

NewSumm 2025

**Proceedings of The 5th New Frontiers in Summarization
Workshop**

Proceedings of EMNLP Workshop

November 8, 2025

©2025 Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-337-1

Message from the Workshop Chairs

The development of intelligent systems capable of producing concise, fluent, and accurate summaries remains a longstanding objective in natural language processing. Despite the remarkable progress brought by large language models, summarization continues to play a critical role in distilling, structuring, and verifying information at scale. Challenges such as faithfulness, controllability, efficiency, and domain adaptation remain open, and summarization research provides valuable insights into addressing these broader issues of generation and understanding. This workshop serves as a forum for the exchange of ideas toward these goals. It brings together experts from various disciplines, including summarization, language generation, and cognitive and psycholinguistics, to discuss key issues in automatic summarization. The agenda covers a wide array of topics, such as innovative paradigms and frameworks, multilingual and cross-lingual setups, shared tasks, information integration, novel evaluation methods, applied research, and future research directions. The workshop is aimed at fostering a cohesive research community, expediting the transfer of knowledge, and developing new tools, datasets, and resources to meet the needs of academia, industry, and government.

This fifth edition of the workshop, following our previous workshops at EMNLP 2017, EMNLP 2019, EMNLP 2021, and EMNLP 2023, received a total of 33 paper submissions with an overall acceptance rate of 42.4

NewSumm Organizing Committee

Yue Dong (University of California, Riverside, USA)

Wen Xiao (Microsoft Azure AI, Canada)

Haopeng Zhang (University of Hawaii at Manoa, USA)

Rui Zhang (Penn State University, USA)

Ori Ernst (McGill University and Mila, Canada)

Lu Wang (University of Michigan, USA)

Fei Liu (Emory University, USA)

Organizing Committee

General Chair, Publication Chair

Yue Dong, University of California, Riverside, USA

Vice General Chair, Presentation Chair

Wen Xiao, Microsoft Azure AI, Canada

Program Chair, Publicity Chair

Haopeng Zhang, University of Hawaii at Manoa, USA

Program Chair, Invited Speakers Chair

Rui Zhang, Penn State University, USA

Program Chair, Technical and Web Chair

Ori Ernst, McGill University and Mila, Canada

Advisory Chairs

Wang Lu, University of Michigan, USA

Fei Liu, Emory University, USA

Program Committee

Program Chairs

Yue Dong, University of California, Riverside and McGill University
Ori Ernst, Mila - Quebec Artificial Intelligence Institute
Wen Xiao, Microsoft
Rui Zhang, Pennsylvania State University
Haopeng Zhang, University of Hawaii System

Reviewers

Shmuel Amar

Florian Boudin

Avi Caciularu, Arie Cattan, Hou Pong Chan, Khaoula Chehbouni, Ziling Cheng, Jackie CK Cheung

Maxime Darrin, Felice Dell'Orletta

Ron Eliav

Tobias Falke, Lorenzo Jaime Yu Flores, Yu Fu

Eran Hirsch, Zhe Hu, Xinyu Hua, Patrick Huber

Hayate Iso

Ayal Klein, Wojciech Maciej Kryscinski, Eugene Kwek

Elena Lloret

Margot Mieskes

Manabu Okumura, Jessica Ouyang

G M Shahariar, Haz Sameen Shahgir, Ori Shapira, Aviv Slobodkin

Esau Villatoro-tello

David Wan

Haohan Yuan

Nan Zhang, Shiyue Zhang, Yusen Zhang, Ming Zhong, Xiyuan Zou

Keynote Talk
Invited Speaker 1

Mohit Bansal
UNC Chapel Hill

Keynote Talk
Invited Speaker 2

Arman Cohan
Yale University

Keynote Talk
Invited Speaker 3

Greg Durrett
New York University

Keynote Talk
Invited Speaker 4

Alexander R. Fabbri
Meta

Keynote Talk
Invited Speaker 5

Mirella Lapata
University of Edinburgh

Keynote Talk
Invited Speaker 6

Jey Han Lau
University of Melbourne

Table of Contents

<i>LLM-as-a-Judge Failures at Automating the Identification of Poor Quality Outputs in Free-Form Texts</i> Zongxia Li, Xiyang Wu, Ishani Mondal, Alexa Siu, Jordan Lee Boyd-Graber and Ani Nenkova .	1
<i>Hierarchical Attention Adapter for Abstractive Dialogue Summarization</i> Raymond Li, Chuyuan Li, Gabriel Murray and Giuseppe Carenini	17
<i>CS-Sum: A Benchmark for Code-Switching Dialogue Summarization and the Limits of Large Language Models</i> Sathya Krishnan Suresh, Tanmay Surana, Lim Zhi Hao and Eng Siong Chng	31
<i>Beyond Paraphrasing: Analyzing Summarization Abstractiveness and Reasoning</i> Nathan Zeweniuk, Ori Ernst and Jackie CK Cheung	48
<i>Improving Aspect-Based Summarization via Contrastive Learning with Anchored Negative Examples</i> Elizabeth Palmieri and Yangfeng Ji	59
<i>REFER: Mitigating Bias in Opinion Summarisation via Frequency Framed Prompting</i> Nannan Huang, Haytham M. Fayek and Xiuzhen Zhang	74
<i>DACP: Domain-Adaptive Continual Pre-Training of Large Language Models for Phone Conversation Summarization</i> Xue-Yong Fu, Elena Khasanova, Md Tahmid Rahman Laskar, Harsh Saini and Shashi Bhushan TN	94
<i>From Keyterms to Context: Exploring Topic Description Generation in Scientific Corpora</i> Pierre Achkar, Satiyabooshan Murugaboopathy, Anne Kreuter, Tim Gollub, Martin Potthast and Yuri Campbell	102
<i>HalluTree: Explainable Multi-Hop Hallucination Detection for Abstractive Summarization</i> Daniel Orshansky, Oskar Oomen, Naaisha Agarwal and Ryan Lagasse	123
<i>Multi2: Multi-Agent Test-Time Scalable Framework for Multi-Document Processing</i> Juntai Cao, Xiang Zhang, Raymond Li, Jiaqi Wei, Chuyuan Li, Shafiq Joty and Giuseppe Carenini	135
<i>Bridging Multimodal and Video Summarization: A Unified Survey</i> Haopeng Zhang	157
<i>AdvSumm: Adversarial Training for Bias Mitigation in Text Summarization</i> Mukur Gupta, Nikhil Reddy Varimalla, Nicholas Deas, Melanie Subbiah and Kathleen McKeown	172
<i>NSF-SciFy: Mining the NSF Awards Database for Scientific Claims</i> Delip Rao, Weiqiu You, Eric Wong and Chris Callison-Burch	183
<i>QA-prompting: Improving Summarization with Large Language Models using Question-Answering</i> Neelabh Sinha	199

Program

Saturday, November 8, 2025

- 08:50 - 09:00 *Opening Remarks*
- 09:00 - 09:45 *Invited Talk - Mohit Bansal*
- 09:45 - 10:30 *Invited Talk - Greg Durrett*
- 10:30 - 11:00 *Coffee Break*
- 11:00 - 11:45 *Invited Talk - Arman Cohan*
- 11:45 - 12:30 *Invited Talk - Alexander R. Fabbri*
- 12:30 - 14:00 *Lunch Break*
- 14:00 - 15:30 *Lightning Talk + Poster Session (Workshop papers)*
- 15:30 - 16:00 *Coffee Break*
- 16:00 - 16:45 *Invited Talk - Jey Han Lau*
- 16:45 - 17:30 *Invited Talk - Mirella Lapata*
- 17:30 - 17:35 *Final Remarks*

NSF-SciFY: Mining the NSF Awards Database for Scientific Claims

Delip Rao^{*†}, Weiqiu You[†], Eric Wong, Chris Callison-Burch

University of Pennsylvania

Philadelphia, PA, USA

{delip, weiqiuy, exwong, ccb}@seas.upenn.edu

Abstract

We introduce NSF-SciFY, a comprehensive dataset of scientific claims and investigation proposals extracted from National Science Foundation award abstracts. While previous scientific claim verification datasets have been limited in size and scope, NSF-SciFY represents a significant advance with 2.8 million claims from 400,000 abstracts spanning all science and mathematics disciplines. We present two focused subsets: NSF-SciFY-MATSCI with 114,000 claims from materials science awards, and NSF-SciFY-20K with 135,000 claims across five NSF directorates. Using zero-shot prompting, we develop a scalable approach for joint extraction of scientific claims and investigation proposals. We demonstrate the dataset’s utility through three downstream tasks: non-technical abstract generation, claim extraction, and investigation proposal extraction. Fine-tuning language models on our dataset yields substantial improvements, with relative gains often exceeding 100%, particularly for claim and proposal extraction tasks. Our error analysis reveals that extracted claims exhibit high precision but lower recall, suggesting opportunities for further methodological refinement. NSF-SciFY enables new research directions in large-scale claim verification, scientific discovery tracking, and meta-scientific analysis.

1 Introduction

The overall growth rate of scientific publications is estimated to be 4% annually, with a doubling time of 17 years (Bornmann et al., 2021). Within this deluge, researchers, reviewers, and the general public struggle to separate substantiated claims from spurious ones—whether it is the “quantum supremacy” assertions in computing, the short-lived excitement over LK-99 superconductors³,

^{*}Corresponding author, [†]co-first author

³for an entertaining digression c.f., <https://en.wikipedia.org/wiki/LK-99>

```
{
  "award_id": 2321365,
  "title": "Electrically Conductive 2D Metal-Organic Frameworks and Cov...",
  "technical_abstract": "Owing to their diverse potentials to serve as e...",
  "non_technical_abstract": "Sustaining the rapid advances of modern ele...",
  "verifiable_claims": [
    "MOFs and COFs have synthetic accessibility, structural modularity,",
    "Electrical conductivity remains one of the most elusive traits of M...",
    "In 2D MOFs, electronic conduction can occur within the planes thro...",
    "In 2D COFs, pi-stacked layers represent the primary transport path...",
    ...
  ],
  "investigation_proposals": [
    "Develop and implement a new design strategy to promote long-range c...",
    "Incorporate built-in alternating pi-donor/acceptor stacks inside c...",
    "Investigate how pi-donor/acceptor stacks consisting of different c...",
    "Create a new design strategy for next-generation electrically cond...",
    "Produce novel electrically conductive 2D MOFs and COFs with unique...",
    ...
  ],
  "publications": [
    {
      "doi": "10.1021/acs.inorgchem.3c02647",
      "status": "resolved",
      "title": "From a Collapse-Prone, Insulating Ni-MOF-74 Analogue to...",
      "abstract": "Electrically conductive porous metal-organic framewo..."
    }
    ...
  ]
}
```

Figure 1: A sample record from our dataset. Each record contains 1) Award ID and title, 2) NSF Directorate, 3) Technical and non-technical abstracts, 4) Scientific Claims, 5) Investigation Proposals, and 6) Associated publications, when present.

or the misunderstanding surrounding microplastic leaches from black plastic spatulas⁴. Manual verification of ever growing body of scientific claims has become intractable, yet the economic and societal consequences of unverified claims are increasingly severe.

