever, summarizer, and dataset. For our baselines, we use two popular long-context evaluation benchmarks, RULER (Hsieh et al., 2024) and HELMET (Yen et al., 2025). They benchmark models on a suite of tasks with inputs of increasing lengths. RULER focuses on synthetic retrieval and aggregation tasks such as multi-hop tracing and counting on synthetically generated data. HELMET expands on RULER and further includes tasks such as LongQA and summarization over real-world data. Although these provide *effective* context length estimates for individual LMs, these estimates are often agnostic to the downstream dataset and the retrievers when used in the RAG setting.

Given a dataset, we first create a subset of representative examples by random sampling. We then use a panel of LMs to compile a candidate set of silver reference summaries. In our panel, we include LMs from the Qwen (Qwen et al., 2025), Llama (Grattafiori et al., 2024) and Jamba (Team et al., 2024b) series. From the pool of candidate silver references, we use Minimum Bayes Risk decoding (Kumar and Byrne, 2004) to identify the top silver reference summaries. For a given combination of retriever and summarizer models, we perform a search over context lengths on this silver subset to estimate the optimal retrieval length. Unlike baseline methods, our approach is customized to the specific experiment configuration (dataset, retriever, and summarizer). Our method is based on two key observations. First, larger LMs show robust performance across a broad range of context lengths. This is mainly due to their enhanced ability to deal with noise in the retrieved input (Jin et al., 2025). Second, to identify a task-specific estimate, we can approximate the gold summaries with silver candidates sampled from strong long-context LMs.

We evaluated our method for the multi-document summarization task using the SummHay dataset (Laban et al., 2024a). Our results show that all retrieval-based methods (baselines and ours) significantly outperform full-context. Our method performs the best in most settings, followed by HELMET- and RULER-based estimates. Although HELMET-based estimates sometimes perform comparable to our method, neither the LongQA nor summarization task-based HELMET estimates consistently perform better. Notably, our method performs much better on very long-context LMs such as Qwen 2.5 1M and ProLong 512k. Our analysis also shows that our method generalizes well to model classes outside of our panel (e.g., Phi-3). We also perform ablation experiments on our LM panel as well as the size of our sampled subset.

6.2 Estimating Optimal Context Length for Retrieval

For the multi-document summarization task, given a long input and a query, we have two possible systems. First, the entire input is fed directly into a long-context summarizer that supports such lengths (full-context). Second, we use the query to rank the documents and only pass the top-k relevant documents to the summarizer (RAG). Previous work has shown that long-context models are not effective at their claimed context windows, and RAG can help improve task performance (Yu et al., 2024b; Pratapa and Mitamura, 2025b).

Benchmarks such as RULER and HELMET provide a comprehensive evaluation of long-context models across a suite of NLP tasks, including QA and summarization. However, these benchmarks focus solely on the model and do not study the effects of unseen downstream datasets and the retrievers used in RAG settings. Previous work has briefly studied this with Jin et al. (2025) noting significant variance in long-context RAG performance depending on the choice of LM and retriever. Yu et al. (2024b) noted similar behavior for question-answering tasks. Giorgi et al. (2023) studied the effects of retriever and summarizer for short-context open-domain multi-document summarization. Therefore, we hypothesize that the optimal context length estimate for a RAG system should be a function of the retriever, summarizer, and specific downstream task.

Our proposed method is centered on two key observations. First, large LMs show robust performance across a broad range of context lengths because of their enhanced ability to deal with noise in the retrieved input. Jin et al. (2025) observed this behavior for long input QA tasks. Second, gold references can be approximated by silver references sampled from strong long-context LMs. For a given dataset (D), retriever (R), and summarizer (S), our method involves the following steps. See Figure 6.1 for an overview of our method.

1. We sample a subset of the dataset (D). Each example in this subset constitutes a set of documents and a query. (§6.2.2)
2. We used a panel of LMs (§6.2.1) to generate summaries for this subset. These summaries serve as our candidate silver references. (§6.2.2)

3. We use Minimum Bayes Risk decoding to identify the best silver references. (§6.2.2)
4. We perform a search for retrieval lengths (8k to 80k) by comparing the system-generated summary (using R & S) against the silver references. This search gives us the optimal retrieval length estimate for our RAG setup. (§6.2.3)
5. Finally, on the full dataset, we retrieve the top-k documents that fit into this length estimate (using R) before generating a summary (using S).

6.2.1 LM panel

In our LM panel, we include a diverse class of models. Panels of diverse LMs have previously been explored for evaluation and are considered a strong alternative to a single LM evaluator (Verga et al., 2024).

Large LMs: We choose Qwen-2.5 72B (Qwen et al., 2025), Llama-3.3 70B (Grattafiori et al., 2024), and Jamba-1.5 Mini (Team et al., 2024b). These are the largest models from each class that we could run locally.[2]

Long-context LMs: We include two smaller LMs that are specifically trained for long-context tasks, Qwen-2.5-1M 14B (Yang et al., 2025) and ProLong 512K (Gao et al., 2024). ProLong is continually trained on long texts starting from the Llama-3 8B model.

In our pool, we focussed on including diverse models while being within our compute budget to run these models locally. Our panel can be easily modified with newer variants of these models as well as include API-based models.

6.2.2 Generating silver references

To begin, we select a fraction of the examples from the dataset (default: 25%) using a uniform sampling algorithm (without replacement). We run our silver reference generation on this data subset. We leave out the gold references and do not use them in our context-length estimation procedure.

Silver references: We run our LM panel to create a pool of candidate silver references. We used temperature sampling ($\tau = 0.5$) to generate three candidate summaries for each LM. We use LMs in a full-context setup and do not assume any optimal context length.

Pooling: We experiment with two ways to collect our final set of silver references. First, we used a single LM from the panel and select the three sampled candidates as our silver references. Second, we collect many candidates by pooling outputs from all LMs in our panel. In this scenario, we use Minimum Bayes Risk (MBR) decoding to identify the three highest scoring candidates. We follow previous work (Suzgun et al., 2023; Bertsch et al., 2023) to compute the similarity between each pair of candidates and obtain the alignment scores among the candidates. To be consistent with our downstream evaluation metric, we use the A3CU F1 score as our utility metric in MBR decoding.

Our use of MBR decoding here borrows ideas from previous summarization works, specifically post-ensemble (Kobayashi, 2018) and crowd sampling (Suzgun et al., 2023). Similarly to Kobayashi (2018), we use a model ensemble in the post-processing stage. We follow Suzgun

[2]We couldn't run Llama 405B and Jamba 1.5 Large (400B) locally on our setup.

et al. (2023) to use temperature sampling and a neural utility metric. However, our utility metric differs from the BLEURT and BERTScore used in Suzgun et al. (2023).

6.2.3 Search for optimal retrieval length

To identify the optimal length for the retrieval step, we search a wide spectrum of context lengths from 8K to 80K tokens in 8K intervals. For each context length C, we run the RAG pipeline on the silver dataset by retrieving up to C tokens (see §6.3.2) and passing them to the summarizer. We evaluate the system generated summaries against the silver references. We generate three predictions per example using temperature sampling ($\tau = 0.5$) and take the average A3CU F1 score (see §6.3.1). For efficiency reasons, we choose the smallest context length that falls within a standard deviation of the maximum score as our optimal context length. Previous works RULER and HELMET use coarser intervals for context lengths (multiples of 8K).

Yue et al. (2025) is closely related to our work. For the long input question answering task, they propose an iterative RAG method that uses inference-time scaling. Unlike traditional RAG, their method iteratively generates subqueries and retrieves additional documents as needed before generating the final answer. They present a computation allocation model that optimizes task performance based on three parameters: number of documents, number of demonstrations, and maximum number of iterations. Our setting differs considerably from that work. For multi-document summarization task, we have a fixed set of documents, and including demonstrations in the prompt is often infeasible. We believe that our single-step retrieval solution can be combined with such iterative methods to further improve task performance. We leave this extension to future work.

6.3 Experimental Setup

In this section, we describe our dataset, the evaluation metric, baselines, and the systems used for the retrieval and summarization tasks.

6.3.1 Dataset & Metric

SummHay: Proposed by Laban et al. (2024a), this is a multi-document summarization curated using GPT-3.5 and GPT-4o, starting with summary insights followed by document generation. Each input typically consists of 100 documents (avg. length 884 words), and the summary consists of an average of 185 words. This dataset includes 92 examples that cover the news and conversational domains.

Metric: For the summarization task, we report the F1 score of the reference-based Atomic Content Unit (A3CU) metric (Liu et al., 2023c). This model-based metric is trained to predict a score that measures the overlap of atomic content units (Liu et al., 2023b) between the system-generated and reference summaries. Previous work has found that this metric is strongly correlated with human evaluation for both single (Liu et al., 2023c) and multi-document summarization (Pratapa and Mitamura, 2025b).

6.3.2 Retrieval systems

For our retrieval task, we use entire documents as retrieval units and obtain document embeddings using Qwen-2-based GTE models (Li et al., 2023). We then compute cosine similarity between document and query embeddings and pick the top-k documents that fit within the specified context length.

Jin et al. (2025) analyzed the effect of the retriever on optimal context lengths in RAG settings and found that the stronger retriever has shorter optimal lengths than the weaker retrievers. To see the impact of this in our setting, we experiment with two sizes of GTE embeddings, Qwen-2-1.5B[3] and Qwen-2-7B.[4]

We acknowledge the impact of chunking strategies on RAG performance (Chen et al., 2024), however, shorter chunks might need additional recontextualization.[5] We leave the exploration of fine-grained chunking strategies to future work.

6.3.3 Summarization systems

For the summarization task, we use the instruction fine-tuned variants from Qwen-2.5, Llama-3, ProLong, and Phi-3 series of models.

Qwen-2.5: We experiment with multiple sizes from this series including 0.5B, 1.5B, 3B, 7B, 14B, 32B, and 72B (Qwen et al., 2025). The smaller models ($\leq$3B) only support a context length of 32K, while the larger models support up to 128K tokens. For the smaller models, we report RAG @ 32K as their full-context performance.

Qwen-2.5-1M: These are long-context variants of the Qwen-2.5 7B and 14B models (Yang et al., 2025) supporting up to a context length of 1M tokens.

Llama-3: We include 1B, 3B, 8B and 70B models in our experiments. All models support a context length of 128K tokens (Grattafiori et al., 2024).

ProLong: Gao et al. (2024) continually fine-tuned Llama-3-8B-Instruct on long texts of up to 512K tokens. They are first trained on 20B training tokens of 64K data, followed by another 20B training tokens of 512K data. We experiment with the 64K and 512K variants.

Phi-3: We use three model sizes, Mini (3.8B), Small (7B) and Medium (14B). All of these models support context lengths of up to 128K tokens (Abdin et al., 2024).

We use vLLM (Kwon et al., 2023a) for our inference runs, using up to four 48GB L40S GPUs in our experiments. For each set of input documents, we sample three summaries using temperature sampling ($\tau = 0.5$). To provide a fair comparison of our systems, we limit all of our inputs to a maximum of 128K tokens. See Appendix §6.9.2 for additional details about our task prompts, tokenization, truncation strategies, and summary lengths.

6.3.4 Baselines

Full-context: In this setup, we utilize the full context window supported by the summarization model. Typically, larger models also tend to perform well in long-context tasks. To study this be-

[3]`https://huggingface.co/Alibaba-NLP/gte-Qwen2-1.5B-instruct`

[4]`https://huggingface.co/Alibaba-NLP/gte-Qwen2-7B-instruct`

[5]`https://www.anthropic.com/news/contextual-retrieval`

havior, we include models of varying sizes in our experiments. Inputs longer than the supported context window are truncated starting with the longest documents.

For our RAG baselines, we rely on widely used long-context benchmarks RULER and HELMET that estimate efficient context windows for language models. For these baselines, we limit the number of tokens retrieved to an efficient context window of the corresponding summarization model.

RULER (Hsieh et al., 2024) benchmark consists of a collection of synthetic retrieval tasks at varying input lengths (8K, 16K, 32K, 64K and 128K). For a given LM, this benchmark evaluates its retrieval performance at these input lengths and determines an effective context window by using the performance of Llama-2-7B @ 4K as the threshold. We used the effective context windows reported in previous work as our baseline estimates.

HELMET (Yen et al., 2025) benchmark covers a suite of NLP tasks, with multiple datasets included in each task. The tasks are recall, RAG, citation, re-ranking, ICL, LongQA, and summarization. For each dataset, they evaluate system performance at varying input lengths (same set as RULER). They report task averages, as well as a HELMET average. As our baseline, we select the two most relevant subtasks, LongQA and summarization. For each task and LM, we choose the context length with the highest task average as the effective context window for LM.

Note that both RULER and HELMET benchmarks evaluate model in a full-context setting but often find the optimal context window to be shorter than the context window claimed for (or supported) by the LM. In our experiments, we used previously reported scores on the RULER and HELMET benchmarks. See Table 6.7 in Appendix §6.9.1 for a full list of context length estimates from our baselines.

6.4 Results

In Table 6.1, we compare our method with the baselines on the SummHay dataset. All RAG-based systems (baselines and ours) outperform full-context setup. Our method consistently shows strong performance across model classes, sizes, and retrievers. Although the RULER- or HELMET-based estimates do well in specific instances, neither is consistently better across all settings. Among our baselines, we find that the LongQA-based estimate from HELMET performs the best. In Table 6.8 in the Appendix, we report the context window estimates used in each experiment setting as well as the standard deviation across three random seeds.

6.5 Discussion & Analysis

We now analyze the effectiveness of our method in various settings. In §6.5.1, we look at very long context LMs (>500K). In §6.5.2, we evaluate the generalization of our estimation method to a model class not included in our LM panel. In §6.5.3, we contrast our pooled estimate with those obtained using silver references from a single large LM. We also evaluate the effect of the dataset sampling ratio on the quality of the estimated context length (§6.5.4). Finally, in §6.5.5, we discuss the performance and efficiency gains with our RAG setup.

6.5.1 Very long-context LMs

As LMs improve their long-context reasoning, there is often a reduced need for RAG. Recent work (Yu et al., 2024b) argues for the combination of long-context models and RAG, and our results in Table 6.1 reinforce this argument. However, we want to test the effectiveness of our method on LMs carefully trained for long-context reasoning. For this analysis, we chose Qwen 2.5 1M models (Yang et al., 2025) (7B, 14B), and ProLong 512K (Gao et al., 2024). These models are continually trained on long texts and show almost perfect performance at 128K context length on HELMET. We report results in Table 6.2. Our method consistently outperforms the baselines. We leave the exploration of closed-weight API-based models such as Gemini 1.5 Pro to future work.

6.5.2 Generalization to new models

In our LM panel, we included a mixture of Qwen, Llama, and Jamba models (§6.2.1). To test the generalization of our method to a new model class, we report the performance for the Phi-3 series (Abdin et al., 2024). In Table 6.3, we compare our proposed method with the baseline using GTE 1.5B and 7B retrievers. We find that RULER estimates perform the best and our method is a close second. In contrast to Table 6.1, the HELMET summarization estimate is better than its LongQA-based estimate, but both underperform our method.

6.5.3 Effectiveness of system pooling

To test the effectiveness of pooling systems using MBR decoding (§6.2.2), we compared the pooled estimate of the system against two variants based on silver references from a single LM. We experiment with Qwen-2.5 72B and Llama-3.3 70B. In Table 6.4, we compare the effectiveness of silver summaries. Notably, we find that the Qwen 72B-based estimate fares better than both the Llama 70B-based and pooled estimates. This could be because Qwen-2.5 provides slightly full-context performance compared to Llama-3.3 70B (see Table 6.1).

Based on these results, we perform further analysis of our silver references in the pooling setup. In Table 6.5, we report the counts for how often each silver LM is chosen in the top-3 post-MBR decoding. The notable outliers here are Qwen-2.5 72B (picked least often) and Llama-3.3 70B (picked most often). This shows a potential limitation of our pooling-based estimate. Although MBR decoding allows us to make better use of the target summary space, it is possible that low-quality summaries in the pool could adversely impact the overall performance, albeit only by a small margin. An interesting future work direction would be to explore Best-of-N sampling as an alternative to MBR decoding.

6.5.4 Effect of sample size

As we describe in §6.2.2, we sample a subset of the dataset before generating silver references using our LM panel. To understand the effect of this sample size, we compare various sampling ratios in Table 6.6. Our results show that even a very small sample (10% $\approx$ 9 examples) is sufficient for our estimation and shows superior performance to baselines.

6.5.5 Performance & Efficiency

Small LMs tend to have very limited effective context windows; therefore, optimal RAG is necessary for improved task performance. For large LMs with longer effective context windows, optimal RAG can provide efficiency gains while maintaining or improving task performance.

Performance: Our results from Table 6.1, Table 6.2, and Table 6.3 show the effectiveness of our method in models ranging from 0.5B to 72B parameters. For a given downstream task, the user can pick a model size that is most suited to their computing budget. For example, Qwen-2.5 $\leq$ 7B can run on a single 48GB GPU, while larger models would require up to 4$\times$48GB GPUs.

Efficiency: Compared to the baselines, our method often provides a significantly shorter context length estimate (see Table 6.8 in the Appendix §6.9.1). Therefore, the final summarization run on the full dataset is much more efficient with our method. However, we acknowledge that our method requires task-specific additional inference time compute to determine the optimal context length. Similar compute is also needed for benchmarks such as RULER and HELMET that compute task averages. In Table 6.6, we showed that our estimation requires a very small sample of the dataset, so the marginal cost of our method would be lower as the size of the dataset increases.

6.5.6 Limitations of Optimal Estimate

Our optimal retrieval length estimation method implicitly considers multiple factors to balance precision and recall for a given downstream task. Unlike the baseline methods, our method factors in the retreiver quality and the information aggregation needs of the downstream task. However, we notice certain limitations of our methods. First, in certain scenarios, our method could not provide a tight context-length estimate (see Llama-3.1 8B and 70B in subsection 5.6.3) leading to suboptimal performance. Second, our method estimates the retrieval length for an entire dataset and does not account for variance in the questions. Questions can have varying amounts of information needs, where for some questions the system needs to retrieve specific information from a small set of documents. For other questions, the system needs to aggregate the information from all parts of the input. Ideally, depending on the question and the source documents, the system should be able to adjust the retrieval length to improve task performance and efficiency. The size of the sampled subset should be adjusted on the basis of the diversity of questions in the downstream task.

6.5.7 Retrieval vs Iterative Compression

In this work, we showed the impact of using retrieval-based methods to improve both performance and efficiency on long-context tasks. Some alternatives to retrieval-based methods include iterative compression methods such as hierarchical and incremental summarization (chapter 5). Compared to the full-text setting, all compression methods suffer from information loss. If the retrieved documents do not contain enough information needed to generate the correct response, the error cascades to the downstream task performance. As we see in this chapter, identifying the optimal retrieval length is a trade-off between recall and precision. For the iterative methods of chapter 5, a corresponding factor would be to control the compression ratio.

The relative effectiveness of retrieval-based methods and iterative text compression would depend on the nature of the input and the user question. For questions seeking specific information from the input, retrieval-based methods could be beneficial. However, for more abstract questions, iterative text compression methods could be more effective. Additionally, the retrieval step is computationally less expensive compared to a large language model inference call; therefore, retrieval-based methods could be a cost-effective option.

6.6 Conclusion & Future Work

In this work, we presented a methodology for estimating optimal context length for RAG-based summarization systems. Unlike traditional long-context benchmarks, our method is geared to a specific downstream dataset and models the estimate as a function of the entire experimental configuration. We show the superior performance of our method across model classes and sizes. We show a generalization of our method to new model classes, as well as its effectiveness on models with very long context windows (>500K). In future work, we plan to apply our method to other tasks, such as open-domain multi-document QA and long-document summarization (Zhou et al., 2023). Previous work has also shown that the relative performance of long context and retrieval varies between examples (Karpinska et al., 2024b; Pratapa and Mitamura, 2025b), so another future direction is to identify the optimal retrieval context length for each example. Using open-weight models allowed us to analyze our method across various model sizes within a reasonable compute budget. We expect future work to expand our LM panel to include larger API-based models such as Gemini or GPT.

Another line of work studies input compression methods (Jiang et al., 2024; Xu et al., 2024) that fit long inputs to a fixed context length. Although these are a promising alternative to full-context setup, they may suffer irreversible information loss (Pratapa and Mitamura, 2025b). In this paper, we focus on the strengths of RAG while taking advantage of the long-context reasoning capabilities of recent LMs. We leave the exploration of input compression with long-context methods to future work.

6.7 Limitations

We discuss the limitations of our estimation method and the potential ways for future work to improve them. We use a silver panel consisting of medium-sized open-weight models (§6.2.1). This silver panel might not work as effectively to estimate the optimal context length for a much larger model such as Gemini-2.5 Pro. In such situations, we believe the silver panel should consist of models with similar capacity. Additionally, we use a full context setup to get silver summaries (§6.2.2) and this might not work as well if the inputs are much longer than the context windows supported by the LMs in our silver panel. An option is to perform RAG by retrieving tokens up to the LM's supported context window (similar to our approach, smaller Qwen models in §6.3.3). Our analysis in §6.5.3 also highlighted the limitations of system pooling, and future work could explore the use of Best-of-N sampling to improve the pooling mechanism. Finally, our method relies on the availability of at least a few examples from the downstream dataset and might not work well if this sample is not representative of the downstream task.

6.8 Ethics Statement

In this work, we limit our focus to the content selection evaluation of our summarization systems. However, we acknowledge that the factual accuracy of summaries is of great practical importance and point the reader to related work on hallucination in text summarization. We believe that our RAG-based estimation procedure does not increase (or decrease) the chances of possible hallucination in RAG-based text summarization systems.

6.9 Appendix

6.9.1 Context length estimates

In Table 6.7, we list our models and the context length estimates from our baselines. In Table 6.8, we present our full set of results, including the standard deviation across the summaries of the three sampled systems summaries and the context length for each setting.

6.9.2 Experiment details

Dataset

We truncate the input documents to 128K tokens. We start by truncating the longest documents first. Due to slight differences in the tokenization methods between model classes, we calibrate the maximum number of summary tokens across models. We first get the 80th percentile of summary length (in NLTK tokens) and use the model-specific word-to-token ratio to set the max summary tokens.

We use the following prompt for the summarization task,

```
{document}

Question: {question}

Answer the question based on the provided document.
Be concise and directly address only the specific question asked.
Limit your response to a maximum of {num_words} words.
```

Generation

For summary generation, we used temperature sampling (0.5) and generated three summaries for each input. All the results we report are the average scores across three runs. For the retrieval task, we limit the length of each document to 1024 tokens.

Compute

We use a single L40S GPU for all our retrieval runs. For our summarization task, we use up to four L40S GPUs.

Retriever	Summarizer	Full-context	RULER	HELMET		Ours
				Summ	LongQA	
GTE 1.5B	Qwen-2.5 0.5B	16.7	-	-	-	**20.6**
	Qwen-2.5 1.5B	26.3	-	26.3	**28.7**	27.4
	Qwen-2.5 3B	29.5	-	29.5	29.5	**30.0**
	Qwen-2.5 7B	34.1	36.4	34.5	**37.6**	37.2
	Qwen-2.5 14B	35.7	35.6	-	-	**37.4**
	Qwen-2.5 32B	33.9	35.1	-	-	**36.6**
	Qwen-2.5 72B	32.5	32.5	35.0	35.0	**36.3**
	Llama-3.2 1B	17.7	-	24.6	24.6	**25.8**
	Llama-3.2 3B	28.7	-	28.7	**31.1**	30.3
	Llama-3.1 8B	33.3	**34.9**	**34.9**	34.0	34.5
	Llama-3.3 70B	31.9	33.2	35.8	33.2	**35.9**
	ProLong 64K	24.9	-	-	-	**32.2**
GTE 7B	Qwen-2.5 0.5B	17.3	-	-	-	**21.3**
	Qwen-2.5 1.5B	26.8	-	26.8	27.7	**28.2**
	Qwen-2.5 3B	30.2	-	30.2	30.2	**32.7**
	Qwen-2.5 7B	34.1	36.8	34.9	**36.9**	**36.9**
	Qwen-2.5 14B	35.7	35.4	-	-	**36.2**
	Qwen-2.5 32B	33.9	34.6	-	-	**37.2**
	Qwen-2.5 72B	32.5	32.5	**35.9**	**35.9**	35.3
	Llama-3.2 1B	17.7	-	24.9	24.9	**25.4**
	Llama-3.2 3B	28.7	-	28.7	29.7	**31.4**
	Llama-3.1 8B	33.3	**35.1**	**35.1**	33.7	33.7
	Llama-3.3 70B	31.9	34.4	**35.8**	34.4	33.3
	ProLong 64K	25.9	-	-	-	**32.3**

Table 6.1: Comparison of our method against the baselines on the SummHay dataset. We report average A3CU F1 scores across three sampled summaries. For the baselines, we only report scores for models with context length estimates previously reported in prior work. See Table 6.8 in Appendix for context window estimate used in each experiment.

Retriever	Summarizer	Full-context	RULER	HELMET		Ours
				Summ	LongQA	
GTE 1.5B	Qwen-2.5-1M 7B	32.1	33.3	32.1	32.1	**33.6**
	Qwen-2.5-1M 14B	35.6	35.6	35.6	35.6	**37.4**
	ProLong 512K	31.0	-	31.0	31.0	**32.3**
GTE 7B	Qwen-2.5-1M 7B	32.1	**32.9**	32.1	32.1	**32.9**
	Qwen-2.5-1M 14B	35.6	35.6	35.6	35.6	**36.6**
	ProLong 512K	31.0	-	31.0	31.0	**32.5**

Table 6.2: A comparison of our method against the baselines for very long-context LMs. Except for RULER on Qwen-2.5-1M 7B, all baselines estimate a full 128K context length.

Retriever	Summarizer	Full-context	RULER	HELMET		Ours
				Summ	LongQA	
GTE 1.5B	Phi-3 Mini	11.0	**30.6**	30.4	30.4	**30.6**
	Phi-3 Small	27.8	-	31.1	30.3	**31.9**
	Phi-3 Medium	29.4	**30.7**	29.9	29.4	**30.7**
GTE 7B	Phi-3 Mini	11.0	**29.9**	28.3	28.3	**29.9**
	Phi-3 Small	27.8	-	**32.4**	30.6	31.5
	Phi-3 Medium	29.4	**30.7**	30.5	29.4	30.3

Table 6.3: A comparison of our method against the baselines for the Phi-3 series.