Wadden et al. (2020) introduced the task of scientific claim verification with the SciFACT dataset, focusing primarily on automatic verification of scientific claims. Follow up works (see Section 2 for a detailed account) have mostly focused on the healthcare, building datasets from scientific publications, and modest-sized dataset creation. In this work, we relax all of these aspects and look at building at least an order of magnitude large-scale

⁴c.f., <https://nationalpost.com/news/canada/black-plastic>

Dataset	# claims	# docs	Evidence Source	Domain
SciFACT (Wadden et al., 2020)	1.4K	5K	Research papers	Biomedical
PubHEALTH (Kotonya and Toni, 2020)	11.8K	11.8K	Fact-checking sites	Public health
CLIMATE-FEVER (Diggelmann et al., 2020)	1.5K	7.5K	Wikipedia articles	Climate change
HealthVer (Sarrouiti et al., 2021)	1.8K	738	Research papers	Healthcare
COVID-Fact (Saakyan et al., 2021)	4K	4K	Research, news	COVID
CoVERT (Mohr et al., 2022)	300	300	Research, news	Biomedical
SciFACT-Open (Wadden et al., 2022)	279	500K	Research papers	Biomedical
NSF-SciFY-MATSCI (ours)	114K	16K	NSF award abstracts	Material Science
NSF-SciFY-20K (ours)	135K	20K	NSF award abstracts	All Science & Math
NSF-SciFY (ours)	2.8M	400K	NSF award abstracts	All Science & Math

Table 1: This table comparison clearly illustrates the scale advantage of NSF-SciFy over existing scientific claim verification datasets. While previous datasets like SciFACT and PubHEALTH contain at most thousands of claims from published research papers or fact-checking sources, our NSF-SciFY-MATSCI and NSF-SciFY-20K datasets individually contribute more than 100K claims. The full NSF-SciFY dataset represents an order-of-magnitude increase with 2.8M claims across 400K abstracts spanning all science & math disciplines. This work introduces grant abstracts as a novel, untapped source for scientific claim extraction, complementing existing approaches that focus on published literature, news articles, or social media content.

scientific claim dataset covering all of basic science. We envision building of such large-scale, scientific claim datasets to help future work on robust scientific claim verification systems.

We introduce NSF-SciFY¹, a comprehensive dataset of claims and investigation proposals extracted from National Science Foundation (NSF) award abstracts. We choose NSF abstracts as our source material for several reasons:

1. NSF is a primary driver of U.S. scientific innovation, funding approximately 25% of all federally supported basic research, spanning the entirety of science and math areas, with an annual budget of \$9.9 billion (FY 2023). Any claim dataset derived from the NSF awards database should faithfully represent the scientific Zeitgeist.
2. NSF’s rigorous subject matter expert-review process provides an high-quality filter for the claims made in funded proposals.
3. The public availability and permissive usage terms of the NSF awards database makes it an excellent resource for open science research.
4. Previous datasets on scientific claims have been derived from scientific papers, but claims in scientific grants, and particularly investigation proposals, remain unstudied.

While not this focus of this paper, grant award abstracts, additionally, provide a unique opportunity to study the relationship between what researchers claim and what they propose to investigate. This could offer valuable insights into scientific practice

¹Short for “NSF SCientific FeasibilitY”.

and the evolution of research questions.

In this paper, we make the following contributions: (1) We introduce NSF-SciFY, the largest scientific claim dataset to date with 2.8M claims extracted from 400K NSF award abstracts, establishing grant proposals as a novel source for scientific claim extraction; (2) We create NSF-SciFY-MATSCI focusing exclusively on materials science with 114K extracted claims from 16K abstracts. This is the first materials science claim dataset and, in number of extracted claims, this alone is an order of magnitude bigger than the largest publicly available claim dataset; In addition, we also create NSF-SciFY-20K with 135K claims spanning five NSF directorates. (3) We develop a zero-shot prompting approach for joint extraction of scientific claims and investigation proposals as a scalable way to bootstrap high-precision, large-scale scientific claim datasets; (4) We present novel evaluation metrics for claim/proposal extraction based on LLM judgments, showing that fine-tuned models significantly outperform base models; and (5) Finally, we release all datasets and trained models from our work for unfettered research and commercial use. Our dataset and methods enable new opportunities for large-scale claim verification, scientific discovery tracking, and meta-scientific research.

2 Related Work

Scientific claim extraction and verification has emerged as an important research area as the volume of scientific literature continues to grow expo-

nentially. Previous work has primarily focused on claims from published papers, fact-checking sites, and news articles.

Scientific Claim Datasets Several datasets have been developed for scientific claim verification, but all have focused on claims from published literature, while we undertake the study of grant award abstracts. SciFACT (Wadden et al., 2020) contains 1,400 scientific claims derived from research papers in the biomedical domain. PubHEALTH (Kotonya and Toni, 2020) includes 11,800 claims from journalists and fact-checkers in public health. CLIMATE-FEVER (Diggelmann et al., 2020) compiled 1,500 claims from news articles about climate change. HealthVer (Sarrouti et al., 2021) extracted 1,800 claims from search queries related to health topics. COVID-Fact (Saakyan et al., 2021) and CoVERT (Mohr et al., 2022) focused on COVID-19 related claims from social media. SciFact-Open (Wadden et al., 2022) expanded the original SciFact dataset using information retrieval pooling, yet it still remains health-care focused and a few orders of magnitude smaller than our largest dataset.

Table 1 situates existing scientific claim datasets with our NSF-SCIFY datasets, highlighting the significantly larger scale of our contribution (2.8 million claims in NSF-SCIFY, 135,000 claims in NSF-SCIFY-20K and 114,000 claims in NSF-SCIFY-MATSCI), broad topic coverage (all of science and math), and novelty of data source (grant abstracts). See Figure 2.

Meta Science and Social Science Previous works have examined grants data in social science and meta-science contexts. For example, Park et al. (2024) examine the relationship between interdisciplinary grants and the impact of papers they support and Xu et al. (2022) study the influence of research funding on team structure using grant data. While these are tenuously connected to our work, we list them for the sake of completeness.

3 Building NSF-SCIFY

3.1 Data Collection

We downloaded the entire NSF Awards database² in XML format, containing more than 0.5 million awards from 1970 through September 2024. After parsing, we obtained 412,155 parseable awards, which we call NSF-SCIFY.

²<https://www.nsf.gov/awardsearch/advancedSearch.jsp>

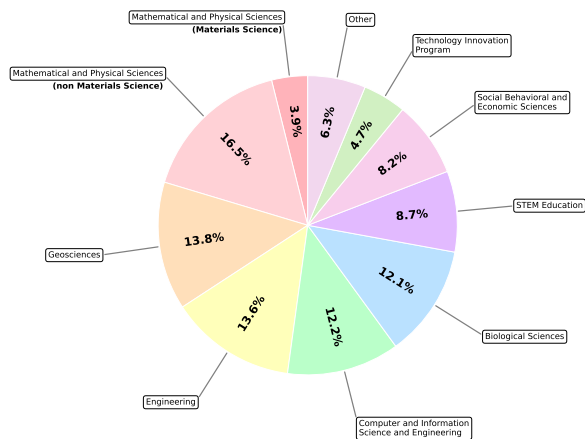


Figure 2: Distribution of awards areas as represented by the National Science Foundation directorates in NSF-SCIFY, illustrating the breadth and comprehensiveness of scientific claims in our dataset. The NSF-SCIFY-MATSCI subset spanning all of materials science awards represents 3.9% of the entire dataset.

In this paper, we focus on all awards from the Division of Materials Research (DMR), which is responsible for most materials science awards at the NSF. This subset, called NSF-SCIFY-MATSCI, contains 16,031 awards, representing approximately 3.2% of the entire NSF awards database. We chose materials science as our focus due to its interdisciplinary nature and technological importance. In addition, we build NSF-SCIFY-20K, a different subset of 20K awards spanning 5 NSF directorates — Mathematical and Physical Sciences (MPS), Geological Sciences (GEO), Engineering (ENG), Computer and Information Science and Engineering (CSE), and Biological Sciences (BIO).

3.2 Data Processing

As Figure 1 illustrates, each record in NSF-SCIFY-MATSCI typically contains:

1. Award ID, title, and year.
2. Directorate and division information
3. Technical abstract
4. Non-technical abstract (present in $\sim 81\%$ of awards)
5. Scientific claims made in the abstracts
6. Investigation proposals in the abstracts
7. Publications resulting from the grant (when available)

The practice of updating awards with resulting publications is relatively recent, primarily occurring from 2014 onwards. For awards where publications are present, we extracted the DOIs and

resolved them to obtain titles, abstracts, and publication URLs.

3.3 Claim and Investigation Proposal Extraction

To extract scientific claims and investigation proposals from the award abstracts, we developed a zero-shot prompting approach using Anthropic’s Claude-3.5³ model. Our prompt instructed the model to identify two types of statements:

1. **Verifiable claims:** Statements that the abstract claims to be true or states as assumptions, either explicitly or implicitly.
2. **Investigation proposals:** Forward-looking statements that propose specific research activities as part of the award.

We structured the prompt to return a JSON object containing the award ID, technical abstract, non-technical abstract, a list of verifiable claims, and a list of investigation proposals. To maintain consistency and quality, we set temperature to zero for all extractions. See Appendix A for the exact prompt and Appendix G for sample claims and investigation proposals.

We performed qualitative experiments with several prompt variants and our analysis showed that jointly extracting claims and investigation proposals helped maintain the relevance of extracted claims. When claims were extracted without also extracting investigation proposals, the model often confused forward-looking statements about proposed investigations as factual claims.

4 Dataset Analysis

NSF-SciFY The full dataset contains 412,155 award abstracts spanning from 1970 to 2024, with 2.8 million scientific claims and corresponding investigation proposals.

NSF-SciFY-MATSCI This materials science subset, which is the focus of this preprint, contains:

- 16,042 awards with each with a technical and non-technical abstract
- 114K extracted scientific claims (average of 7 ± 2 claims per abstract-pair)
- 145K extracted investigation proposals (average of 9 ± 3 proposals per abstract-pair)
- 2,953 awards with linked publications (18.4% of the dataset). Such awards had anywhere between 1 – 4 publications.

³Claude-3.5-Sonnet-20240620 accessed between Sep-Oct. 2024, to be specific

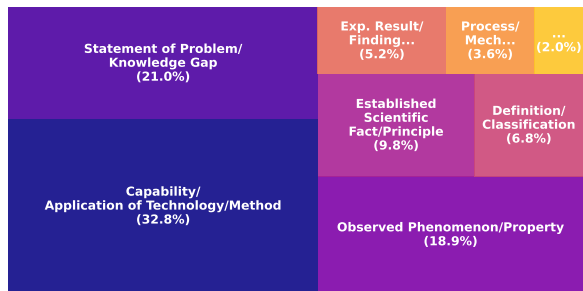


Figure 3: A treemap of the scientific claim categories in NSF awards. See Section 4.2 for descriptions of these categories.

NSF-SciFY-20K For building models across all NSF directorates, we take 20,000 sample subset of NSF-SciFY, by stratifying across 5 directorates.

- 20,001 awards with each with a technical and non-technical abstract
- 135K extracted scientific claims (average of 7 ± 2 claims per abstract-pair)
- 139K extracted investigation proposals (average of 7 ± 2 proposals per abstract-pair)

4.1 Technical vs. Non-Technical Abstracts

We investigated the differences between technical and non-technical abstracts in our dataset. Using a symmetric BLEU score to measure textual similarity between paired abstracts, we found that only 202 (1.5%) out of 13,025 technical/non-technical abstract pairs had a similarity score greater than 0.6, suggesting that the non-technical abstracts are not simply copied from the technical abstracts.

Since grant abstracts are previously unexamined in literature, we further investigated the stylistic differences between technical and non-technical abstracts using pre-trained document embedding models. Figure 5 compares content embeddings from SPECTER (Cohan et al., 2020) and style embeddings from STEL (Patel et al., 2025). Using these embeddings with a linear SVM classifier, we achieved F1 scores of 90.99 (SPECTER), 88.42 (STEL), and 89.99 (concatenated), demonstrating that the abstracts are distinguishable both in content and style.

4.2 A Taxonomy of Scientific Claims in NSF Award Abstracts

To understand the nature of assertions made in NSF award abstracts, we analyzed and grouped 810 extracted claims over a stratified sample of 120 awards spanning 5 NSF directorates – MPS, GEO, ENG, CSE, and BIO – resulting in the identifica-

tion of eight distinct categories. These categories encompass a range of scientific statements: **Established Scientific Fact/Principle** refers to claims that state well-accepted scientific laws or widely known facts. **Observed Phenomenon/Property** includes claims describing natural phenomena or material properties that have been observed. Claims detailing the function, potential, or application of a specific technology, method, or material fall under **Capability/Application of Technology/Method**. **Hypothesis/Theoretical Prediction** categorizes claims that propose theoretical explanations, predictions, or hypotheses requiring further verification. Specific results, findings, or statements about measurability are grouped under **Experimental Result/Finding/Measurability**. **Statement of Problem/Knowledge Gap** highlights existing problems, limitations, or gaps in current knowledge or technology. Claims that define terms or classify entities are categorized as **Definition/Classification**. Finally, **Process/Mechanism Description** includes claims that describe physical or chemical processes or mechanisms. Figure 3 shows the distribution of these claim types, and Table 10 lists examples for each of these types. As one might expect in grant applications, the most common claim type is “Capability/Application of Technology/Method” (32.8%), followed by “Statement of Problem/Knowledge Gap” (21.0%) and “Observed Phenomenon/Property” (18.9%).

4.3 Evaluating the Extracted Verifiable Claims

To evaluate how good the claim extraction from Section 3.3 is, we annotated the abstracts in the 120 sampled awards (see Section 4.2) with claims and computed precision and recall. Figure 4 summarizes the precision, recall, and F1-scores of the extracted claims for all 6 areas of interest. The areas of interest include, Materials Science (DMR) Mathematical and Physical Sciences except Materials Science (MPS-DMR), Geological Sciences (GEO), Engineering (ENG), Computer and Information Science and Engineering (CSE), and Biological Sciences (BIO). We notice from Figure 4 a near perfect precision for the extracted claims in all areas. Although our LLM-derived dataset suffers from recall, we hope that training on a massive corpus of high quality claims will result in useful claim extraction systems. We consider our findings on claim extraction as directional and skip a similar evaluation for investigation proposals.

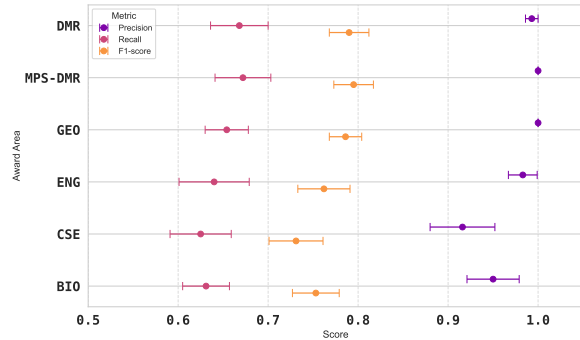


Figure 4: A Cleveland dot plot of precision, recall, and f1-score across different NSF Award Areas for claims extracted via Claude (See Section 3.3). Error bars denote standard deviation (bootstrap N=1000). See Section 4.3 for analysis.

5 Tasks, Metrics, and Experiments

We designed three NLP tasks to evaluate the utility of our dataset:

1. The **Non-technical Abstract Generation** task translates dense, technical grant abstracts into accessible language for broader science communication. Motivated by capturing the core scientific essence while navigating stylistic and content differences between technical and lay summaries, this task uses the dataset’s paired examples (common in NSF awards) to train models for this nuanced transformation.
2. The **Abstract to Scientific Claims Extraction** task automates identifying verifiable assertions—the core of scientific discourse—from grant abstracts, which capture these claims at an early, pre-publication stage. Significant performance gains post-fine-tuning highlight the dataset’s effectiveness in teaching models to pinpoint these crucial statements.
3. The **Abstract to Investigation Proposals Extraction** task distinguishes aspirational research intentions from established claims, offering a novel analysis of scientific texts. This provides a clearer view of the planned research trajectory by identifying intended activities. It complements claim extraction by presenting a fuller picture of proposed work, from assertions to investigative pathways, again showing significant fine-tuning efficacy due to the dataset’s focused nature.

To explore the three tasks, we finetuned two 7B parameter language models:

- Mistral-7B-instruct-v0.3 (Jiang et al., 2023)

Metric	Mistral	Qwen
BERTScore-P	0.8563 (+0.38% ↑)	0.8459 (+0.98% ↑)
BERTScore-R	0.8555 (+0.30% ↑)	0.8597 (+1.61% ↑)
BERTScore-F1	0.8561 (+0.36% ↑)	0.8437 (+0.75% ↑)
ROUGE1	0.2000 (+2.58% ↑)	0.1978 (+1.98% ↑)
ROUGE2	0.0198 (+4.76% ↑)	0.0210 (+3.89% ↑)
ROUGE-L	0.1273 (+2.96% ↑)	0.1466 (+0.65% ↑)
ROUGE-L-sum	0.2166 (+2.45% ↑)	0.2078 (+1.66% ↑)

Table 2: Finetuning performance for Mistral-7B-instruct-v0.3 and Qwen2.5-7B-Instruct models for Technical abstract to Non-technical abstract translation (Task 1), with relative improvements over the corresponding unfinetuned model indicated in green. Error bars for all metrics at 95% confidence intervals range between 0.0000–0.0025. Mistral model outperforms Qwen on almost all metrics for this task regardless of finetuning. We note that the percent changes, while statistically significant, are modest for this task indicating an excellent out-of-the-box performance that’s challenging to improve upon.

- Qwen2.5-7B-Instruct (Yang et al., 2024)

Details about the finetuning can be found in Appendix E.

5.1 Evaluation Metrics

For Task 1 – abstract generation – we employed a comprehensive evaluation framework using both BERTScore (Zhang* et al., 2020) and ROUGE (Lin, 2004) metrics to assess the quality of generated non-technical abstracts. This combination enables us to capture both lexical overlap and structural similarity through the ROUGE variants, while BERTScore provides insights into semantic alignment between the generated texts and reference abstracts. Incorporating such multi-viewed metrics⁴ ensures that the evaluation reflects not only the presence of key words and phrases but also the underlying meaning and narrative coherence of the abstracts.

For Task 2 – claim extraction – we developed a novel evaluation approach using LLM-based comparisons. Previous methods for claim evaluations focused on comparing a single claim against a single document. See Tang et al. (2024), for example. However, our setting required evaluating a set of extracted claims against a gold set of claims.

Towards that end, we defined a boolean function Φ_{claim} using GPT-4o-mini with zero-shot prompt-

⁴For BERTScore we report precision, recall and F1, and for ROUGE we report ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-L-sum.

ing to determine whether a generated claim is supported by a gold standard claim. See Appendix B for prompt details⁵. Using this function, we calculated precision and recall as follows:

$$\text{Precision} = \frac{1}{|S|} \sum_{c \in S} \max_{g \in G} \Phi_{\text{claim}}(c, g)$$

$$\text{Recall} = \frac{1}{|G|} \sum_{g \in G} \max_{c \in S} \Phi_{\text{claim}}(g, c)$$

where S is the set of claims generated from the finetuned model, after removal of any repeats/near-repeats⁶, and G is the gold standard set. We note that this is a variant of precision/recall metrics defined for image captioning in (Deitke et al., 2024), however unlike Deitke et al., we explicitly use Φ_{claim} in computing both precision and recall. This is necessary as we need to accurately penalize any spurious claims generated by the finetuned model. Works by (Gu et al., 2025; Liu et al., 2023) are relevant here.

Analogously, for Task 3 – extraction of investigation proposals – we define precision and recall similarly but use a different pairwise boolean judge function Φ_{IP} *mutatis mutandis*. See Appendix C for prompt details.

6 Results

6.1 Non-technical Abstract Generation

Table 2 shows the results for Task 1. Both Mistral and Qwen models demonstrated strong performance, with fine-tuning providing modest improvements. The Mistral model outperformed Qwen on almost all metrics, achieving a BERTScore-F1 of 0.8561 after fine-tuning (+0.36% relative improvement). ROUGE scores were generally low (0.01–0.22), reflecting the stylistic differences between technical and non-technical abstracts.

6.2 Scientific Claim Extraction

For Task 2 (claim extraction), fine-tuning yielded substantial improvements. As shown in Table 3, the fine-tuned Mistral model achieved a precision of 0.7450 (+116.7% relative improvement), recall of 0.7098 (+59.5%), and F1 of 0.7097 (+101.8%). The Mistral model consistently outperformed Qwen, though both showed significant benefits from fine-tuning.

⁵We tried several slight edits of the prompts and found them to be robust to such changes.

⁶We determine repeats and near-repeats in the generation by thresholding cosine similarity calculated over a TF-IDF representation of the generated claims.

Metric	Mistral	Qwen
Precision	0.7450 (+116.7% ↑)	0.6839 (+107.1% ↑)
Recall	0.7098 (+59.5% ↑)	0.6611 (+7.8% ↑)
F1	0.7097 (+101.8% ↑)	0.6541 (+63.3% ↑)

Table 3: Finetuning performance for Mistral-7B-instruct-v0.3 and Qwen2.5-7B-Instruct models for Claim Extraction from abstracts (Task 2), with relative improvements over the corresponding unfinetuned model indicated in green. Error bars for all metrics at 95% confidence intervals range between 0.0038–0.0055. Mistral model outperforms Qwen on almost all metrics for this task regardless of finetuning. We note the large positive percent changes, sometimes improvements as large as 2x, indicate finetuning is indispensable for claim extraction. Mistral model outperforms Qwen on almost all metrics for this task.