Summarizer	Silver Reference LM(s)		
	System Pooling	Qwen 72B	Llama 70B
Qwen-2.5 0.5B	**21.3**	**21.3**	**21.3**
Qwen-2.5 1.5B	**28.2**	27.7	**28.2**
Qwen-2.5 3B	**32.7**	**32.7**	**32.7**
Qwen-2.5 7B	**36.9**	**36.9**	**36.9**
Qwen-2.5-1M 7B	32.9	**34.8**	32.9
Qwen-2.5 14B	36.2	35.4	**36.6**
Qwen-2.5-1M 14B	**36.6**	**36.6**	**36.6**
Qwen-2.5 32B	37.2	**37.8**	37.2
Qwen-2.5 72B	**35.3**	-	**35.3**
Llama-3.2 1B	25.4	**26.3**	25.4
Llama-3.2 3B	**31.8**	**31.8**	**31.8**
Llama-3.1 8B	33.7	33.7	**35.1**
Llama-3.3 70B	33.3	**34.7**	-
Phi-3 Mini	**29.9**	**29.9**	**29.9**
Phi-3 Small	31.5	**31.8**	28.3
Phi-3 Medium	30.3	**31.5**	27.7
ProLong 64K	32.3	**32.7**	32.6
ProLong 512K	**32.5**	32.0	**32.5**

Table 6.4: A comparison of system pooling against Qwen and Llama-based silver references (GTE 7B retriever). We don't compute estimates for Qwen 2.5 72B based on Qwen 2.5 72B silver references (and similarly for Llama 3.3 70B).

Silver Reference LM (full-context)	Count
Qwen-2.5 72B	33
Llama-3.3 70B	79
Jamba-1.5 Mini	54
Qwen-2.5-1M 14B	51
ProLong 512K	59
Total	276

Table 6.5: Counts of silver summaries from individual LMs post-MBR decoding. We pick top-3 summaries per input, so a total of 276 summaries.

Model	10%	25%	50%	75%	100%
Qwen-2.5 0.5B	21.3	21.3	21.3	21.3	21.3
Qwen-2.5 1.5B	27.7	28.2	27.7	27.7	27.7
Qwen-2.5 3B	32.7	32.7	32.7	32.7	32.7
Qwen-2.5 7B	36.9	36.9	36.9	36.9	36.8
Llama-3.2 1B	26.3	25.4	25.8	25.8	25.8
Llama-3.2 3B	31.8	31.4	31.4	31.4	31.4
Llama-3.1 8B	34.6	33.7	34.6	35.1	35.1

Table 6.6: A comparison of our method at various dataset sampling ratios (GTE 7B retriever).

Summarizer	Size	Supported	Estimated		
			RULER	HELMET	
				Summ	LongQA
Qwen-2.5 0.5B	0.5B	32,768	-	-	-
Qwen-2.5 1.5B	1.5B	32,768	-	32,768	16,384
Qwen-2.5 3B	3B	32,768	-	32,768	32,768
Qwen-2.5 7B	7B	131,072	32,768	65,536	16,384
Qwen-2.5-1M 7B	7B	1,010,000	65,536	131,072	131,072
Qwen-2.5 14B	14B	131,072	65,536	-	-
Qwen-2.5-1M 14B	14B	1,010,000	131,072	131,072	131,072
Qwen-2.5 32B	32B	131,072	65,536	-	-
Qwen-2.5 72B	72B	131,072	131,072	32,768	32,768
Llama-3.2 1B	1B	131,072	-	32,768	32,768
Llama-3.2 3B	3B	131,072	-	131,072	65,536
Llama-3.1 8B	8B	131,072	32,768	32,768	65,536
Llama-3.3 70B	70B	131,072	65,536	32,768	65,536
ProLong 64K	8B	65,536	-	-	-
ProLong 512K	8B	524,288	-	131,072	131,072
Phi-3 Mini	3B	131,072	32,768	65,536	65,536
Phi-3 Small	7B	131,072	-	32,768	65,536
Phi-3 Medium	14B	131,072	32,768	65,536	131,072
Jamba-1.5 Mini	52B-A13B	262,144	-	131,072	131,072

Table 6.7: A summary of LMs used in our work. We report the model size and context lengths (supported and estimated). For RULER and HELMET, we use the results reported in prior works to identify the context length estimates. Since our proposed context length estimate is also dependent on the retriever and dataset, we do not include those numbers here (see Table 6.8).

Summarizer	Full-context	RULER	HELMET Summ	HELMET LongQA	Ours
Retriever: GTE 1.5B					
Qwen-2.5 0.5B	16.7$_{\pm 1.5}$ (32K)	-	-	-	**20.6**$_{\pm 0.9}$ (8K)
Qwen-2.5 1.5B	26.3$_{\pm 0.8}$ (32K)	-	26.3$_{\pm 0.8}$ (32K)	**28.7**$_{\pm 1.2}$ (16K)	27.4$_{\pm 1.1}$ (8K)
Qwen-2.5 3B	29.5$_{\pm 0.2}$ (32K)	-	29.5$_{\pm 0.2}$ (32K)	29.5$_{\pm 0.2}$ (32K)	**30**$_{\pm 0.6}$ (8K)
Qwen-2.5 7B	34.1$_{\pm 1.1}$ (128K)	36.4$_{\pm 1.0}$ (32K)	34.5$_{\pm 0.9}$ (64K)	**37.6**$_{\pm 0.3}$ (16K)	37.2$_{\pm 0.9}$ (24K)
Qwen-2.5-1M 7B	32.1$_{\pm 0.3}$ (128K)	33.3$_{\pm 0.6}$ (64K)	32.1$_{\pm 0.3}$ (128K)	32.1$_{\pm 0.3}$ (128K)	**33.6**$_{\pm 0.4}$ (56K)
Qwen-2.5 14B	35.7$_{\pm 0.6}$ (128K)	35.6$_{\pm 0.7}$ (64K)	-	-	**37.4**$_{\pm 0.3}$ (24K)
Qwen-2.5-1M 14B	35.6$_{\pm 1.2}$ (128K)	35.6$_{\pm 1.2}$ (128K)	35.6$_{\pm 1.2}$ (128K)	35.6$_{\pm 1.2}$ (128K)	**37.4**$_{\pm 0.7}$ (24K)
Qwen-2.5 32B	33.9$_{\pm 0.7}$ (128K)	35.1$_{\pm 0.7}$ (64K)	-	-	**36.6**$_{\pm 0.6}$ (16K)
Qwen-2.5 72B	32.5$_{\pm 0.5}$ (128K)	32.5$_{\pm 0.5}$ (128K)	35$_{\pm 0.8}$ (32K)	35$_{\pm 0.8}$ (32K)	**36.3**$_{\pm 0.3}$ (24K)
Llama-3.2 1B	17.7$_{\pm 0.2}$ (128K)	-	24.6$_{\pm 0.6}$ (32K)	24.6$_{\pm 0.6}$ (32K)	**25.8**$_{\pm 1.9}$ (8K)
Llama-3.2 3B	28.7$_{\pm 1.4}$ (128K)	-	28.7$_{\pm 1.4}$ (128K)	**31.1**$_{\pm 0.5}$ (64K)	30.3$_{\pm 0.7}$ (56K)
Llama-3.1 8B	33.3$_{\pm 0.9}$ (128K)	**34.9**$_{\pm 0.8}$ (32K)	**34.9**$_{\pm 0.8}$ (32K)	34$_{\pm 0.5}$ (64K)	34.5$_{\pm 0.5}$ (40K)
Llama-3.3 70B	31.9$_{\pm 0.8}$ (128K)	33.2$_{\pm 0.4}$ (64K)	35.8$_{\pm 0.2}$ (32K)	33.2$_{\pm 0.4}$ (64K)	**35.9**$_{\pm 0.2}$ (40K)
ProLong 64K	24.9$_{\pm 0.6}$ (64K)	-	-	-	**32.2**$_{\pm 0.4}$ (16K)
ProLong 512K	31$_{\pm 0.8}$ (128K)	-	31$_{\pm 0.8}$ (128K)	31$_{\pm 0.8}$ (128K)	**32.3**$_{\pm 0.3}$ (48K)
Phi-3 Mini	11$_{\pm 0.3}$ (128K)	**30.6**$_{\pm 0.5}$ (32K)	30.4$_{\pm 0.1}$ (64K)	30.4$_{\pm 0.1}$ (64K)	**30.6**$_{\pm 0.4}$ (16K)
Phi-3 Small	27.8$_{\pm 1.3}$ (128K)	-	31.1$_{\pm 0.1}$ (32K)	30.3$_{\pm 0.9}$ (64K)	**31.9**$_{\pm 0.2}$ (48K)
Phi-3 Medium	29.4$_{\pm 1.3}$ (128K)	**30.7**$_{\pm 1.2}$ (32K)	29.9$_{\pm 0.1}$ (64K)	29.4$_{\pm 1.3}$ (128K)	**30.7**$_{\pm 1.2}$ (32K)
Retriever: GTE 7B					
Qwen-2.5 0.5B	17.3$_{\pm 0.4}$ (32K)	-	-	-	**21.3**$_{\pm 0.4}$ (8K)
Qwen-2.5 1.5B	26.8$_{\pm 0.3}$ (32K)	-	26.8$_{\pm 0.3}$ (32K)	27.7$_{\pm 0.6}$ (16K)	**28.2**$_{\pm 0.6}$ (24K)
Qwen-2.5 3B	30.2$_{\pm 0.2}$ (32K)	-	30.2$_{\pm 0.2}$ (32K)	30.2$_{\pm 0.2}$ (32K)	**32.7**$_{\pm 1.1}$ (16K)
Qwen-2.5 7B	34.1$_{\pm 1.1}$ (128K)	36.8$_{\pm 0.5}$ (32K)	34.9$_{\pm 0.4}$ (64K)	**36.9**$_{\pm 1.2}$ (16K)	**36.9**$_{\pm 1.2}$ (16K)
Qwen-2.5-1M 7B	32.1$_{\pm 0.3}$ (128K)	**32.9**$_{\pm 0.2}$ (64K)	32.1$_{\pm 0.3}$ (128K)	32.1$_{\pm 0.3}$ (128K)	**32.9**$_{\pm 0.2}$ (64K)
Qwen-2.5 14B	35.7$_{\pm 0.6}$ (128K)	35.4$_{\pm 0.9}$ (64K)	-	-	**36.2**$_{\pm 0.4}$ (16K)
Qwen-2.5-1M 14B	35.6$_{\pm 1.2}$ (128K)	35.6$_{\pm 1.2}$ (128K)	35.6$_{\pm 1.2}$ (128K)	35.6$_{\pm 1.2}$ (128K)	**36.6**$_{\pm 0.1}$ (48K)
Qwen-2.5 32B	33.9$_{\pm 0.7}$ (128K)	34.6$_{\pm 0.2}$ (64K)	-	-	**37.2**$_{\pm 0.7}$ (32K)
Qwen-2.5 72B	32.5$_{\pm 0.5}$ (128K)	32.5$_{\pm 0.5}$ (128K)	**35.9**$_{\pm 0.4}$ (32K)	**35.9**$_{\pm 0.4}$ (32K)	35.3$_{\pm 0.1}$ (24K)
Llama-3.2 1B	17.7$_{\pm 0.2}$ (128K)	-	24.9$_{\pm 0.3}$ (32K)	24.9$_{\pm 0.3}$ (32K)	**25.4**$_{\pm 0.7}$ (16K)
Llama-3.2 3B	28.7$_{\pm 1.4}$ (128K)	-	28.7$_{\pm 1.4}$ (128K)	29.7$_{\pm 0.2}$ (64K)	**31.4**$_{\pm 0.5}$ (32K)
Llama-3.1 8B	33.3$_{\pm 0.9}$ (128K)	**35.1**$_{\pm 0.2}$ (32K)	**35.1**$_{\pm 0.2}$ (32K)	33.7$_{\pm 0.4}$ (64K)	33.7$_{\pm 0.4}$ (56K)
Llama-3.3 70B	31.9$_{\pm 0.8}$ (128K)	34.4$_{\pm 0.5}$ (64K)	**35.8**$_{\pm 0.8}$ (32K)	34.4$_{\pm 0.5}$ (64K)	33.3$_{\pm 0.6}$ (80K)
ProLong 64K	25.9$_{\pm 0.6}$ (64K)	-	-	-	**32.3**$_{\pm 0.7}$ (32K)
ProLong 512K	31$_{\pm 0.8}$ (128K)	-	31$_{\pm 0.8}$ (128K)	31$_{\pm 0.8}$ (128K)	**32.5**$_{\pm 0.6}$ (32K)
Phi-3 Mini	11$_{\pm 0.3}$ (128K)	**29.9**$_{\pm 0.4}$ (32K)	28.3$_{\pm 1.4}$ (64K)	28.3$_{\pm 1.4}$ (64K)	**29.9**$_{\pm 0.4}$ (32K)
Phi-3 Small	27.8$_{\pm 1.3}$ (128K)	-	**32.4**$_{\pm 0.7}$ (32K)	30.6$_{\pm 1.3}$ (64K)	31.5$_{\pm 0.6}$ (24K)
Phi-3 Medium	29.4$_{\pm 1.3}$ (128K)	**30.7**$_{\pm 0.9}$ (32K)	30.5$_{\pm 0.5}$ (64K)	29.4$_{\pm 1.3}$ (128K)	30.3$_{\pm 1.4}$ (80K)

Table 6.8: Full set of results on the SummHay dataset. For each system, we report the average score and standard deviation across three runs. We also provide the (optimal) context length estimate used for each experiment configuration in parantheses.

Chapter 7

Context Augmented LM Judges for Long Text Summarization

7.1 Introduction

For complex user queries, NLP systems extract information by retrieving and synthesizing content from the Web. Such tasks are ubiquitous across domains, including news and scientific discovery. Progress in automated knowledge discovery has been accelerated with large language models. These systems are actively being used to answer general web queries, as well as to synthesize research reports in technical domains. For complex tasks, such as understanding the scientific literature, a reliable evaluation of model-generated responses is critical. Reference-based evaluation was the gold standard for text summarization, but due to the difficulty in obtaining high-quality references (Goyal et al., 2022b; Zhang et al., 2023), recent work pivoted to preference-based evaluation frameworks. In this setup, an LM or human judge labels a winner from a pair of model responses. Preference-based evaluations are scalable, do not rely on human-written references, and provide rankings of test models using Bradley-Terry or Elo ratings (Arena; (Chiang et al., 2024)). Although preference-based evaluation has been used as a common paradigm for many years, recent adoption was to evaluate the alignment of model responses with human preferences (Chiang et al., 2024), code completion (Chi et al., 2025), search (Miroyan et al., 2025), and scientific literature understanding and synthesis (Zhao et al., 2025).

Although preference-based evaluations are quite effective, the judge has limited context (if any) about the source documents used to generate the model responses. For knowledge-intensive tasks, it is important that the model response contains the necessary and sufficient information from the source documents. Using an expert-written reference is ideal, but it might be expensive to collect for long input tasks (Kim et al., 2024b). To this end, we present *Contextual-Arena*, which augments the LM judges with context information from the source documents. To counteract the long input lengths, we present two methods to compress the source documents into a context for the LM judge prompt: 1) summarize the source documents into a short text for use as explicit context and 2) rewrite the question by augmenting information from the source documents. Our idea of using source documents shares similarities with reference-free QA-style evaluation methods for text summarization (Scialom et al., 2021; Pratapa et al., 2023). Instead of generating a fixed set of QA pairs, we let the LM judge use the provided context in conjunction with question and the two responses.

In our experiments, we perform meta-evaluation of two preference-based evaluation frameworks, standard Arena and our proposed Contextual Arena. For a user question, the standard evaluation framework selects a winner from a pair of model responses. Our proposed framework also includes context information drawn from the source documents. In both Arenas, we autogenerate leaderboards for a given set of test models (summarization systems) by estimating Bradley-Terry ratings. We also computed confidence intervals for each rating using bootstrapping. We evaluated the leaderboards using standard metrics from the literature. Specifically, we adopt three metrics from Arena-Hard (Li et al., 2024a): model separability with confidence, agreement with confidence, and pair-rank Brier score. Separability measures the fraction of model pairs whose ratings are separable with confidence. Agreement measures the alignment (with confidence) between leaderboard ratings and the ratings estimated from a human (or any reference) judge. Pair rank Brier score measures the rank correlation (with confidence) between the estimated ratings and the human (or any reference) ratings. Together, these three metrics provide a holistic comparison of the baseline and the Contextual Arena.

We evaluated on two preference-based benchmarks: SciArena-Eval and SQuALITY-pref. SciArena-Eval (Zhao et al., 2025) consists of 2000 pairs of model responses to questions related to scientific literature search. SQuALITY-pref is a synthetic dataset that we compiled using a human-written reference summary to approximate human preference labels. We use an active sampling algorithm to select a set of model battles from the SQuALITY dataset (Wang et al., 2022) that provide sufficient model separability. We follow the taxonomy from Liu et al. (2025b) to select LM judges covering different paradigms of generation (scalar, generative) and scoring patterns (pointwise, pairwise). As our LM judges, we select variants of Skywork-v2 (Liu et al., 2025a), DeepSeek (DeepSeek-AI, 2024, 2025), and GPT-5 (OpenAI, 2025).

On the challenging SciArena-Eval benchmark, our results show that Contextual Arena significantly boosts model separability, while providing competitive agreement scores. This highlights its ability to separate models with fewer battles per model. On the SQuALITY-pref dataset, Contextual Arena achieves strong gains in agreement while matching reference model separability. Our results show varying effects of context across different LM judges.

7.2 Augmenting LM Judges with Compressed Context

To evaluate knowledge discovery tasks, model responses can be compared against an expert-written reference or directly against other model responses. For complex tasks with often long inputs, reference-based evaluation can be expensive (Kim et al., 2024b; Samaya-AI, 2025). Pairwise evaluations present a useful alternative in which a human or a LM judge contrasts two model-generated responses for the same user question. These evaluations are commonplace in Arena-style benchmarks (Chatbot Arena; (Chiang et al., 2024)) and have since been used for both model evaluation and reward modeling during post-training. In this work, we identify a gap in current pairwise evaluations for knowledge discovery tasks. Current methods only consider the user question when comparing two model-generated responses. This does not allow the judge (human or LM) access to the source documents that were originally used the models for their response generation. We hypothesize that adding context to the evaluation task by compiling information from these source documents will improve the effectiveness of the judges. We explore multiple text compression methods to augment context to the evaluation task. In this section, we give a brief overview of the standard pairwise evaluation task used in Arena-style benchmarks (subsection 7.2.1). Then, we describe our augmentation methods (subsection 7.2.2) and our LM judges (subsection 7.2.3).

7.2.1 Standard Arena: Pairwise Evaluation Task

Given a user question and two model-generated responses (A, B), a judge (human or LM) picks a winning response based on their accuracy and relevance to the question. This style of evaluation has been widely adopted by Arena-style benchmarks (Chiang et al., 2024; Chi et al., 2025). These benchmarks simulate battles between pairs of model responses, and use the results of these battles to estimate Elo or Bradley-Terry ratings (Bradley and Terry, 1952) for the models. We are particularly interested in search-augmented Arena benchmarks such as Search Arena (Miroyan et al., 2025) and SciArena (Zhao et al., 2025). In these evaluations, the systems first retrieve the relevant source documents using the question. The final responses are then generated from these retrieved documents.

Pairwise evaluations have been explored for summary evaluation, often including more than two model-generated summaries. These evaluations use the best-worst ratings to rank the model responses to the same question (Goyal et al., 2022b; Pratapa et al., 2023). In line with Arena-style benchmarks, we limit our focus to pairwise evaluations, but our methods can be extended to work with more than two responses.

7.2.2 Contextual Arena: Context-Augmented Pairwise Evaluation Task

The pairwise evaluation methodology used in Arena-style benchmarks does not include the source documents. Although this works well for evaluating instruction following abilities, lack of context about the source documents can be limiting for knowledge discovery tasks. The additional context could be useful to help augment underspecified questions, as well as to ground the model comparisons to content from the source documents. A related idea was explored by (Malaviya et al., 2025) for under-specified queries. They augment each question with a set of sub-question and answer pairs to improve the quality of model judgments. Instead, we utilize the underlying source documents that provide a direct signal to the judge. Since the total length of the source could be long, we explored two text compression-based methods to curate our context.

Context summary: We generate a short text summary of the source documents and use this as a context. This summary generation is agnostic to the question and is intended to provide the judge with an overview of the source documents. Note that for retrieval-based benchmarks such as SciArena, source documents are already filtered from a large document index. We consider this as an explicit context.

Augmented question: To closely resemble the standard pairwise evaluation task, we also explore including the context information directly in the user question. To achieve this, we rewrite the original user question to augment the information from the context summary generated above. Compared to the above method, this setup uses higher compression on the source documents. We consider this an an implicit context.

We use Qwen3 235B-A22B for both context summarization and question rewriting. See Table 7.10 in the Appendix for the system prompts.

7.2.3 LM judges

LM judges vary significantly in their use of question, model responses, and scoring methods. (Liu et al., 2025b) presents a taxonomy of reward models on two axes, the generation paradigm, and scoring patterns. In generation paradigms, reward models can be scalar, semi-scalar, or generative. For scoring the model responses, reward models can score individual responses (pointwise) or pairs of responses (pairwise). In this work, we use reward models as LM judges to score responses.

Generation paradigm: For a given query and responses, scalar models directly generate a score for the responses. Generative models generate a critique that could be used to select winning responses. Semi-scalar models generate both a critique and a score.

Scoring patterns: Pointwise methods provide a score separately for each response, while pairwise methods provide a relative score comparing the two responses.

In our experiments, we evaluated the effectiveness of our proposed context-augmented evaluation task in scalar/generative and pointwise/pairwise settings (subsection 7.3.4).

7.3 Experimental Setup

Given a set of model battles, our goal is to compare the standard and contextual arenas. Standard Arena uses the default pairwise evaluation task (question, responseA, responseB $\rightarrow$ winner), while Contextual Arena uses a context-augmented task (question|context, responseA, responseB $\rightarrow$ winner). First, we follow the standard procedures in the Arena-style benchmarks to estimate model ratings based on the results of pairwise model battles. We use the Bradley-Terry model (Bradley and Terry, 1952) to estimate model ratings and compute 95% confidence intervals using bootstrapping.[1] Finally, we use the model ratings to perform a meta-evaluation of the two arenas.

7.3.1 Metrics

For our meta-evaluation, we adopt Arena-Hard metrics (Li et al., 2024a). Given the LM judge model ratings (Standard, Contextual) and the human judge rating ratings, we calculate the model separability (with confidence), agreement with the human ratings (with confidence), and rank correlation between the LM and human ratings.

Separability: the percentage of model pairs with non-overlapping confidence intervals. The higher the separability of the model, the more useful the benchmark to distinguish close models.

[1] We use the Search Arena codebase to compute BT ratings and confidence intervals via boostrapping, `https://github.com/lmarena/search-arena`.

	Domain	# Battles	# Models	Model Separability
SciArena-Eval	Scientific literature	2000	23	37.5
SQuALITY-pref	Books	2500	10	91.1

Table 7.1: An overview of our preference datasets. We report the model separability score with reference preference labels.

Agreement: the agreement between two leaderboards (LM judge vs. human judge) to confidently distinguish two test models with the same ordering. If the two leaderboards can confidently separate two test models, they get a score of +1 (same preference) and -1 (opposite preference). If the LM judge cannot separate the two test models with confidence, they get a score of 0. The average score across all unique model pairs (separable by a human judge) is taken as the final score.

Pair rank Brier score: evaluates the ranking of pairs of competing models where higher confidence is rewarded for correct ranking and penalizing confidence for incorrect rankings. Unlike Spearman correlation, this metric factors the confidence intervals in the two leaderboards.

Following Li et al. (2024a), we use a combination of these metrics to get a holistic view of the pairwise evaluation tasks, Standard Arena and Contextual Arena.

7.3.2 Preference Dataset: SciArena-Eval

We used SciArena-Eval, an arena-style benchmark of 2000 battles between model responses collected using the SciArena platform (Zhao et al., 2025). Each battle has a winning response chosen by an expert human judge. The dataset spans 23 test models that include a mix of open-weight (Qwen, DeepSeek, Llama) and proprietary LMs (Gemini, Claude, GPT). Examples include model responses to scientific queries from four disciplines: natural sciences, healthcare, humanities & social sciences, and engineering. Each discipline contains 500 examples, with an equal distribution of votes between models A and B in the battles.

For a user-provided question, the authors used the ScholarQA engine to select up to 30 snippets of relevant publicly accessible scientific papers. These paper snippets could come from the abstract or the main text of a scientific paper. These snippets are passed to the test models to generate the final responses. For our augmented evaluation task, we use the set of paper snippets as the source documents.[2]

Other relevant datasets include Search Arena (Miroyan et al., 2025), BrowseComp (Wei et al., 2025). In this work, we focus on SciArena-Eval due to ease of availablity of the source documents used for model response generation.

7.3.3 Synthetic Preference Dataset: SQuALITY-pref

In addition to SciArena, we also compile a synthetic preference dataset using SQuALITY (Wang et al., 2022). SQuALITY is a long-document summarization dataset with question-focused sum-

[2] https://huggingface.co/datasets/yale-nlp/SciArena-with-paperbank

Model	Rating	CI Lower	CI Upper	Num Battles	Ranking	Response length
GPT5	1429.3	1367.6	1491.6	400	1	359 ± 58
GPT5 Mini	1302.8	1251.5	1350.7	355	2	340.6 ± 31.8
Llama4 Maverick	1209.7	1173.5	1251	521	3	300.6 ± 16.2
Qwen3 235B-A22B	1159.5	1127.3	1193.4	772	3	294.2 ± 15.4
Kimi K2	1111.1	1072.4	1139.2	702	4	240.5 ± 41.2
DeepSeek V3.1	1081.6	1048.3	1124.6	416	5	292.5 ± 17
Qwen3 30B-A3B	997.4	969.7	1028	426	7	300.1 ± 14.6
GPT5 Nano	929.3	891.7	963.2	581	8	356.6 ± 52.1
Qwen3 4B	837.2	791.8	880.8	586	9	304.1 ± 13.8
Llama4 Scout	719.5	650.5	798.9	241	9	302.2 ± 16.5

Table 7.2: Reference model ratings after cold-start and active sampling stages of our synthetic preference dataset (SQuALITY-pref). We also report the 95% confidence intervals.

maries of short stories from Project Gutenberg. Each input is annotated with multiple summaries from skilled writers. To curate a synthetic preference dataset, we simulate model battles in the SQuALITY dataset and pick a winner by prompting GPT-5-mini to compare system responses against a human-written reference (see Table 7.11 in the Appendix for the system prompt). This allows us to indirectly factor the source documents (aka context) into the decision process of picking a winner in a model battle.