6.3 Investigation Proposal Extraction

Similarly, Task 3 (proposal extraction) showed dramatic improvements with fine-tuning. As shown in Table 4, the Mistral model achieved a precision of 0.7351 (+18.24%), recall of 0.7539 (+127.24%), and F1 of 0.7261 (+90.97%) after fine-tuning. The relative improvements were even larger for the Qwen model, though Mistral still performed better overall.

Metric	Mistral	Qwen
Precision	0.7351 (+18.24% ↑)	0.7245 (+70.07% ↑)
Recall	0.7539 (+127.24% ↑)	0.6865 (+81.57% ↑)
F1	0.7261 (+90.97% ↑)	0.6827 (+112.60% ↑)

Table 4: Finetuning performance for Mistral-7B-instruct-v0.3 and Qwen2.5-7B-Instruct models for extraction of Investigation Proposals from award abstracts (Task 3), with relative improvements over the corresponding unfinetuned model indicated in green. Error bars for all metrics at 95% confidence intervals range between 0.0036–0.0073. Mistral model outperforms Qwen on almost all metrics for this task regardless of finetuning. We note the large positive percent changes, sometimes improvements as large as 2x, indicate finetuning is indispensable for this task. Mistral model outperforms Qwen on almost all metrics for this task.

Since Mistral models seemed to have an edge over the Qwen2.5 models for these tasks, we also trained a Mistral only version of on the NSF-SCIFY-20K subset which spans all NSF directorates. The results for that can be found in Appendix F.

7 Error Analysis

Of the three proposed tasks, we consider the claim extraction task as a canonical task for performing error analysis. To do so, we consider another 120 awards from the test portion of NSF-SCIFY-MATSCI and NSF-SCIFY-20K. These were stratified sampled across the five areas of interest (similar to Section 4.3). We then generate the claims using a Mistral-7B model finetuned on NSF-SCIFY-20K, resulting in 802 claims. A careful examination revealed around 2.6% of the generated claims were incorrect. To dive deeper, we categorized the erroneous claims into 5 categories. We list them here with examples:

1. Overconfidence: The claim can be overconfident about information that has qualifiers in the supporting document text (award abstract).

Award ID: 9820570

Extracted Claim: The research areas include knot theory, immiscible fluids and geodesic nets, ergodic theory, commutative algebra and vector-valued forms.

Analysis: The abstract states 'probably in the areas of,' indicating potential areas, not certainty.

2. Mixing Information: The claim can mix information from two sentences together to form wrong information.

Award ID: 1205671

Extracted Claim: The SEAQUEST experiment at Fermilab has successfully measured the asymmetry of up and down anti-quarks in the nucleon.

Analysis: The abstract mentions that SEAQUEST will follow the successful E866 measurement with more precise data, and thus it does not say SEAQUEST has already successfully measured that, but the success is describing the previous E866.

3. Overgeneralization: The claim can overgeneralize what the supporting document implies.

Award ID: 0957482

Extracted Claim: The methodology is potentially environmentally benign.

Analysis: The abstract mentions non-dangerous chemicals but does not specifically state that the methodology is environmentally benign.

4. Information Omission: The claim might omit important information from the abstract and thus the meaning is changed.

Award ID: 9409461

Extracted Claim: Frequency-domain techniques can display trade-offs between output performance and sensitivity reduction.

Analysis: The claim frames output performance and sensitivity reduction as two separate quantities and leaves out bandwidth, so it does not accurately reflect the abstract.

5. Hallucinations about Administrative Metadata:

The model can sometimes hallucinate claims regarding where the funding is from and which institutions are included. While hallucination is a serious issue, it is worth noting that for this dataset and model scientific claims seem to be rarely hallucinated. In our study, all hallucinations were connected with administrative metadata.

Award ID: 0542751

Claim: The award is funded under the American Recovery and Reinvestment Act of 2009 (Public Law 111-5).

Reasoning: This claim is not mentioned in the abstract.

To mitigate these errors, uncertainty calibration and prompting strategies can reduce overconfidence and overgeneralization, encouraging the model to reflect source qualifiers. Fine-tuning with more annotated data and enforcing stricter alignment between claims and source text can address mixing information and omission issues. Retrieval-augmented generation and chain-of-thought prompting may also promote better grounding. For hallucinations about administrative metadata, entity verification or output constraints based on structured data can help. Combining these approaches with human-in-the-loop evaluation might further improve claim extraction reliability.

We performed a similar error analysis on claims extracted from Claude (See section 3.3). Our findings revealed a smaller error-rate (2.1% as opposed to 2.6%), and of the only 10 erroneous claims, 5 were hallucinations of administrative data.

8 Discussion and Conclusion

We introduced NSF-SCIFY, a large dataset of 2.8 million scientific claims and proposals from 400,000 NSF grant abstracts across all science and mathematics disciplines. Focused subsets include NSF-SCIFY-MATSCI (114,000 materials science claims) and NSF-SCIFY-20K (135,000

claims from five directorates). Experiments demonstrate that fine-tuning language models on NSF-SCIFY significantly improves scientific claim and proposal extraction, with relative performance gains often exceeding 100%. Non-technical abstract generation saw modest improvements due to strong baselines. Stylistic differences between technical and non-technical abstracts offer potential for science communication. Our claim taxonomy identifies prevalent assertion types like capability/application and problem/knowledge gap statements. NSF-SCIFY’s unique advantages include its vast scale, high quality from NSF expert review, comprehensive coverage of scientific domains, a temporal span from 1970-2024 enabling longitudinal studies, and, for recent grants, links to resulting publications. Error analysis showed high precision but lower recall in claim extraction, with error categories guiding future improvements. NSF-SCIFY opens new research avenues in large-scale claim verification, scientific discovery tracking, and meta-scientific analysis, a key resource for understanding scientific assertions at their origin.

9 Limitations

Source Material Scope. The dataset, derived from NSF award abstracts, offers insights into early-stage scientific claims from a rigorously reviewed, cross-disciplinary source. However, it currently excludes claims from unfunded proposals or international contexts. Future work may expand to other agencies and sources.

Extraction Methodology. Our approach utilizes zero-shot prompting with large language models, refined by prompt engineering and selective human validation. While manual evaluation shows high precision, some residual errors in recall and claim boundary identification may exist, especially for nuanced assertions. Fine-tuning on this dataset significantly improves extraction, indicating potential for further methodological refinement and broader annotation.

Evaluation Design. We introduced LLM-based metrics for evaluating claims and investigation proposals, offering a nuanced assessment beyond lexical overlap. These metrics correlate well with human judgment in samples, but broader validation across more scientific domains is needed to confirm their robustness. The public dataset and code aim to facilitate such community efforts.

Temporal and Linked Data Coverage. Spanning over five decades and including recent linked publication metadata, the dataset’s systematic outcome tracking is limited for older awards. This restricts longitudinal analysis of claim evolution from proposal to publication. Broader, consistent outcome reporting could enrich NSF-SCIFY for deeper research trajectory studies.

Generalizability. While designed and validated for National Science Foundation abstracts, whose structure may differ from other scientific communications, the general framework is adaptable. It could be extended to related corpora like other funding agencies, patent abstracts, or scientific news, creating opportunities for future research.

10 Reproducibility Statement

To foster research on large-scale claim extraction, we are releasing our datasets, training code, and trained models:

- NSF-SCIFY-MATSCI: Materials Science subset with extracted claims, investigation proposals, and resolved publication information.
- NSF-SCIFY: Similar in content to NSF-SCIFY-MATSCI, but a larger superset spanning all of NSF awards database. The key difference is the claims and investigation proposals are extracted from our finetuned models instead of frontier LLMs.
- Our best finetuned model checkpoints for extraction of claims and investigation proposals at <https://huggingface.co/darpa-scify/nsf-scify-matsci-claims>.

References

Lutz Bornmann, Robin Haunschild, and Rüdiger Mutz. 2021. [Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases.](#) *Humanities and Social Sciences Communications*, 8(1).

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. [SPECTER: Document-level representation learning using citation-informed transformers.](#) In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom,

Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. 2024. [Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models.](#) *Preprint*, arXiv:2409.17146.

Thomas Diggelmann, Jordan L. Boyd-Graber, Janis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. [CLIMATE-FEVER: A dataset for verification of real-world climate claims.](#) *CoRR*, abs/2012.00614.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. [A survey on llm-as-a-judge.](#) *Preprint*, arXiv:2411.15594.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. [Lora: Low-rank adaptation of large language models.](#) *Preprint*, arXiv:2106.09685.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth e Lacroix, and William El Sayed. 2023. [Mistral 7b.](#) *Preprint*, arXiv:2310.06825.

Neema Kotonya and Francesca Toni. 2020. [Explainable automated fact-checking for public health claims.](#) In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online. Association for Computational Linguistics.

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.

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

Isabelle Mohr, Amelie W uhrl, and Roman Klinger. 2022. [CoVERT: A corpus of fact-checked biomed-](#)

ical COVID-19 tweets. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 244–257, Marseille, France. European Language Resources Association.

Minsu Park, Suman Kalyan Maity, Stefan Wuchty, and Dashun Wang. 2024. **Interdisciplinary papers supported by disciplinary grants garner deep and broad scientific impact.** *Preprint*, arXiv:2303.14732.

Ajay Patel, Jiacheng Zhu, Justin Qiu, Zachary Horvitz, Marianna Apidianaki, Kathleen McKeown, and Chris Callison-Burch. 2025. **StyleDistance: Stronger content-independent style embeddings with synthetic parallel examples.** In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8662–8685, Albuquerque, New Mexico. Association for Computational Linguistics.

Arkadiy Saakyan, Tuhin Chakraborty, and Smaranda Muresan. 2021. **COVID-fact: Fact extraction and verification of real-world claims on COVID-19 pandemic.** In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2116–2129, Online. Association for Computational Linguistics.

Mourad Sarrouiti, Asma Ben Abacha, Yassine Mrabet, and Dina Demner-Fushman. 2021. **Evidence-based fact-checking of health-related claims.** In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3499–3512, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Liyang Tang, Philippe Laban, and Greg Durrett. 2024. **MiniCheck: Efficient fact-checking of LLMs on grounding documents.** In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8818–8847, Miami, Florida, USA. Association for Computational Linguistics.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. **Fact or fiction: Verifying scientific claims.** In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. 2022. **SciFact-open: Towards open-domain scientific claim verification.** In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4719–4734, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Fengli Xu, Lingfei Wu, and James A. Evans. 2022. **Quantifying hierarchy in scientific teams.** *Preprint*, arXiv:2210.05852.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

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*.

Appendix

A Complete Prompt for Extracting Claims and Investigation Proposals

You are an expert materials science researcher. Given an input JSON description of an NSF material science award abstract, parse out the technical and nontechnical abstracts, and identify the verifiable claims and research/investigation proposals the abstract makes. Be thorough. Answer in the following JSON format:

```
{
  "award_id": "", // copied from input
  "technical_abstract": "" // technical
    abstract if present, otherwise
    contents of the abstract field in the
    input
  "non_technical_abstract": /non-technical
    abstract if present, otherwise empty
  "verifiable_claims": [ // list of strings
  ],
  "investigation_proposals": [ // list of
    strings
  ],
}
```

verifiable_claims are statements that the abstract claims to be true or states as an assumption explicitly or implicitly.

investigation_proposals are forward-looking statements that the abstract proposals to investigate as a part of this award.

Ensure that the output is in JSON format and that the JSON is valid.

We manually tested the prompt with a few award abstracts to make sure it was optimal for this task.

B Prompt for Task 2 evaluation function

Φ_{claim}

Check two scientific claims $c1$ and $c2$, if $c1$ is supported by $c2$. If $c2$ includes all the evidences for $c1$, but also includes additional content, then it should still be supported (YES). If not all information of $c1$ is included in $c2$, or if $c2$ contains information that conflicts with information in $c1$, then it should be unsupported (NO). Answer only as a YES or NO.

$c1$: { $c1$ }

$c2$: { $c2$ }

C Prompt for Task 3 evaluation function

Φ_{IP}

Check two investigation proposals $c1$ and $c2$, if $c1$ is supported by $c2$. If $c2$ includes all the investigations proposed by $c1$, but also includes additional proposals, then it should still be supported (YES). If not all proposed investigations by $c1$ is included in $c2$, or if $c2$ contains investigation actions that conflict with investigation actions in $c1$, then it should be unsupported (NO). Answer only as a YES or NO.

$c1$: { $c1$ }

$c2$: { $c2$ }

D Stylistic Differences between Technical and Nontechnical Abstracts

Figure 5 shows stylistic differences between technical and nontechnical abstracts.

E Finetuning Details

E.1 Data Preparation

Starting with 16,042 processed entries in NSF-SCIFY-MATSCI, we removed near-duplicates in technical and non-technical abstracts using trigram Jaccard similarity (threshold > 0.9), resulting in 11,569 data points. We further filtered cases where character-level 10-gram similarity between an entry’s technical and non-technical abstracts exceeded 0.6, yielding 11,141 final data points. We split this dataset into train/validation/test sets with 8,641/500/2,000 examples, respectively.

E.2 Finetuning Details

For fine-tuning, we used LoRA (Hu et al., 2021) with rank=128, lora_alpha=64 and a learning rate

of $1e-5$ scheduled linearly. We updated the query, key, value, and output projection layers, as well as MLP gate, up, and down projections. We ran the finetuning on an A100 GPU for 3 epochs, 100 warmup steps, and a batch size of 2 with 4 accumulated steps.

F Evaluation results for NSF-SCIFY-20K

Tables 5, 6, and 7 summarize the results for the three generation tasks defined in Section 5 on NSF-SCIFY-20K.

Model Metric	Base	Finetuned
BERTScore-F1	0.8514 \pm 0.0003	0.8500 \pm 0.0006
BERTScore-Precision	0.8515 \pm 0.0003	0.8513 \pm 0.0007
BERTScore-Recall	0.8516 \pm 0.0003	0.8496 \pm 0.0005
ROUGE-rouge1	0.3351 \pm 0.0013	0.3141 \pm 0.0023
ROUGE-rouge2	0.0705 \pm 0.0008	0.0936 \pm 0.0016
ROUGE-rougeL	0.1773 \pm 0.0008	0.1967 \pm 0.0016
ROUGE-rougeLsum	0.1982 \pm 0.0010	0.1998 \pm 0.0016

Table 5: Technical to Non-Technical Abstract Task: Mistral-7B

Model	Base	Finetuned
Precision	0.4146 \pm 0.0025	0.7526 \pm 0.0027
Recall	0.8141 \pm 0.0026	0.7354 \pm 0.0026
F-score	0.5247 \pm 0.0025	0.7268 \pm 0.0023

Table 6: Abstract to Claims Task: Mistral-7B

Model	Base	Finetuned
Precision	0.6222 \pm 0.0038	0.7219 \pm 0.0027
Recall	0.6364 \pm 0.0034	0.7359 \pm 0.0029
F1-score	0.5668 \pm 0.0033	0.7039 \pm 0.0026

Table 7: Abstract to Investigation Proposals Task: Mistral-7B

G Examples of Extracted Claims and Investigation Proposals

Tables 8 and 9 provide a sampling of the extracted claims and investigation proposals.

H Examples of Scientific Claim Categories

Please see Table 10 for the examples.

I AI Writing/Coding Assistance Disclosure

In accordance with the ACL Policy on AI Writing Assistance⁷, the authors attest that we used generative AI tools for assistance purely with the language of the paper, including spell checking, grammar fixes, and proof reading. Additionally, we used GPT-4o to fix LaTeX issues, and to generate LaTeX tables from spreadsheets. In all such uses, the outputs were verified by the first author for correctness.

⁷https://www.aclweb.org/adminwiki/index.php/ACL_Policy_on_Publication_Ethics#Guidelines_for_Generative_Assistance_in_Authorship

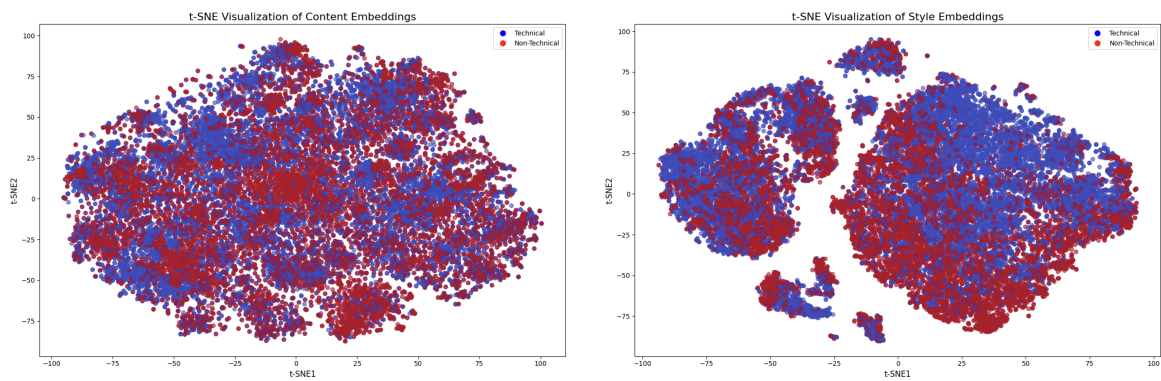


Figure 5: The t-SNE plot of comparing content embeddings from SPECTER (Cohan et al., 2020) and style embeddings from STEL (Patel et al., 2025) for technical and non-technical abstracts in NSF-SciFY-MATSCI. The somewhat clear separation between technical and non-technical abstracts when using style embeddings indicate marked stylistic differences between the two kinds abstracts.

Award ID	Title	Extracted Claims
2324035	DMREF: Developing and Harnessing the Platform of Quasi-One-Dimensional Topological Materials for Novel Functionalities and Devices	<p>[</p> <p>"Topological insulators are electrically insulating in the bulk but host conductive surface states that are immune to impurities.",</p> <p>"Current TI materials face critical challenges that limit their potential.",</p> <p>"Quasi-1D structures promise to overcome challenges faced by current TI materials.",</p> <p>"Most identified topological insulators are either strongly bonded bulk materials or layered van der Waals materials.",</p> <p>...</p> <p>]</p>
9814055	Kinks and Surface Potentials	<p>[</p> <p>"Atomically flat terraced surfaces for thin TEM samples can be prepared under moderate (10^{-7} Torr) vacuum conditions by annealing in oxygen or vacuum for materials such as sapphire, SiC and MgO.",</p> <p>"\Forbidden\" Bragg reflections arise from the stacking fault between partial dislocations.",</p> <p>"The surface potential is critical for chemical reactions at surfaces, adsorption, catalysis, epitaxy, diffusion bonding process, oxidation, and semiconductor crystal growth.",</p> <p>...</p> <p>]</p>
0821136	MRI: Acquisition of an Imaging Spherical Aberration Corrector and a Lorentz Lens for Magnetic Materials Characterization	<p>[</p> <p>"The attainable spatial resolution of uncorrected Lorentz instruments is in the range 10-15 nm.",</p> <p>"Delocalization effects cause significant image blurring in uncorrected Lorentz microscopes.",</p> <p>"Recent developments in aberration correction make it possible to correct the spherical aberration of a Lorentz lens.",</p> <p>"The size of written bits in state-of-the-art magnetic recording media is comparable to the magnetic resolution of uncorrected Lorentz microscopes.",</p> <p>"Transmission electron microscopes have suffered from lens aberration since their invention in the 1930s.",</p> <p>"The Hubble space telescope suffered from a similar aberration when first launched.",</p> <p>...</p> <p>]</p>

Table 8: A sample of extracted claims from the NSF-SciFY-MATSCI dataset. Award IDs are hyperlinked to the NSF's Award database.

Award ID	Title	Extracted Investigation Proposals
2324035	DMREF: Developing and Harnessing the Platform of Quasi-One-Dimensional Topological Materials for Novel Functionalities and Devices	<p>[</p> <p>"Predict, design, synthesize, and control topological phases in quasi-1D topological materials.",</p> <p>"Design and demonstrate emergent materials, functionalities, and devices, including moir\'e quasi-1D TIs, stable and high temperature quantum spin Hall (QSH) insulators, and quantum intelligent sensors.",</p> <p>"Expand research to include other selected quasi-1D materials families through collaborations.",</p> <p>"Discover or realize novel topological materials and phases.",</p> <p>"Study topological phase transitions and control.",</p> <p>"Investigate room-temperature QSH effect.",</p> <p>...</p> <p>]</p>
9814055	Kinks and Surface Potentials	<p>[</p> <p>"Observe dislocation kinks by atomic resolution TEM in materials such as sapphire, SiC and MgO.",</p> <p>"Use \"forbidden\" Bragg reflections to form lattice images without surface noise.",</p> <p>"Determine which process (kink formation, kink migration or obstacles along the dislocation line) limits kink (and hence dislocation) velocity, for given conditions of temperature and stress.",</p> <p>"Extend quantitative convergent-beam TEM measurements of bonding in crystals to the RHEED geometry to refine the electrostatic potential extending into the vacuum from ceramic surfaces.",</p> <p>"Measure modifications to the surface potential resulting from the deposition of a monolayer or more of atoms.",</p> <p>...</p> <p>]</p>
0821136	MRI: Acquisition of an Imaging Spherical Aberration Corrector and a Lorentz Lens for Magnetic Materials Characterization	<p>[</p> <p>"Acquire an imaging spherical aberration corrector and a Lorentz lens for magnetic materials characterization.",</p> <p>"Add these components to an existing FEI Titan 80-300 TEM.",</p> <p>"Bring the spatial resolution in Lorentz mode down to less than 1 nm, with negligible delocalization effects.",</p> <p>"Enable direct quantitative study of magnetic features at a length scale of around 1 nm",</p> <p>"Obtain new scientific results on material systems for which these observations were previously impossible.",</p> <p>"Impact a large number of research groups within CMU, as well as collaborations with local industry and several national laboratories.",</p> <p>...</p> <p>]</p>

Table 9: A sample of extracted investigation proposals from the NSF-SCIFY-MATSCI dataset. Award IDs are hyperlinked to the NSF's Award database.