For our synthetic dataset, we chose ten models that include variants of Qwen3, Llama4, GPT5, DeepSeek and Kimi-K2 (see Table 7.2). To create a diverse pool, we include open-weight and proprietary models of varying sizes. We first generate summaries for all the three splits of the SQuALITY dataset using these models. We then compile a subset of model battles using an active sampling algorithm that maximizes model separability in Bradley-Terry ratings. We briefly describe our sampling algorithm below.

Cold start: We uniformly sample a subset of examples from the dataset (N=400, without replacement), and for each example, we uniformly sample a pair of models for a battle. We estimate our initial BT ratings based on the results of these cold-start battles.

Active sampling: Given the current ratings, we estimate the relative ranking uncertainty between two models using the absolute overlap of their confidence intervals. At each sampling step, we select a model pair for battle according to a probability distribution proportional to these uncertainty estimates. For the chosen pair, we then sample an example uniformly from the SQuALITY dataset to simulate a battle. We repeat this step to collect 100 new battles, which we add to our synthetic preference dataset. Finally, we update the model ratings by reestimating with the expanded dataset. We repeat this procedure until we satisfy the following stop condition.

Stop condition: We stop the active sampling algorithm when we reach a pre-defined threshold for model separability (90%).

Through this procedure, we compiled a total of 2500 synthetic preference pairs from the train, validation, and test splits of the SQuALITY v1.3 dataset.

7.3.4 LM judges

We use the taxonomy of the reward models from (Liu et al., 2025b) to guide our selection of LM judges. We choose models that cover multiple reward generation paradigms (scalar and generative) and scoring patterns (pointwise and pairwise) (see Table 7.3). As a pointwise scalar judge, we pick Skywork-Llama-3.1, the best performing model in RewardBench v2 (as of Aug 28, 2025). For our pairwise generative models, we use DeepSeek V3.1 and R1(-0528), GPT-5-nano and GPT-5-mini.

LM-as-a-judge	Reward generation	Scoring pattern	Source
Skywork Llama-3.1	Scalar	Pointwise	Open-weight
DeepSeek V3.1	Generative	Pairwise	API
DeepSeek R1	Generative	Pairwise	API
GPT-5-nano	Generative	Pairwise	API
GPT-5-mini	Generative	Pairwise	API

Table 7.3: Our selection of LM judges. We follow the taxonomy from (Liu et al., 2025b) to categorize the reward generation and scoring patterns.

7.4 Results

On the SciArena-Eval and SQuALITY-pref benchmarks, we compare the Standard Arena setup (Q) with Contextual Arena, augmented with context summary (Q+C) and augmented question (Q^C). We report on the three metrics, separability, agreement, and the Brier score. Table 7.4 and Table 7.5 present the results on SciArena-Eval and SQuALITY-pref, respectively.

LM-as-a-judge	Separability ↑			Agreement ↑			Brier Score ↓		
	Q	Q+C	Q^C	Q	Q+C	Q^C	Q	Q+C	Q^C
Skywork Llama-3.1	49.8	–	60.9	77.9	–	87.4	0.27	–	0.27
DeepSeek V3.1	52.2	55.3	69.2	85.3	89.5	90.5	0.19	0.20	0.21
DeepSeek R1	57.7	56.1	68.8	92.6	88.4	87.4	0.18	0.22	0.22
GPT-5-nano	50.6	51.8	47.8	90.5	90.5	84.2	0.17	0.16	0.16
GPT-5-mini	49.4	60.5	55.7	93.7	89.5	92.6	0.17	0.23	0.21

Table 7.4: Results on SciArena-Eval dataset comparing the Standard Arena (Q) and our proposed Contextual Arena (Q+C, Q^C).

On the challenging SciArena-Eval benchmark with 23 test models, Contextual Arena shows significant improvements in model separability. For reference, the separability with human judgments is only 37.5 (Table 7.1). This illustrates the effectiveness of the Contextual Arena in

LM-as-a-judge	Separability ↑			Agreement ↑			Brier Score ↓		
	Q	Q+C	Q^C	Q	Q+C	Q^C	Q	Q+C	Q^C
Skywork Llama-3.1	86.7	–	86.7	61.0	–	63.4	0.24	–	0.23
DeepSeek V3.1	77.8	82.2	88.9	61.0	68.3	56.1	0.22	0.17	0.25
DeepSeek R1	86.7	80.0	88.9	58.5	68.3	53.7	0.26	0.19	0.25
GPT-5-nano	86.7	84.4	77.8	41.5	36.6	51.2	0.32	0.31	0.22
GPT-5-mini	71.1	86.7	86.7	65.9	80.5	65.9	0.11	0.14	0.19

Table 7.5: Results on SQuALITY-pref dataset comparing the Standard Arena (Q) and our proposed Contextual Arena (Q+C, Q^C).

achieving better separability with fewer battles per model. On SQuALITY-pref, with fewer models and high reference separability (Table 7.1), Contextual Arena almost matches the reference separability, while it gains substantial improvements in agreement compared to the Standard Arena baseline.

On SciArena-Eval, we generally see improvements in separability with Contextual Arena across all LM judges. With DeepSeek-V3.1 and Skywork, we also see improvements in agreement. We find that Contextual Arena performs poorly on the Brier score. On SQuALITY-pref, we again see improvments in separability with the exceptions of DeepSeek-R1 and GPT-5-nano. We see considerable improvements in agreement and Brier score with the Contextual Arena (Q+C) setup.

For Skywork, a point-wise and scalar reward model, the additional context provided by Contextual Arena results in gains on both datasets. We see mixed results with generative and pairwise reward models. In some cases, an explicit context (Q+C) helps, while in other cases an implicit context (Q^C) is more beneficial.

7.5 Analysis

In this section, we analyze the effectiveness of our Contextual Arena in various settings.

SciArena-Eval vs. SQuALITY-pref In addition to the differences described in Table 7.1, these two datasets also differ in the source documents used for context compression. In SciArena, we use the documents retrieved by the ScholarQA engine, whereas we use the full-story text for SQuALITY-pref. Therefore, the context generated from SciArena tends to be more directly related to the user question and leads to better gains.

Pointwise vs. Pairwise Pointwise LM judges score a model response based solely on the user's question. On the other hand, pairwise LM judges can contrast two model responses to pick a winner. Since the original question only provides a limited signal, we found that augmenting context with compressed source documents can lead to large gains (Skywork+Q^C in Table 7.8, Table 7.9). We found significant gains in SciArena with separability and agreement. On SQuALITY-pref, we match the baseline separability while providing small gains in the agreement. The differences in performance gains across the two datasets could be attributed to our previous comparison of the source documents in the two datasets.

Effect of context on model judgment We perform a qualitative analysis of the reasoning trace of generative LM judges to understand the impact of the provided context. We inspect SciArena examples with GPT-5-mini and DeepSeek-V3.1 judges and observe how the judges use the explicit context (see Table 7.6). The reasoning traces highlight the alignment between the winning response and the provided context, often superceding other limitations. In Table 7.7, we also report the impact of the implicit context. We find that the implicit context could lead to an erroneous judgment.

Context for saliency vs. faithfulness Faithfulness evaluation methods compare the response generated by the model with the source documents. Although our compressed context (subsection 7.2.2) could be a useful signal for the evaluation of faithfulness, it is not sufficient to verify all the facts in the response to the model. We refer the reader to previous work on faithfulness evaluation (Kim et al., 2024b).

7.6 Conclusion & Future Work

In this work, we propose a modification to contextualize pairwise evaluations in standard Arena-style benchmarks. We present two methods to augment compressed information from the source documents for the pairwise evaluation task. On SciArena-Eval benchmark, we show that our method improves model separability but suffers from lower agreement with human judgments. Some future work directions include evaluating the impact of context when collecting human judgments and training reward models for knowledge discovery tasks that include context information in the training examples.

7.7 Appendix

7.7.1 Computing ratings and metrics

We used the Bradley-Terry model to estimate model ratings. For this, we use the implementation of Search Arena (Miroyan et al., 2025).[3] For our metrics, we use the implementation of Arena-Hard (Li et al., 2024a).[4]

7.7.2 Additional Metrics

Although we are primarily interested in the overall effectiveness of Standard Arena and Contextual Arena benchmarks, we also report the accuracy of LM judge predictions for individual

[3] `https://github.com/lmarena/search-arena`

[4] `https://github.com/lmarena/arena-hard-auto`

pairs of responses. In Table 7.8 and Table 7.9, we report the accuracy scores in SciArena and SQuALITY-pref, respectively.

7.7.3 Compression

In Table 7.10, we provide the prompts used for context summarization and question rewrite tasks.

7.7.4 Synthetic preference dataset

Table 7.11 provides our system prompt to collect labels for the synthetic preference dataset (subsection 7.3.3). We only include the first human-written reference summary. Future work could look at augmenting more than one reference in the system prompt.

7.7.5 LM judges

Table 7.12 provides the prompts used with our LM-as-a-judge setup on SciArena-Eval. For the SQuALITY-pref, we make a slight modification at the start of the prompt to: 'Your task is to evaluate the quality of two AI-generated responses to a user's question. Assess both responses for relevance, accuracy and clarity....'

Example 1: explicit context helps
→ Original judgment Both outputs are relevant and clear. - Output A provides more detail on the physical-digital consumption trade-off (a nuanced point), but its citations are highly repetitive and potentially misleading. - Output B is slightly more concise, includes an additional source (Liu et al.), and covers a broader range of challenges (e.g., commodification). However, it also has citation issues. The citation practices in both outputs are flawed, but Output B is marginally better due to the inclusion of Liu et al. (2024a) and slightly more balanced coverage of benefits and challenges. ### Final Answer: \boxed{B}
→ Augmented judgment Output A is superior because it is more focused on consumption, accurately synthesizes the context (especially the physical-digital dynamic), and uses citations more appropriately to support specific claims. \boxed{A}
Example 2: explicit context helps
→ Original judgment While Output B has broader coverage, the citation inaccuracies make it less trustworthy. Output A, though narrower, is more precise and correctly attributed. Therefore, **Output A is the better response**. \boxed{A}
→ Augmented judgment Output B is **more comprehensive** (covers S2L and complementarity, which are important in the context) and thus better addresses the breadth of the question. However, its citation formatting is flawed. But since the content is accurate and the user question asks for "various strategies," Output B's broader coverage wins. \boxed{B}

Table 7.6: Qualitative analysis of LM judge's reasoning trace. In both these examples, the explicit context provided to the LM judge was critical for the judgment task.

Example 1: a failure case for using implicit context
$\rightarrow$ Question How has the trade-off between language model performance and computational efficiency evolved in recent research?
$\rightarrow$ Original judgment Output A is superior due to its clearer structure, more credible citations (no anachronistic 2025 references), tighter focus on the core trade-off, and better synthesis of recent research trends. Output B, while detailed, suffers from questionable citations and less coherent organization. \boxed{A}
$\rightarrow$ Augmented Question How has the balance between language model performance, computational efficiency, and adversarial robustness evolved with recent advancements in efficient architectures and evaluation frameworks?
$\rightarrow$ Augmented judgment Output B is superior because it directly addresses all three components of the user's question (performance, computational efficiency, and adversarial robustness), whereas Output A ignores adversarial robustness entirely. Output B also provides more quantitative details and a broader perspective on the trade-offs, including privacy and component-level efficiency. \boxed{B}

Table 7.7: Qualitative analysis of LM judge's reasoning trace when provided with implicit context.

LM-as-a-judge	Accuracy ↑		
	Q	Q+C	Q^C
Skywork Llama-3.1	59.0	–	58.8
DeepSeek V3.1	62.0	62.5	59.5
DeepSeek R1	60.4	60.2	59.2
GPT-5-nano	56.7	55.6	55.6
GPT-5-mini	59.4	61.0	59.5

Table 7.8: Results on SciArena-Eval dataset comparing the Standard Arena (Q) and our proposed Contextual Arena (Q+C, Q^C).

LM-as-a-judge	Accuracy ↑ Q	Q+C	Q^C
Skywork Llama-3.1	58.4	–	54.4
DeepSeek V3.1	49.4	55.2	50.3
DeepSeek R1	58.4	49.0	57.5
GPT-5-nano	53.5	54.8	54.5
GPT-5-mini	75.2	70.5	69.7

Table 7.9: Results on SQuALITY-pref dataset comparing the Standard Arena (Q) and our proposed Contextual Arena (Q+C, Q^C).

Context compression prompt
{document} Write a summary for the above document. Generate a {num words} word response.
Question rewrite prompt
{context summary} Question: {question} Using the content from the text above, rewrite the question to make it more contextual. Only output the question, no additional text.

Table 7.10: Prompts to Qwen3-235B-A22B model for context compression and question rewrite tasks (subsection 7.2.2).

Your task is to evaluate the quality of two AI-generated responses to a user's question. You are also provided with a reference response written by an expert. Assess both responses for relevance, accuracy and clarity. Then, select the response, Output A or Output B, that best address the user's question. User Question: {question} Reference Output: {reference} Output A: {response A} Output B: {response B} Which is best, Output A or Output B? Please reason step by step, and put your final answer (A or B) within `\boxed{}`.

Table 7.11: Prompt for GPT judge to curate synthetic preference dataset. In contrast to our evaluation prompts to LM judges in Table 7.12, we provide an expert-written reference summary to the GPT judge. We use the responses from this reference-augmented GPT judge as our reference labels for our SQuALITY-pref dataset (subsection 7.3.3).