Category: Capability/Application of Technology/Method

Memory-centric computing capitalizes on extensive parallelism in memory arrays.
The Illinois group has joined the fixed target COMPASS experiment at CERN.
An electronics company is involved in the project, making imaging products in this energy regime.

Category: Definition/Classification

The RV Weatherbird II is owned and operated by the Bermuda Biological Station for Research (BBSR), Inc.
The program will include topics such as dark matter, dark energy, inflation, and gravitational waves.
The shear zone in question is the Cuyamaca-Laguna Mountains shear zone.

Category: Statement of Problem/Knowledge Gap

Current efforts on analyzing tree-informed compositional data are primarily designed for individual applications.
CU began the Guerrero GPS project in 1997.
High pressure-low temperature metamorphism is often obscured by post-tectonic thermal equilibration or later deformation and mineral growth.

Category: Experimental Result/Finding/Measurability

Lattice QCD has made important progress.
RBP repression is absent when an oncoprotein is present.
Over 100 of 650 U.S. electronics fabricators have gone out of business in the past five years, according to a 1999 White Paper by the Interconnection Technology Research Institute.

Category: Established Scientific Fact/Principle

Dynamic programming includes well-known search algorithms like breadth-first search, Dijkstra's algorithm, A*, value iteration and policy iteration for Markov decision processes.
The electron carries a magnetic moment.
Stars in clusters evolve off the main sequence, become red giants, and ultimately horizontal branch stars.

Category: Observed Phenomenon/Property

The lake level of Laguna Paron was artificially lowered in 1985.
Laminated sediments are exposed in Laguna Paron, Peru.
The study sites exhibit extreme differences (1 to 2 orders of magnitude) in larval settlement.

Category: Process/Mechanism Description

Exciton-phonon and exciton-exciton interactions contribute to decoherence at finite temperatures.
The fidelity of translation is determined by the accuracy of aminoacyl-tRNA selection by ribosomes and synthesis of cognate amino acid/tRNA pairs by aminoacyl-tRNA synthetases.
The evaluation process includes both direct and indirect measures of student success and learning.

Category: Hypothesis/Theoretical Prediction

Assemblages that combine human-technology partnerships are stronger than individual humans or machines.
Mating advantage in guppies appears to result from female sexual responses to unusual males.
The long wavelength part of the CBR spectrum is important for constraining the evolution of the intergalactic medium.

Table 10: Scientific claim categories found in NSF-SCIFY and 3 randomly selected examples for each category.

Hierarchical Attention Adapter for Abstractive Dialogue Summarization

Raymond Li[†], Chuyuan Li[†], Gabriel Murray[‡], Giuseppe Carenini[†]

[†] University of British Columbia, Vancouver, BC, Canada

[‡] University of Fraser Valley, Abbotsford, BC, Canada

{raymondli, carenini}@cs.ubc.ca, chuyuan.li@ubc.ca
gabriel.murray@ufv.ca

Abstract

Dialogue summarization is still a very challenging task even for large language models (LLMs). On the one hand, some previous approaches have pre-trained language models specifically for dialogue understanding and summarization, but they have been limited to relatively small language models such as BART and T5. On the other hand, other works have tried to directly exploit the dialogue semantics and discourse structures in their modeling effort, but by construction, they require access to those structures, which is in itself a largely unsolved problem. In this paper, we synergistically combine these two ideas in an approach that can be seamlessly integrated into the decoder-only architecture adopted by the most state-of-the-art LLMs. In particular, our novel solution leverages the parameter-efficient fine-tuning (PEFT) paradigm to model the hierarchical structure of dialogues, where input sequences are naturally segmented into dialogue turns, and then fine-tune the model for abstractive summarization. From experiments on two datasets, we find that Hierarchical Attention Adapter outperforms all baseline adapter methods on SummScreen, where our approach can also be combined with LoRA to achieve the best performance on SamSum.

1 Introduction

The explosion in real-time messaging, consultation forums, and online meetings has resulted in a vast amount of conversational data, necessitating more efficient methods for understanding and extracting key information. Dialogue summarization, which aims to automatically distill salient information from dialogues, has been widely applied across various scenarios in different domains. These include task-oriented dialogues such as customer service (Feigenblat et al., 2021), law (Duan et al., 2019), medical care (Joshi et al., 2020), and open-ended dialogues like chit-chat (Chen et al., 2021b), screen

plays (Chen et al., 2021a), and forum discussions (Chowdhury and Chakraborty, 2019).

Conventional dialogue summarization models typically approach the task as a sequence-to-sequence problem and fine-tune encoder-decoder models such as BART (Lewis et al., 2020a) and T5 (Raffel et al., 2020). Although these methods show promising results (Liu et al., 2021a; Wang et al., 2023; Zhong et al., 2022; Cho et al., 2024), they are limited to relatively small Transformer architecture models and cannot be easily employed on large decoder-only models. Another line of research examines dialogue specific features such as speaker marker, discourse structure, topic changes, and co-reference information (Chen and Yang, 2021a; Liu et al., 2021b; Cho et al., 2024) and incorporates dialogue semantics and structural information into model pre-training, oftentimes involving an intermediate stage for dialogue-specific information extraction. However, these methods require additional effort and the results constructed from such information may not be accurate.

On the other hand, recent advances in Large Language Models (LLMs) have revolutionized the field of NLP and have become an essential building block in various intelligent user-facing applications (Bang et al., 2023; Bubeck et al., 2023), resulting in a shift in focus from relatively small encoder-decoder models to large-scale decoder-only models. The remarkable achievements of these LLMs can largely be attributed to research on model scaling (Brown et al., 2020; Chowdhery et al., 2023; Workshop et al., 2022), where increasing the number of model parameters and the volume of pre-training data can lead to significant enhancements of their capabilities to understand and generate human language. However, despite the success of sophisticated prompting (Wei et al., 2022; Zhou et al., 2023) and demonstration selection (Lewis et al., 2020b; Rubin et al., 2022) strategies, there remains a noticeable performance gap compared to

fine-tuning (Liu et al., 2022; Mosbach et al., 2023), especially for tasks such as dialogue summarization where the input sequences can be long and possess hierarchical structures.

In this paper, we synergistically combine the ideas of using modern LLMs with dialogue structural information in dialogue summarization task, without explicitly injecting rigid linguistic structures. To this end, we select GPT-style decoder-only architecture as our LLM backbone. This simplified architecture allows for more efficient pre-training through the language modeling objective, where the model can quickly process and generate tokens without first transforming the input sequence into an abstract representation by the encoder. However, an innate drawback of such architecture is the token-level unidirectional flow, which limits the ability to model the full context of the dialogue, especially the nuances and dependencies that emerge from future turn back to past ones. By contrast, encoder-decoder architecture encodes the input sequence using bidirectional conditioning and results in inherently stronger representations compared to the causal conditioning representation. In order to fine-tune decoder-only models more effectively, we propose a novel parameter-efficient fine-tuning (PEFT) architecture that encodes the input with bidirectional contextualization.

Another great challenge in dialogue summarization is the length of text, which can sometimes exceed the model’s input limit. While recent efforts have managed to increase the context window of the model by scaling the positional embeddings (Press et al., 2022; Su et al., 2024) and reducing the complexity of attention through sparsity (Child et al., 2019; Jaszczur et al., 2021), the theoretical support for long context are often measured with the language modeling loss and synthetic tasks, which do not comprehensively demonstrate their effectiveness in practical applications. Inspired by previous studies (Nguyen et al., 2020; Madaan et al., 2023; Du et al., 2023) demonstrating that hierarchical structures of input can significantly enhance downstream performance, especially for tasks with naturally segmented input sequences, we propose encoding dialogues at both the speech turn level and the dialogue level. This is achieved through two attention layers in our adapter module. Our proposed adapter, called **Hierarchical Attention Adapter**, can incorporate the interactions of speech turns in dialogues naturally without the need for external structural integration.

To summarize, our contributions are threefold: (1) We propose a novel PEFT architecture: Hierarchical Attention Adapter, that incorporates bidirectional contextualization to model the hierarchical structure of the dialogue sequence; (2) Our experiments on two dialogue summarization datasets demonstrate the effectiveness of our approach, where we achieve the best overall performance over the baseline either using our method directly or combining with LoRA; (3) We analyze the importance of each layer of adapters and ways to represent dialogue rounds, and provide useful insights for future PEFT approaches.

2 Related Work

2.1 Dialogue Summarization

Dialogue summarization, aimed at distilling the salient information of dialogue into a concise summary, has received more attention as virtual conversations have become increasingly prevalent (Jia et al., 2023). While the standard approach for neural abstractive summarization follows the sequence-to-sequence generation paradigm (Sutskever et al., 2014), where an autoregressive model generates the summary conditioned on the input text (Rush et al., 2015; Nallapati et al., 2016; See et al., 2017; Lewis et al., 2020a; Zhang et al., 2020), studies on dialogue summarization often exploited the semantic and discourse graph structures of the input by explicitly incorporating those structures to model the high-level interactions between dialogue segments (Hardy and Vlachos, 2018; Chen and Yang, 2021b; Hua et al., 2023). However, these methods typically require additional steps to obtain the dialogue graphs making them impractical for most application scenarios. Other studies have proposed pre-training language models specifically for dialogue understanding and summarization (Gu et al., 2021; Zhong et al., 2022), their scales are order-of-magnitude smaller compared to the general-purpose Large Language Models (Touvron et al., 2023b; Jiang et al., 2023) resulting in often inferior performance in poor generalizability. In contrast, our approach can be seamlessly integrated into the decoder-only architecture adopted by the most state-of-the-art LLMs.