Arena-1 (Q): LM judge prompt
You are an expert in scientific literature synthesis. Your task is to evaluate the quality of two AI-generated citation-attributed responses to a user's question. Assess both responses for relevance, accuracy, clarity, and appropriate use of citations. Then, select the response, Output A or Output B, that best address the user's question. User Question: {question} Output A: {response A} Output B: {response B} Which is best, Output A or Output B? Please reason step by step, and put your final answer (A or B) within `\boxed{}`.
Arena-2 (Q+C): LM judge prompt with augmented context summary
You are an expert in scientific literature synthesis. Your task is to evaluate the quality of two AI-generated citation-attributed responses to a user's question. You are also provided with a summary of the context that was available to the AI systems. Assess both responses for relevance, accuracy, clarity, and appropriate use of citations. Then, select the response, Output A or Output B, that best address the user's question. User Question: {question} Context: {context} Output A: {response A} Output B: {response B} Which is best, Output A or Output B? Please reason step by step, and put your final answer (A or B) within `\boxed{}`.
Arena-2 (Q^C): LM judge prompt with context-augmented question
You are an expert in scientific literature synthesis. Your task is to evaluate the quality of two AI-generated citation-attributed responses to a user's question. Assess both responses for relevance, accuracy, clarity, and appropriate use of citations. Then, select the response, Output A or Output B, that best address the user's question. User Question: {augmented question} Output A: {response A} Output B: {response B} Which is best, Output A or Output B? Please reason step by step, and put your final answer (A or B) within `\boxed{}`.

Table 7.12: Prompts used with LM judges for the SciArena-Eval dataset.

Chapter 8

Conclusion & Future Work

In this thesis, we studied multiple subproblems relating to event modeling in multi-document settings. In the early chapters, we defined the identity relationship between event mentions and explored methods to capture such event-event relations across documents. We presented coreference resolution and grounding as tools to jointly model within-document and cross-document relations. In the latter chapters, we utilized recent advancements in language modeling to tackle complex multi-document tasks that require search and synthesis over long inputs. Specifically, we look at query-focused text summarization tasks. We presented a novel benchmark that requires generating background summaries for complex news events. For such complex multi-document summarization tasks, we contrasted the effectiveness of input compression and long-context LMs. Our results highlighted the need for hybrid approaches, and we further proposed a hybrid RAG method that combines the strengths of retrieval with long-context windows supported by recent LMs. Finally, we present methods to improve preference-based evaluation of long text summarization systems. We evaluated the effectiveness of including context from source documents in Arena-style evaluations.

The works presented in this thesis highlight some key findings. First, there is considerable variance in the nature of multi-document tasks. Some important dimensions include the length of the input, the spread of the information, the specificity of the question, and the domain. Model performance varies widely across these dimensions. Second, we find evidence that input compression remains useful in the current model settings. Models vary considerably in their long-context capabilities, and verifying their performance on both standard benchmarks and specific downstream tasks is important. Input compression could also serve as a useful tool to improve input comprehension in evaluation tasks. Third, we need better content selection evaluation methods for knowledge-intensive tasks. Some options here include QA-based or context-augmented preference-based evaluations.

There are many directions for future work that can build on the work presented in this thesis. Here, we highlight some open questions: 1) building efficient multi-agent systems that combine search, reasoning, long context, and compression for tasks with varying levels of complexity; 2) memory-augmented systems that are customizable to individual users; 3) reliable evaluation methods (human or automated) for complex search and synthesis tasks to use at various stages of model development.

With recent improvements in long-context and agentic capabilities of LMs, a single system (with multiple underlying LMs) can be deployed for complex search and synthesis tasks. This provides new opportunities to tackle event understanding in a large-scale multi-document setting. Top proprietary LMs such as GPT, Gemini, and Claude support such capabilities through their Research tools. Given a user question, these systems first retrieve relevant information from the Web before summarizing the findings into a long-form response. Typically, these systems go through multiple iterations of retrieval and reasoning before generating the final response. However, these systems have some limitations. First, multi-agent systems like Claude Research consist of a combination of a lead agent (Opus) and multiple subagents (Sonnet), and utilize 15 times more tokens compared to a chat-based system.[1] Future work could improve token efficiency by selectively choosing between long-context and compression (chapter 5, chapter 6). Second, evaluating multi-agent systems for search and synthesis tasks remains a challenging problem. Current evaluation methods such as BrowseComp (Wei et al., 2025) consist of complex queries with easy-to-verify responses. For hard-to-verify responses, expert-curated checklists are an alternative. To curate checklists, we need tools that help experts digest long inputs in a timely manner. Our work in chapter 7 that uses input compression is a useful step in this direction. Overall, long-context LMs equipped with text compression tools could provide a scalable solution for deploying LMs for complex multi-document understanding tasks.

Another important direction would be to customize document understanding systems for individual users. For example, a background summary of a news event (chapter 4) should ideally factor in information already known to the user. This requires the system to keep track of each user's knowledge about the event through their previous interactions with the system. Memory-augmented LMs are a step forward in this direction that allows a LM to look into its memory of past interactions and user-specified preferences and combine it with the new information from the source documents. A unique challenge here will be to perform multi-document reasoning over a heterogeneous set of documents. For example, an LM deployed in an enterprise setting with access to user data could merge information from previous chats, calendar, emails, and any additional documents stored by the user. For memory-augmented systems, we will need a combination of text compression and long-context reasoning from LMs (chapter 6). Deciding the ratio of past and new information mix into a single context window of an LM could have a considerable impact on downstream system performance.

A third future work direction will be to develop reliable evaluation methods for use in different stages of model development. Automated, fast, and verifiable evaluations are key during model development. During model deployment, comprehensive evaluations that accurately estimate real-world performance are critical. For multi-document summarization, we explored LM-based evaluation through novel QA-style metrics (chapter 4) and Arena-style evaluations with LM-as-a-judge (chapter 7). Although both methodologies are valuable during model development, future work could focus on comprehensive human-in-the-loop methods for use in model deployment. As we highlighted earlier, a key challenge with such evaluations would be the length of source documents used during model generation. To help humans assimilate long inputs, we could provide search tools that use input compression to present the necessary information to the human judge in an accessible manner (chapter 7). Another interesting challenge here would be aligning the evaluation objectives used in the development and deployment stages. With LMs be-

[1] https://www.anthropic.com/engineering/multi-agent-research-system

coming ubiquitous in real-world user interfaces, especially in technical and potentially sensitive domains, reliable evaluations are critical to tracking system performance. Therefore, this future work direction is the most important among the three directions highlighted in this chapter.

Bibliography

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219. 6.3.3, 6.5.2

James Allan, Jaime G. Carbonell, George R. Doddington, Jonathan Yamron, and Yiming Yang. 1998. Topic detection and tracking pilot study final report. 1, 5.1

Jun Araki and Teruko Mitamura. 2018. Open-domain event detection using distant supervision. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 878–891, Santa Fe, New Mexico, USA. Association for Computational Linguistics. 2.1, 2.3

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *Preprint*, arXiv:1907.05019. 3.5

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596. 2.4.3

Akari Asai, Jungo Kasai, Jonathan Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2021. XOR QA: Cross-lingual open-retrieval question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 547–564, Online. Association for Computational Linguistics. 3.2.4

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*. 6.1

Javed Aslam, Fernando Diaz, Matthew Ekstrand-Abueg, Richard McCreadie, Virgil Pavlu, and Tetsuya Sakai. 2015. Trec 2015 temporal summarization track overview. In *TREC*. 1, 4.1, 4.2, 5.1

Giusepppe Attardi. 2015. WikiExtractor. 3.3.1

Cosmin Bejan and Sanda Harabagiu. 2008. A linguistic resource for discovering event structures and resolving event coreference. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA). 1, 2.2

Amanda Bertsch, Alex Xie, Graham Neubig, and Matthew Gormley. 2023. It‘s MBR all the way down: Modern generation techniques through the lens of minimum Bayes risk. In *Proceedings of the Big Picture Workshop*. Association for Computational Linguistics. 6.2.2

Giang Binh Tran, Mohammad Alrifai, and Dat Quoc Nguyen. 2013. Predicting relevant news events for timeline summaries. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13 Companion, page 91–92, New York, NY, USA. Association for Computing Machinery. 1, 4.1, 4.2, 4.3.2, 5.6.1

Leo Born, Maximilian Bacher, and Katja Markert. 2020. Dataset reproducibility and IR methods in timeline summarization. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1763–1771, Marseille, France. European Language Resources Association. 4.2

Jan A. Botha, Zifei Shan, and Daniel Gillick. 2020. Entity Linking in 100 Languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7833–7845, Online. Association for Computational Linguistics. 3.1, 3.2.1, 3.2.4, 3.3, 3.3.1, 3.5.3

Ralph Allan Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345. 7.2.1, 7.3

Dorian Brown. 2020. Rank-BM25: A Collection of BM25 Algorithms in Python. 3.4.1

Michael Bugert, Nils Reimers, Shany Barhom, Ido Dagan, and Iryna Gurevych. 2020. Breaking the subtopic barrier in cross-document event coreference resolution. In *Proceedings of Text2Story - Third Workshop on Narrative Extraction From Texts co-located with 42nd European Conference on Information Retrieval, Text2Story@ECIR 2020, Lisbon, Portugal, April*

14th, 2020 [online only], volume 2593 of *CEUR Workshop Proceedings*, pages 23–29. CEUR-WS.org. 2.2

Razvan Bunescu and Marius Paşca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 9–16, Trento, Italy. Association for Computational Linguistics. 3.2.1

Taylor Cassidy, Bill McDowell, Nathanael Chambers, and Steven Bethard. 2014. An annotation framework for dense event ordering. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 501–506, Baltimore, Maryland. Association for Computational Linguistics. 2.1

Arie Cattan, Alon Eirew, Gabriel Stanovsky, Mandar Joshi, and Ido Dagan. 2021. Realistic evaluation principles for cross-document coreference resolution. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 143–151, Online. Association for Computational Linguistics. 20

Khyathi Raghavi Chandu, Yonatan Bisk, and Alan W Black. 2021. Grounding 'grounding' in NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4283–4305, Online. Association for Computational Linguistics. 3.1

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. Booookscore: A systematic exploration of book-length summarization in the era of LLMs. In *The Twelfth International Conference on Learning Representations*. 5.2.2, 5.3.1, 5.3.6

Mingda Chen, Zewei Chu, Sam Wiseman, and Kevin Gimpel. 2022. SummScreen: A dataset for abstractive screenplay summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8602–8615, Dublin, Ireland. Association for Computational Linguistics. 4.2, 4.2, 4.6.3

Muhao Chen, Hongming Zhang, Qiang Ning, Manling Li, Heng Ji, Kathleen McKeown, and Dan Roth. 2021. Event-centric natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Tutorial Abstracts*, pages 6–14, Online. Association for Computational Linguistics. 1, 4.1, 4.2

Tong Chen, Hongwei Wang, Sihao Chen, Wenhao Yu, Kaixin Ma, Xinran Zhao, Hongming Zhang, and Dong Yu. 2024. Dense X retrieval: What retrieval granularity should we use? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. 6.3.2

Wayne Chi, Valerie Chen, Anastasios Nikolas Angelopoulos, Wei-Lin Chiang, Aditya Mittal, Naman Jain, Tianjun Zhang, Ion Stoica, Chris Donahue, and Ameet Talwalkar. 2025. Copilot arena: A platform for code LLM evaluation in the wild. In *Forty-second International Conference on Machine Learning*. 7.1, 7.2.1

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. 4.3.5

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *Preprint*, arXiv:2403.04132. 7.1, 7.2, 7.2.1

Hai Leong Chieu and Yoong Keok Lee. 2004. Query Based Event Extraction along a Timeline. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, page 425–432, New York, NY, USA. Association for Computing Machinery. 1, 4.1, 4.2, 5.1

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *Preprint*, arXiv:2210.11416. 4.1, 4.4

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470. 3.2.4

Cohere For AI. 2024. c4ai-command-r-08-2024. 5.2.3

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics. 3.1, 3.4.2

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics. 3.2.4

Agata Cybulska and Piek Vossen. 2014. Using a sledgehammer to crack a nut? lexical diversity and event coreference resolution. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4545–4552, Reykjavik, Iceland. European Language Resources Association (ELRA). 1, 2.2

Hoa Dang and Karolina Owczarzak. 2009. Overview of the TAC 2008 Update Summarization Task. 1, 4.1, 4.2, 5.1

Hoa Trang Dang. 2005. Overview of duc 2005. 4.1

Hoa Trang Dang. 2007. Document understanding conference. 1

DeepSeek-AI. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437. 7.1

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948. 7.1

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics. 3.1, 3.4.2

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. 2024. LongroPE: Extending LLM context window beyond 2 million tokens. In *Forty-first International Conference on Machine Learning*. 5.3.3

Seth Ebner, Patrick Xia, Ryan Culkin, Kyle Rawlins, and Benjamin Van Durme. 2020. Multi-sentence argument linking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8057–8077, Online. Association for Computational Linguistics. 2.6

Alon Eirew, Arie Cattan, and Ido Dagan. 2021. WEC: Deriving a large-scale cross-document event coreference dataset from Wikipedia. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2498–2510, Online. Association for Computational Linguistics. 2.2, 3.1, 3.2.3, 3.3

Yotam Eshel, Noam Cohen, Kira Radinsky, Shaul Markovitch, Ikuya Yamada, and Omer Levy. 2017. Named entity disambiguation for noisy text. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 58–68, Vancouver, Canada. Association for Computational Linguistics. 3.1, 3.2.1