2.2 Modeling Structures

While Graph Neural Networks (GNNs) have traditionally been the de-facto standard for graph representation, the Transformer model has garnered a

lot of popularity due to its superior performance in modeling graph structures (Hussain et al., 2022; Wu et al., 2022, 2023). Similar to GNNs, the self-attention mechanism in the Transformer encoder aggregates the node embeddings to update the representation of each node in a fully connected graph. In the domain of NLP, one popular area of research is to model explicitly defined linguistic structures within the input sequence. While a majority of studies have focused on sentence-level dependency or constituency trees (Wu et al., 2018; Hao et al., 2019; Strubell et al., 2018; Wang et al., 2019a,b), multi-sentential discourse structures have also been found to be beneficial for more practical downstream tasks such as summarization (Xiao et al., 2020; Xu et al., 2020; Feng et al., 2020; Dong et al., 2021). For instance, Chen and Yang (2021b) incorporated discourse relations between utterances and action for dialogue summarization, while Du et al. (2023) utilized discourse structures to propagate hidden representation for question answering. Meanwhile, another line of work aims to implicitly learn the structure of language through architectural design, allowing the model to have more flexibility in learning representations beneficial for the task (Nguyen et al., 2020; Madaan et al., 2023). Given the substantial proficiency of LLMs in representing and comprehending natural language, it is questionable whether the integration of rigid linguistic structures, which often require multi-stage pipelines to acquire, can further enhance the model’s capabilities in downstream tasks. Therefore, our proposed technique aims to allow the model to learn multi-sentential structures based on the internal representations of the LLM.

2.3 Large Language Models

While earlier work on masked language models are often designed to encode a contextualized representation of the input sequence (Peters et al., 2018; Devlin et al., 2019), autoregressive pre-training was found to be much effective for language generation tasks (Lewis et al., 2020a; Raffel et al., 2020). The current lineage of large language models can trace their origins to the GPT-family (Radford et al., 2018, 2020; Brown et al., 2020), where they find that scaling the decoder-only architecture can lead to an improved model capacity on downstream tasks (following scaling laws (Kaplan et al., 2020; Hoffmann et al., 2022)) while displaying emergent abilities to solve complex tasks through few-shot demonstrations (Brown et al.,

2020) and chain-of-thought prompting (Wei et al., 2022). Following the success of proprietary LLMs (e.g., ChatGPT, Claude, etc.) as general-purpose task solvers through instruction-tuning (Ouyang et al., 2022), the open-source community has also released a large number of publicly available models for researchers to work on (Zhang et al., 2022; Bai et al., 2023; Almazrouei et al., 2023). For example, Llama (Touvron et al., 2023a) is one of the first collections of open-sourced models pre-trained exclusively using publicly available datasets, Llama-2 (Touvron et al., 2023b) improved upon the Llama by training on more data and doubling the context length from 2K to 4K tokens, XGen (Nijkamp et al., 2023) further increased to context length to 8K through pre-training in stages with increasing sequence length, while Mistral (Jiang et al., 2023) used sliding window attention (Child et al., 2019) to support a theoretical attention span of approximately 131K tokens. While other techniques have also been proposed to improve the long-context capabilities of LLMs through position encoding refinement and continual pre-training (Xu et al., 2023; Xiong et al., 2023), these techniques require tuning the full set of parameters which is far too expensive for practical applications.

3 Method

3.1 Hierarchical Attention Adapter

An overview of the architecture is presented in Figure 1. In our hierarchical attention adapter, we first project the LLM hidden states to dimension d using a single linear layer. At each output time step j , we decompose the projected hidden states based on input and out tokens where $H = [h_i; h_o]$, with $h_i \in \mathbb{R}^{n \times d}$ and $h_o \in \mathbb{R}^{j \times d}$ representing the hidden states of input and output tokens respectively. We model the hierarchical structure of the input dialogue sequence using a hierarchical self-attention module. To compute the dialogue turn embeddings $H_t \in \mathbb{R}^{t \times d}$, where t is the number of dialogue turns, we first apply an attention encoder to the hidden states of each dialogue turn. Finally, to model the coarse interactions between dialogue turns, we use another layer of attention layer on the turn embeddings to obtain the hierarchical representation H_d of the dialogue sequence. This module allows us to construct representations of the input sequence with bidirectional contextualization while incorporating the hierarchical structure of the dialogue. The hierarchical representation is ex-

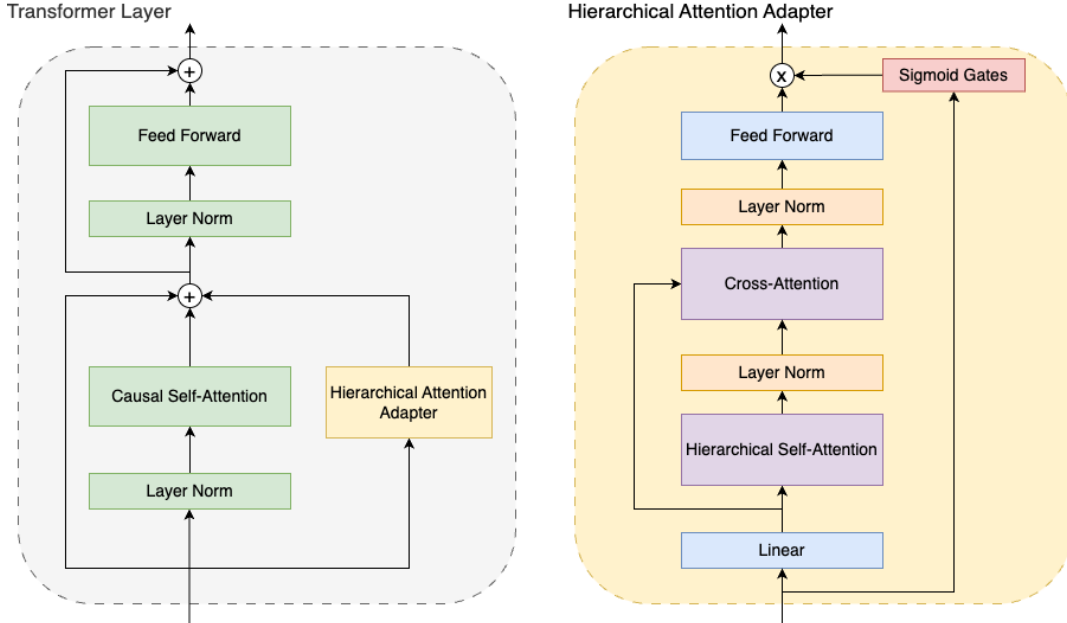


Figure 1: Overview of the Hierarchical Attention Adapter inserted in-parallel to the causal self-attention layer of the decoder transformer model. Specifically, the Hierarchical Self-Attention contains two attention layers: one encodes the hidden state of individual speech turn and one encodes the interactions of different speech turns.

pressed in Equation 1, where M_i denotes the mask for selecting the i^{th} dialogue turn.

$$H_t = [\text{Attn}(M_i \odot h^{(i)}); \dots; \text{Attn}(M_t \odot h^{(t)})] \quad (1)$$

$$H_d = \text{Attn}(H_t) \quad (2)$$

To enable the decoder to use the hierarchical representation of the input sequence, we follow the intuition of the cross-attention mechanism in the original transformer architecture (Vaswani et al., 2017). Specifically, we compute the weighted sum of the projected hidden states using a cross-attention layer with the hierarchical representation H_d as query and H as key and value. Finally, we use a two-layer feed-forward module to project the adapter representation to the dimension of LLM hidden states. The final adapter representation H_a can be expressed in Equation 3.

$$H_a = \text{FF}(\text{Attn}(Q = H^{(d)}, K = V = H)) \quad (3)$$

3.2 Gated Addition

Following the work by Mao et al. (2022), we use a gated additive method to inject the adapter output into the hidden states of the LLM. To do so, our learnable gate consists of a two-layer feed-forward module that takes the previous LLM hidden layer as input and uses a sigmoid activation function to map the final output to a value between 0 and 1.

The gate acts as a learnable scaling factor that estimates the importance of our module based on the input hidden states. We insert our adapter in parallel with the LLM self-attention layer, where the gated adapter output is combined with the attention output and residual connection through element-wise addition.

4 Experiments

4.1 Model

We adopt Mistral-7B (Jiang et al., 2023) as the base model for all our experiments. This model has 32 hidden layers (32 heads per layer) and uses the sliding window attention (SWA) (Child et al., 2019) with a window size of 4096 to support sequences of up to 4096×32 tokens. The main reason for choosing this model is due to the memory efficiency of SWA, as the standard self-attention mechanism has a quadratic complexity w.r.t. the sequence length.

4.2 Datasets and Metrics

We choose two widely used dialogue datasets: SamSum (chit-chat) (Gliwa et al., 2019) and SummScreen (screen plays) (Chen et al., 2022). From the statistics in Table 1, we can see that SamSum is an easier dataset with both short dialogues and reference summaries, while SummScreen is a much harder dataset with both extremely long dialogues and summaries requiring the model to learn the

Dataset	Domain	Dialogue	Summary	Train Size	Validation Size	Test Size
SamSum	Chat	83.9	20.3	14,731	818	819
SummScreen	Screenplay	6,612.5	337.4	18,915	1,795	1,793

Table 1: Statistics of the three datasets used in our experiments. The number of sentences in original dialogues and reference summaries, the number of documents in train, validation, and test sets are reported.

long-term dependencies between turns to generate a coherent summary. For SummScreen, we remove ultra-long examples (>16,000 tokens) to avoid out-of-memory during training.

For evaluation metrics, we use the popular ROUGE (Lin, 2004), which measures n-gram overlaps with the reference summary, as well as GPT3Score (Fu et al., 2023), which employs generative pre-trained models to evaluate text quality by calculating the length-normalized conditional log probability of the evaluated text (reference given candidate and candidate given reference) given task-specific prompts and aspect definitions. For ROUGE metric, we compute ROUGE-1, ROUGE-2, and ROUGE-Lsum (all are F1 scores) using the `rouge-score`¹ package, which respectively measures the overlap of unigram, bigram, and the longest common sub-sequence for each sentence. Following Grusky (2023), we compute the ROUGE scores without stemming and stopword removal, which is consistent with the original ROUGE-1.5.5 implementation by Lin (2004). For GPTScore, we use `text-davinci-002` (Brown et al., 2020) since it is currently the most powerful text completion model accessing through the OpenAI API² that supports token probabilities and has shown to be highly correlated with human judgment (Fu et al., 2023). We compute the harmonic mean of the conditional probability for the candidate summary predicted by the reference and vice versa. The conditional probabilities are computed based on three aspects, namely, Informativeness (I), Naturalness (N), and overall Quality (Q).

4.3 Baselines

To evaluate the performance of our proposal, we compare against three baseline PEFT methods, namely, Low-Rank Adaptation (LoRA) (Hu et al., 2022), Bottleneck Adapter (Houlsby et al., 2019), and the standard Attention Adapter. In particular, LoRA injects trainable rank decomposition matrices to approximate the gradient updates during fine-

tuning, Bottleneck Adapter injects two-layer MLPs sequential to the self-attention and feed-forward modules of the LLM, while Attention Adapter is a standard decoder attention layer with casual self-attention followed by a feed-forward layer. From results presented in prior studies (Yu et al., 2023), we expect both LoRA and the Bottleneck Adapter to be competitive baselines.