Tianyu Gao, Alexander Wettig, Howard Yen, and Danqi Chen. 2024. How to train long-context language models (effectively). *Preprint*, arXiv:2410.02660. 6.2.1, 6.3.3, 6.5.1

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics. 2.3

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 *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1302–1308, Online. Association for Computational Linguistics. 5.1, 5.2.1

Demian Gholipour Ghalandari and Georgiana Ifrim. 2020. Examining the state-of-the-art in news timeline summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1322–1334, Online. Association for Computational Linguistics. 4.2

John Giorgi, Luca Soldaini, Bo Wang, Gary Bader, Kyle Lo, Lucy Wang, and Arman Cohan. 2023. Open domain multi-document summarization: A comprehensive study of model brittleness under retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8177–8199, Singapore. Association for Computational Linguistics. 6.2

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022a. News summarization and evaluation in the era of gpt-3. *Preprint*, arXiv:2209.12356. 1, 4.1, 4.3.3, 4.3.5, 4.6, 5.3.5

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022b. News summarization and evaluation in the era of gpt-3. *Preprint*, arXiv:2209.12356. 7.1, 7.2.1

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide

Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh

Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783. 6.1, 6.2.1, 6.3.3

Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2022. LongT5: Efficient text-to-text transformer for long sequences. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 724–736, Seattle, United States. Association for Computational Linguistics. 4.1, 4.4

Qi Guo, Fernando Diaz, and Elad Yom-Tov. 2013. Updating users about time critical events. In *Advances in Information Retrieval*, pages 483–494, Berlin, Heidelberg. Springer Berlin Heidelberg. 1, 4.2

Hiroaki Hayashi, Wojciech Kryściński, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2020. What's new? summarizing contributions in scientific literature. *Preprint*, arXiv:2011.03161. 4.2, 4.6.3

Eduard Hovy, Teruko Mitamura, Felisa Verdejo, Jun Araki, and Andrew Philpot. 2013. Events are not simple: Identity, non-identity, and quasi-identity. In *Workshop on Events: Definition, Detection, Coreference, and Representation*, pages 21–28, Atlanta, Georgia. Association for Computational Linguistics. 1, 2.1, 2.2, 2.4.1, 16, 2.5

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. 2024. RULER: What's the real context size of your long-context language models? In *First Conference on Language Modeling*. 5.1, 5.2.2, 5.2.3, 5.3.6, 6.1, 6.1, 6.3.4

Kung-Hsiang Huang, Philippe Laban, Alexander Fabbri, Prafulla Kumar Choubey, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. 2024. Embrace divergence for richer insights: A multi-document summarization benchmark and a case study on summarizing diverse information from news articles. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 570–593, Mexico City, Mexico. Association for Computational Linguistics. 5.3.1

Heng Ji and Ralph Grishman. 2011. Knowledge base population: Successful approaches and challenges. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1148–1158, Portland, Oregon, USA. Association for Computational Linguistics. 3.1

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2024. LongLLMLingua: Accelerating and enhancing LLMs in long context scenarios via prompt compression. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1658–1677, Bangkok, Thailand. Association for Computational Linguistics. 5.3.6, 6.6

Bowen Jin, Jinsung Yoon, Jiawei Han, and Sercan O Arik. 2025. Long-context LLMs meet RAG: Overcoming challenges for long inputs in RAG. In *The Thirteenth International Conference on Learning Representations*. 6.1, 6.1, 6.2, 6.3.2

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics. 3.2.4

Greg Kamradt. 2023. Needle in a haystack - pressure testing llms. 5.1

Marzena Karpinska, Katherine Thai, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024a. One thousand and one pairs: A "novel" challenge for long-context language models. *Preprint*, arXiv:2406.16264. 5.3.4, 5.3.6, 5.5

Marzena Karpinska, Katherine Thai, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024b. One thousand and one pairs: A “novel” challenge for long-context language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. 6.6

Daniel Khashabi, Gabriel Stanovsky, Jonathan Bragg, Nicholas Lourie, Jungo Kasai, Yejin Choi, Noah A. Smith, and Daniel Weld. 2022. GENIE: Toward reproducible and standardized human evaluation for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11444–11458, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. 4.6

Yekyung Kim, Yapei Chang, Marzena Karpinska, Aparna Garimella, Varun Manjunatha, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024a. FABLES: Evaluating faithfulness and content selection in book-length summarization. In *First Conference on Language Modeling*. 5.3.1

Yekyung Kim, Yapei Chang, Marzena Karpinska, Aparna Garimella, Varun Manjunatha, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024b. FABLES: Evaluating faithfulness and content selection in book-length summarization. In *First Conference on Language Modeling*. 7.1, 7.2, 7.5

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470, Vancouver, Canada. Association for Computational Linguistics. 4.5.2, 5.3.5

Hayato Kobayashi. 2018. Frustratingly easy model ensemble for abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 6.2.2

Kalpesh Krishna, Erin Bransom, Bailey Kuehl, Mohit Iyyer, Pradeep Dasigi, Arman Cohan, and Kyle Lo. 2023. LongEval: Guidelines for human evaluation of faithfulness in long-form summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1650–1669, Dubrovnik, Croatia. Association for Computational Linguistics. 4.2

Shankar Kumar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 169–176, Boston, Massachusetts, USA. Association for Computational Linguistics. 6.1

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023a. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*. Association for Computing Machinery. 6.3.3

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023b. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*. 5.6.2

Philippe Laban, Alexander Fabbri, Caiming Xiong, and Chien-Sheng Wu. 2024a. Summary of a haystack: A challenge to long-context LLMs and RAG systems. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. 6.1, 6.1, 6.3.1

Philippe Laban, Alexander R. Fabbri, Caiming Xiong, and Chien-Sheng Wu. 2024b. Summary of a haystack: A challenge to long-context llms and rag systems. *Preprint*, arXiv:2407.01370. 5.1, 5.2.1, 5.3.1, 5.6.5

Adam Lerer, Ledell Wu, Jiajun Shen, Timothee Lacroix, Luca Wehrstedt, Abhijit Bose, and Alex Peysakhovich. 2019. Pytorch-biggraph: A large scale graph embedding system. In *Proceedings of Machine Learning and Systems*, volume 1, pages 120–131. 3.4.2

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. 2024a. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. *Preprint*, arXiv:2406.11939. 7.1, 7.3.1, 7.7.1

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *Preprint*, arXiv:2308.03281. 6.3.2

Zhuowan Li, Cheng Li, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. 2024b. Retrieval augmented generation or long-context LLMs? a comprehensive study and hybrid approach. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*. 6.1

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics. 4.5.1, 5.6.3

Bang Liu, Di Niu, Kunfeng Lai, Linglong Kong, and Yu Xu. 2017. Growing story forest online from massive breaking news. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, page 777–785. Association for Computing Machinery. 1, 4.2, 4.3

Chris Yuhao Liu, Liang Zeng, Yuzhen Xiao, Jujie He, Jiacai Liu, Chaojie Wang, Rui Yan, Wei Shen, Fuxiang Zhang, Jiacheng Xu, Yang Liu, and Yahui Zhou. 2025a. Skywork-reward-v2: Scaling preference data curation via human-ai synergy. *arXiv preprint arXiv:2507.01352*. 7.1

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023a. G-eval: Nlg evaluation using gpt-4 with better human alignment. *Preprint*, arXiv:2303.16634. 4.4, 4.6.1

Yixin Liu, Alex Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023b. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4140–4170, Toronto, Canada. Association for Computational Linguistics. 5.3.1, 6.3.1

Yixin Liu, Alexander Fabbri, Yilun Zhao, Pengfei Liu, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023c. Towards interpretable and efficient automatic reference-based summarization evaluation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16360–16368, Singapore. Association for Computational Linguistics. 5.1, 5.3.1, 6.3.1

Zhengzhong Liu, Guanxiong Ding, Avinash Bukkittu, Mansi Gupta, Pengzhi Gao, Atif Ahmed, Shikun Zhang, Xin Gao, Swapnil Singhavi, Linwei Li, Wei Wei, Zecong Hu, Haoran Shi, Xiaodan Liang, Teruko Mitamura, Eric Xing, and Zhiting Hu. 2020. A data-centric framework for composable NLP workflows. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 197–204, Online. Association for Computational Linguistics. 2.3, 2.4.2

Zijun Liu, Peiyi Wang, Runxin Xu, Shirong Ma, Chong Ruan, Peng Li, Yang Liu, and Yu Wu. 2025b. Inference-time scaling for generalist reward modeling. *Preprint*, arXiv:2504.02495. 7.1, 7.2.3, 7.3.4, 7.3

Robert Logan, Nelson F. Liu, Matthew E. Peters, Matt Gardner, and Sameer Singh. 2019. Barack's wife hillary: Using knowledge graphs for fact-aware language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5962–5971, Florence, Italy. Association for Computational Linguistics. 3.1

Lajanugen Logeswaran, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Jacob Devlin, and Honglak Lee. 2019. Zero-shot entity linking by reading entity descriptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3449–3460, Florence, Italy. Association for Computational Linguistics. 3.1, 3.2.1, 3.3.1, 3.3.1, 3.3.2, 12, 3.5

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*. 2.6, 3.7.3

Annie Louis. 2014. A Bayesian method to incorporate background knowledge during automatic text summarization. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 333–338, Baltimore, Maryland. Association for Computational Linguistics. 4.2

Annie Louis and Ani Nenkova. 2013. Automatically assessing machine summary content without a gold standard. *Computational Linguistics*, 39(2):267–300. 4.3.5

Yi Lu, Jing Nathan Yan, Songlin Yang, Justin T. Chiu, Siyu Ren, Fei Yuan, Wenting Zhao, Zhiyong Wu, and Alexander M. Rush. 2024. A controlled study on long context extension and generalization in llms. *Preprint*, arXiv:2409.12181. 5.3.3

Yuanhua Lv and ChengXiang Zhai. 2011a. Lower-bounding term frequency normalization. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, page 7–16, New York, NY, USA. Association for Computing Machinery. 3.1, 3.4.1

Yuanhua Lv and ChengXiang Zhai. 2011b. When documents are very long, bm25 fails! In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, page 1103–1104, New York, NY, USA. Association for Computing Machinery. 3.4.1

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics. 2.4.2

Chaitanya Malaviya, Joseph Chee Chang, Dan Roth, Mohit Iyyer, Mark Yatskar, and Kyle Lo. 2025. Contextualized evaluations: Judging language model responses to underspecified queries. *Transactions of the Association for Computational Linguistics*, 13:878–900. 7.2.2

Rui Meng*, Ye Liu*, Shafiq Rayhan Joty, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. 2024. Sfr-embedding-2: Advanced text embedding with multi-stage training. 5.2.2

Meta. 2024. Llama 3.1 model card. 5.2.3

Rada Mihalcea and Andras Csomai. 2007. Wikify! linking documents to encyclopedic knowledge. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, page 233–242, New York, NY, USA. Association for Computing Machinery. 3.2.1

Anne-Lyse Minard, Manuela Speranza, Ruben Urizar, Begoña Altuna, Marieke van Erp, Anneleen Schoen, and Chantal van Son. 2016. MEANTIME, the NewsReader multilingual event and time corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4417–4422, Portorož, Slovenia. European Language Resources Association (ELRA). 2.2, 3.2.3

Mihran Miroyan, Tsung-Han Wu, Logan King, Tianle Li, Jiayi Pan, Xinyan Hu, Wei-Lin Chiang, Anastasios N. Angelopoulos, Trevor Darrell, Narges Norouzi, and Joseph E. Gonzalez. 2025. Search arena: Analyzing search-augmented llms. *Preprint*, arXiv:2506.05334. 7.1, 7.2.1, 7.3.2, 7.7.1

Teruko Mitamura, Zhengzhong Liu, and Eduard H. Hovy. 2017. Events detection, coreference and sequencing: What's next? overview of the tac kbp 2017 event track. *Theory and Applications of Categories*. 1, 2.2

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 145–152, Boston, Massachusetts, USA. Association for Computational Linguistics. 5.3.1

Joel Nothman, James R. Curran, and Tara Murphy. 2008. Transforming Wikipedia into named entity training data. In *Proceedings of the Australasian Language Technology Association Workshop 2008*, pages 124–132, Hobart, Australia. 3.2.1

Joel Nothman, Matthew Honnibal, Ben Hachey, and James R. Curran. 2012. Event linking: Grounding event reference in a news archive. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 228–232, Jeju Island, Korea. Association for Computational Linguistics. 3.1, 3.2.2, 3.2.3

OpenAI. 2022. Introducing chatgpt. 4.1, 4.4

OpenAI. 2025. Gpt-5 system card. 7.1

Jiefu Ou, Adithya Pratapa, Rishubh Gupta, and Teruko Mitamura. 2023. Hierarchical event grounding. In *AAAI*. 1

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155. 4.4

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics. 4.5.1

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2023. Yarn: Efficient context window extension of large language models. *Preprint*, arXiv:2309.00071. 5.3.3

Maxime Peyrard. 2019. Studying summarization evaluation metrics in the appropriate scoring range. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5093–5100, Florence, Italy. Association for Computational Linguistics. 4.3.5

Maxime Peyrard and Robert West. 2020. KLearn: Background knowledge inference from summarization data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2073–2085, Online. Association for Computational Linguistics. 4.2

Horst Pöttker. 2003. News and its communicative quality: the inverted pyramid—when and why did it appear? *Journalism Studies*, 4(4):501–511. 4.3

Adithya Pratapa, Rishubh Gupta, and Teruko Mitamura. 2022. Multilingual event linking to

Wikidata. In *Proceedings of the Workshop on Multilingual Information Access (MIA)*, pages 37–58, Seattle, USA. Association for Computational Linguistics. 1, 3

Adithya Pratapa, Zhengzhong Liu, Kimihiro Hasegawa, Linwei Li, Yukari Yamakawa, Shikun Zhang, and Teruko Mitamura. 2021. Cross-document event identity via dense annotation. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 496–517, Online. Association for Computational Linguistics. 1, 2, 3.1, 3.2.3

Adithya Pratapa and Teruko Mitamura. 2025a. Estimating optimal context length for hybrid retrieval-augmented multi-document summarization. In *Second Conference on Language Modeling*. 1, 6

Adithya Pratapa and Teruko Mitamura. 2025b. Scaling multi-document event summarization: Evaluating compression vs. full-text approaches. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 514–528, Albuquerque, New Mexico. Association for Computational Linguistics. 1, 5, 6.2, 6.3.1, 6.6

Adithya Pratapa, Kevin Small, and Markus Dreyer. 2023. Background summarization of event timelines. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8111–8136, Singapore. Association for Computational Linguistics. 1, 4, 5.1, 5.2.1, 5.3.5, 7.1, 7.2.1

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115. 6.1, 6.2.1, 6.3.3

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. *Journal of Machine Learning Research*, 21(140):1–67. 4.4

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 3505–3506, New York, NY, USA. Association for Computing Machinery. 4.4

Mathieu Ravaut, Aixin Sun, Nancy Chen, and Shafiq Joty. 2024. On context utilization in summarization with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2764–2781, Bangkok, Thailand. Association for Computational Linguistics. 5.2.2, 5.3.1, 5.3.6

Marta Recasens, E. Hovy, and M. A. Martí. 2011. Identity, non-identity, and near-identity: Addressing the complexity of coreference. *Lingua*, 121:1138–1152. 1, 2.1, 2.2, 2.5, 19

Marta Recasens, M. Antònia Martí, and Constantin Orasan. 2012. Annotating near-identity from coreference disagreements. In *Proceedings of the Eighth International Conference on*

Language Resources and Evaluation (LREC'12), pages 165–172, Istanbul, Turkey. European Language Resources Association (ELRA). 2.1, 2.2, 2.4.1

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at trec-3. In *TREC*. 3.4.1

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. 3.2.4

Samaya-AI. 2025. Criteria-eval: Evaluating long-form answers to complex questions. 7.2

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D Manning. 2024. RAPTOR: Recursive abstractive processing for tree-organized retrieval. In *The Twelfth International Conference on Learning Representations*. 5.3.6

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. 4.3.5, 4.5.1, 7.1

Royal Sequiera, Luchen Tan, and Jimmy Lin. 2018. Overview of the trec 2018 real-time summarization track. In *TREC*. 4.1, 4.2

Ori Shapira, David Gabay, Yang Gao, Hadar Ronen, Ramakanth Pasunuru, Mohit Bansal, Yael Amsterdamer, and Ido Dagan. 2019. Crowdsourcing lightweight pyramids for manual summary evaluation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 682–687, Minneapolis, Minnesota. Association for Computational Linguistics. 5.3.1

Avirup Sil, Ernest Cronin, Penghai Nie, Yinfei Yang, Ana-Maria Popescu, and Alexander Yates. 2012. Linking named entities to any database. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 116–127, Jeju Island, Korea. Association for Computational Linguistics. 3.2.1

Sameer Singh, Amarnag Subramanya, Fernando Pereira, and Andrew McCallum. 2012. Wikilinks: A large-scale cross-document coreference corpus labeled via links to wikipedia. *Technical Report*. 3.1, 3.2.1

Zhiyi Song, Ann Bies, Justin Mott, Xuansong Li, Stephanie Strassel, and Christopher Caruso. 2018. Cross-document, cross-language event coreference annotation using event hoppers. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA). 2.2, 2.4.1

Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan. 2018. Supervised open information extraction. In *Proceedings of the 2018 Conference of the North American Chapter* *of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 885–895, New Orleans, Louisiana. Association for Computational Linguistics. 2.3

Jannik Strötgen and Michael Gertz. 2015. A baseline temporal tagger for all languages. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 541–547, Lisbon, Portugal. Association for Computational Linguistics. 3.3.1

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. 2021. Roformer: Enhanced transformer with rotary position embedding. *Preprint*, arXiv:2104.09864. 5.3.3

Mirac Suzgun, Luke Melas-Kyriazi, and Dan Jurafsky. 2023. Follow the wisdom of the crowd: Effective text generation via minimum Bayes risk decoding. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics. 6.2.2

Jamba Team, Barak Lenz, Alan Arazi, Amir Bergman, Avshalom Manevich, Barak Peleg, Ben Aviram, Chen Almagor, Clara Fridman, Dan Padnos, Daniel Gissin, Daniel Jannai, Dor Muhlgay, Dor Zimberg, Edden M Gerber, Elad Dolev, Eran Krakovsky, Erez Safahi, Erez Schwartz, Gal Cohen, Gal Shachaf, Haim Rozenblum, Hofit Bata, Ido Blass, Inbal Magar, Itay Dalmedigos, Jhonathan Osin, Julie Fadlon, Maria Rozman, Matan Danos, Michael Gokhman, Mor Zusman, Naama Gidron, Nir Ratner, Noam Gat, Noam Rozen, Oded Fried, Ohad Leshno, Omer Antverg, Omri Abend, Opher Lieber, Or Dagan, Orit Cohavi, Raz Alon, Ro'i Belson, Roi Cohen, Rom Gilad, Roman Glozman, Shahar Lev, Shaked Meirom, Tal Delbari, Tal Ness, Tomer Asida, Tom Ben Gal, Tom Braude, Uriya Pumerantz, Yehoshua Cohen, Yonatan Belinkov, Yuval Globerson, Yuval Peleg Levy, and Yoav Shoham. 2024a. Jamba-1.5: Hybrid transformer-mamba models at scale. *Preprint*, arXiv:2408.12570. 5.2.3

Jamba Team, Barak Lenz, Alan Arazi, Amir Bergman, Avshalom Manevich, Barak Peleg, Ben Aviram, Chen Almagor, Clara Fridman, Dan Padnos, Daniel Gissin, Daniel Jannai, Dor Muhlgay, Dor Zimberg, Edden M Gerber, Elad Dolev, Eran Krakovsky, Erez Safahi, Erez Schwartz, Gal Cohen, Gal Shachaf, Haim Rozenblum, Hofit Bata, Ido Blass, Inbal Magar, Itay Dalmedigos, Jhonathan Osin, Julie Fadlon, Maria Rozman, Matan Danos, Michael Gokhman, Mor Zusman, Naama Gidron, Nir Ratner, Noam Gat, Noam Rozen, Oded Fried, Ohad Leshno, Omer Antverg, Omri Abend, Opher Lieber, Or Dagan, Orit Cohavi, Raz Alon, Ro'i Belson, Roi Cohen, Rom Gilad, Roman Glozman, Shahar Lev, Shaked Meirom, Tal Delbari, Tal Ness, Tomer Asida, Tom Ben Gal, Tom Braude, Uriya Pumerantz, Yehoshua Cohen, Yonatan Belinkov, Yuval Globerson, Yuval Peleg Levy, and Yoav Shoham. 2024b. Jamba-1.5: Hybrid transformer-mamba models at scale. *Preprint*, arXiv:2408.12570. 6.1, 6.2.1

Giang Tran, Mohammad Alrifai, and Eelco Herder. 2015. Timeline summarization from relevant headlines. In *Advances in Information Retrieval*, pages 245–256, Cham. Springer International Publishing. 4.1, 4.2, 4.3.2, 5.6.1

Shyam Upadhyay, Nitish Gupta, Christos Christodoulopoulos, and Dan Roth. 2016. Revisiting the evaluation for cross document event coreference. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1949–1958, Osaka, Japan. The COLING 2016 Organizing Committee. 2.6

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. 2024. Replacing judges with juries: Evaluating llm generations with a panel of diverse models. *Preprint*, arXiv:2404.18796. 6.2.1

Jesse Vig, Alexander Fabbri, Wojciech Kryscinski, Chien-Sheng Wu, and Wenhao Liu. 2022. Exploring neural models for query-focused summarization. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1455–1468, Seattle, United States. Association for Computational Linguistics. 5.2.3

Piek Vossen, Filip Ilievski, Marten Postma, and Roxane Segers. 2018. Don't annotate, but validate: a data-to-text method for capturing event data. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA). 2.2

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. ACE 2005 Multilingual Training Corpus. *LDC2006T06, Philadelphia, Penn.: Linguistic Data Consortium*. 2.2

Alex Wang, Richard Yuanzhe Pang, Angelica Chen, Jason Phang, and Samuel R. Bowman. 2022. SQuALITY: Building a long-document summarization dataset the hard way. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1139–1156, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. 7.1, 7.3.3

Lu Wang, Claire Cardie, and Galen Marchetti. 2015. Socially-informed timeline generation for complex events. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1055–1065, Denver, Colorado. Association for Computational Linguistics. 4.1, 4.2, 4.3.2, 5.6.1

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. 2025. Browsecomp: A simple yet challenging benchmark for browsing agents. *Preprint*, arXiv:2504.12516. 7.3.2, 8

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. 2013. OntoNotes Release 5.0. *LDC2013T19, Philadelphia, Penn.: Linguistic Data Consortium*. 2.2, 3.3.1

David Wiggins. 1967. *Identity and Spatio-temporal Continuity*. Library of philosophy and logic. Blackwell. 19

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics. 2.6, 3.5

Kristin Wright-Bettner, Martha Palmer, Guergana Savova, Piet de Groen, and Timothy Miller.

2019. Cross-document coreference: An approach to capturing coreference without context. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 1–10, Hong Kong. Association for Computational Linguistics. 2.2, 2.4.1, 2.5

Jeff Wu, Long Ouyang, Daniel M. Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. 2021. Recursively summarizing books with human feedback. *Preprint*, arXiv:2109.10862. 4.2

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable zero-shot entity linking with dense entity retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6397–6407, Online. Association for Computational Linguistics. 3.1, 3.4, 3.4.2, 3.4.2, 3.5.3

Fangyuan Xu, Weijia Shi, and Eunsol Choi. 2024. RECOMP: Improving retrieval-augmented LMs with context compression and selective augmentation. In *The Twelfth International Conference on Learning Representations*. 5.3.6, 6.6

Yumo Xu and Mirella Lapata. 2022. Document summarization with latent queries. *Transactions of the Association for Computational Linguistics*, 10:623–638. 4.1

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. *Preprint*, arXiv:2501.15383. 6.2.1, 6.3.3, 6.5.1

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*. 6.1, 6.1, 6.3.4

Tan Yu, Anbang Xu, and Rama Akkiraju. 2024a. In defense of rag in the era of long-context language models. *Preprint*, arXiv:2409.01666. 5.2.2

Tan Yu, Anbang Xu, and Rama Akkiraju. 2024b. In defense of rag in the era of long-context language models. *Preprint*, arXiv:2409.01666. 6.1, 6.2, 6.5.1

Xiaodong Yu, Wenpeng Yin, Nitish Gupta, and Dan Roth. 2023. Event linking: Grounding event mentions to Wikipedia. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2679–2688, Dubrovnik, Croatia. Association for Computational Linguistics. 3.2.2

Zhenrui Yue, Honglei Zhuang, Aijun Bai, Kai Hui, Rolf Jagerman, Hansi Zeng, Zhen Qin, Dong Wang, Xuanhui Wang, and Michael Bendersky. 2025. Inference scaling for long-context retrieval augmented generation. In *The Thirteenth International Conference on Learning Representations*. 6.1, 6.2.3

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*. 4.5.1

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori Hashimoto. 2023. Benchmarking large language models for news summarization. *Preprint*, arXiv:2301.13848. 4.1, 4.3.3, 4.3.5, 4.6, 7.1

Yusen Zhang, Ansong Ni, Ziming Mao, Chen Henry Wu, Chenguang Zhu, Budhaditya Deb, Ahmed Awadallah, Dragomir Radev, and Rui Zhang. 2022. Summ^n: A multi-stage summarization framework for long input dialogues and documents. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1592–1604, Dublin, Ireland. Association for Computational Linguistics. 4.2

Yilun Zhao, Kaiyan Zhang, Tiansheng Hu, Sihong Wu, Ronan Le Bras, Taira Anderson, Jonathan Bragg, Joseph Chee Chang, Jesse Dodge, Matt Latzke, Yixin Liu, Charles McGrady, Xiangru Tang, Zihang Wang, Chen Zhao, Hannaneh Hajishirzi, Doug Downey, and Arman Cohan. 2025. Sciarena: An open evaluation platform for foundation models in scientific literature tasks. *Preprint*, arXiv:2507.01001. 7.1, 7.2.1, 7.3.2

Yijie Zhou, Kejian Shi, Wencai Zhang, Yixin Liu, Yilun Zhao, and Arman Cohan. 2023. Odsum: New benchmarks for open domain multi-document summarization. *Preprint*, arXiv:2309.08960. 6.6

Dawei Zhu, Liang Wang, Nan Yang, Yifan Song, Wenhao Wu, Furu Wei, and Sujian Li. 2024. Longembed: Extending embedding models for long context retrieval. *Preprint*, arXiv:2404.12096. 5.6.5