4.4 Hyperparameter Settings

We apply our Hierarchical Attention Adapter in parallel to the self-attention sub-layer of the decoder LLMs. We first project the LLM hidden states to our adapter dimension of $d_{\text{adapter}} = 128$, before applying our hierarchical and cross-attention modules with 4 attention heads each. Finally, we apply the two-layer feed-forward module with the same SiLU activation (Elfwing et al., 2018) as Mistral. Following the settings by Mao et al. (2022), we train our model with AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of $1e^{-4}$ with linear decay to 0 and a batch size of 64. For the baseline methods, we apply rank 32 LoRA vectors to the query and key vectors, following a hyperparameter search for $\text{rank} \in \{8, 16, 32, 64\}$. For both the Bottleneck and Attention Adapter, we use the same hidden size of 128, where the Attention Adapter has the same number of attention heads (4) and learnable gate dimension as our Hierarchical Attention Adapter. All our experiments are conducted on machines with $4 \times$ A100 SXM4 GPU (40GB of memory).

To reduce the number of parameters and improve task performance, we also experiment with combining LoRA with our Hierarchical Attention Adapter. Following the findings by Li et al. (2023), where they find that only the top layers of the pre-trained model can effectively utilize injected adapters, we inject our adapters to the top-4 (excluding the last) layers of the LLM when combined with LoRA.

4.5 Results

From the results presented in Table 2, we see that combining LoRA with Hierarchical Attention

¹<https://pypi.org/project/rouge-score>

²<https://platform.openai.com/docs/models>

Model	SamSum						SummScreen					
	ROUGE			GPTScore			ROUGE			GPTScore		
	R1	R2	RLs	I	N	Q	R1	R2	RLs	I	N	Q
LoRA	.543	.303	.504	-.913	-.901	-.891	.305	.085	.295	-.666	-.667	-.666
Bottleneck	.540	.300	.501	-.909	-.895	-.887	.278	.068	.265	-.780	-.781	-.779
Attention	.519	.275	.480	-.930	-.915	-.907	.263	.078	.249	-.653	-.652	-.652
Ours	.545	.303	.507	-.911	-.897	-.888	.334	.089	.321	-.632	-.633	-.632
Ours + LoRA	.546	.306	.508	-.902	-.889	-.881	.326	.089	.318	-.702	-.703	-.704

Table 2: Evaluation comparisons with baseline models and our methods across three datasets using ROUGE and GPTScore metrics. R1, R2, RLs denote resp. ROUGE-1, ROUGE-2, and ROUGE-Lsum; I, N, and Q denote resp. Informativeness, Naturalness, and Quality.

Model	R1	R2	RL	RLs
ConDigSum (2021a)	.543	.293	.452	-
GPT3-finetuned (2022)	.534	.298	.459	-
ChatGPT (2023)	.408	.137	.315	-
InstructDS (2023)	.553	.313	.467	-
Ours	.545*	.303*	.466*	.507

Table 3: Comparison with state-of-the-art benchmarks on SAMSum dataset. We report ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-LSum metrics. * not significantly different with the best score (in bold).

Model	R1	R2	RL	RLs
DialogLM-sparse (2022)	.358	.083	.187	-
SPECTRUM (2024)	.358	.095	.212	-
Ours	.334	.089	.169	.321

Table 4: Comparison with state-of-the-art benchmarks on SummScreen. We report ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-LSum metrics.

Adapter (*LoRA + Ours*) achieved the best results on SamSum and using only the Hierarchical Attention Adapter achieved the best results on SummScreen. We also see that our method (with and without LoRA) achieves the highest improvement over the baselines on the SummScreen dataset. This is mainly due to the fact that the examples in SummScreen have the longest average input and output length, making the hierarchical representation more beneficial even without LoRA. This finding also confirms our hypothesis that decoder-only architectures struggle with significantly long contexts. On the contrary, we see that our model (Ours) does not outperform the baselines in the SamSum datasets, where we only see improvements when combined with LoRA. We hypothesize that due to the short context of SamSum, the casual self-attention of the decoder can sufficiently encode the input dialogue

for summarization. This could be due to the fact that hierarchical attention helps model the dialogue context while tuning the decoder self-attention (via LoRA) improves the context selection for generation. This leads to long summaries generated by our model, and partially explains the poor performance (lower precision) of our method on SamSum, where the short output length (20 tokens) requires the model to precisely generate the summary based on the context.

In addition, for reference, we present the performances of state-of-the-art models fine-tuned specifically for dialogue summarization task without using the PEFT paradigm. The benchmarks on SAMSum are (1) topic-aware BART model trained with contrastive learning ConDigSum (Liu et al., 2021a), (2) GPT3 fine-tuned with LoRA (Hu et al., 2022), (3) ChatGPT with instruction tuning (Wang et al., 2023), and (4) Flan-T5-XL with instruction tuning (Wang et al., 2023). As shown in Table 3, our method is comparable with the SOTA InstructDS model on all ROUGE metrics, demonstrating the effectiveness of our hierarchical adapter on decoder-only model. On SummScreen, the best performing models are given in Table 4: (1) DialogLM (Zhong et al., 2022) – an encoder-decoder model pre-trained using dialogue-tailored noise; (2) SPECTRUM (Cho et al., 2024), a speaker-enhanced model pre-trained on PEGASUS. These models are specially trained with dialogue data, while our approach can applied to any general-purpose LLMs. With the scaling effect, we expect larger decoder-only models to lead to greater improvements.

	R-1	R-2	R-L	GPTScore (Ave)
Last Token	0.534	0.295	0.496	-0.902
Mean-Pooling	0.541	0.303	0.502	-0.893
Attention	0.546	0.306	0.508	-0.891

Table 5: Results on SamSum for different turn embedding methods. R-1/2/L refer to ROUGE-1/2/Lsum metrics.

5 Analysis

5.1 Dialogue Turn Representations

We first study the techniques for constructing the representations for dialogue turns. Using the SamSum dataset, we use three different turn representations while keeping the rest of the architecture identical. From the results in Table 5, we see that using an additional layer of attention outperforms mean-pooling and using the last token of each turn. It is worth noting that not only is using the last token representation outperforms mean-pooling. We believe that it is mainly due to the autoregressive nature of the decoder-only architecture, where the representation of each token encodes information from all prior tokens.

5.2 Summary Length

We find that adapter-based models tend to generate longer summaries than LoRA. Conversely, overly brief summaries tend to omit salient information; we provide concrete examples in subsection 5.4. In particular, on SamSum, the average number of summary tokens of our Hierarchical Attention Adapter is 23, while LoRA produces an average of 18 tokens. However, when combining LoRA with our method, the predicted summary has an average of 20 tokens. On the SummScreen dataset, the average summary length for LoRA is 460 while the length for Hierarchical Attention Adapter and Hierarchical Attention Adapter + LoRA are 526 and 549, respectively.

5.3 Gate Values

To assess the relative importance of our adapter at each layer of the model, we analyze the gate output by computing the average absolute values over the development set of SamSum. For the results presented in Figure 2, we see that the average gate value gradually increases before drastically dropping off at the last layer. This is in contrast to the findings by previous work on encoder models (Rücklé et al., 2021; Li et al., 2023), where they re-

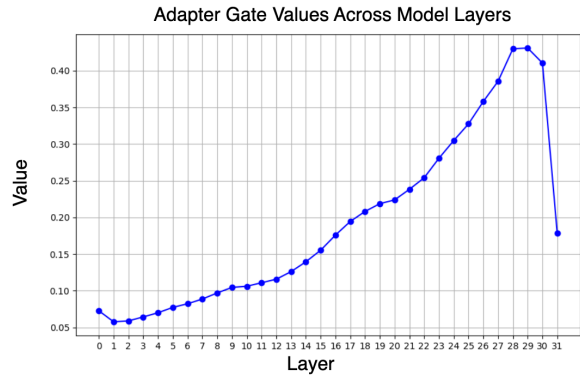


Figure 2: Gate value for our adapter at each layer of the Mistral-7B model. Values are averaged on the development set of SamSum.

ported performance benefits from inserting adapter modules in the final layer of the pre-trained model. We hypothesize the main reason is that since the last layer of the pre-trained decoder is directly used to predict the next word distribution over the vocab, the model cannot effectively learn to use the modified representation (from adapters) to encode the fine-grained contextual nuances required for word prediction. We believe these insights will benefit future studies on developing PEFT methods for pre-trained decoder models.

5.4 Qualitative Comparison

We perform a qualitative comparison between the summary generated by LoRA and our Hierarchical Attention Adapter. From the output snippets presented in Table 6, we see that on SamSum, our model tends to be more verbose during prediction, resulting in longer summaries that contains richer information. For the first example of SummScreen, we see that while both generated summaries do not capture the same content as the reference summary, the summary generated by LoRA begins to repeat towards the end. In contrast, the summary generated by our method captures more aspects of the conversation as it introduced six characters (rather than the three for LoRA). For the second example, we see that while both summaries contain information about *Kendall's trial*, the summary generated by LoRA also becomes repetitive towards the end. We hypothesize that this could be due to the intrinsic capacity of LoRA, as it only uses a small set of parameters to update the self-attention weights.

6 Conclusions and Future Work

In this work, we present Hierarchical Attention Adapter, a novel PEFT architecture that uses bidi-

Summary Type	Generated Summary
<i>SamSum</i>	
Reference	Kate broke her arm and she’s going to the hospital. She’d like to know whether her medical insurance covers hospital costs. Greg suggests her to call Linda or ask someone at the reception about it.
LoRA	Kate broke her arm and is going to the hospital.
Ours	Kate broke her arm and is being sent to the hospital. She wants to know if their medical insurance covers hospital costs.
<hr/>	
Reference	Casey got a new nail polish and did her nails herself. It took her nearly 4 hours, so she won’t do her friends’ nails, as it takes too long.
LoRA	Casey did her nails herself. It took her 4 hours.
Ours	Casey did her nails herself and they took her 4 hours. Amelia and Kristen want her to do their nails too but she refuses.
<hr/>	
<i>SummScreen</i>	
Reference	Adam bails Krystal out of jail and makes her an “offer” in exchange for her telling him where their mutual grandson is. She denies that she knows he’s alive and refuses his offer. But JR gets Bianca to admit that he is alive. . .
LoRA	Jonathan and Greenlee continue to argue about Ryan’s actions. Jonathan wants to know if Greenlee is going to stand by Ryan. Greenlee tells Jonathan that she will stand by Ryan. Jonathan tells Greenlee that he wants to know if she is going to stand by him. . .
Ours	Greenlee and Jonathan argue about Ryan and Kendall. Maggie comes in and tells them that she just saw Bianca. Greenlee and Jonathan argue about Kendall. Maggie comes in and tells them that she just saw Bianca. . .
<hr/>	
Reference	Kendall’s trial begins. Bianca urges Kendall to listen to what she learned in her dream about something she needs to find in order to help Kendall get acquitted. Jack proposed to Erica and she accepts. Edmund suddenly regains consciousness. . .
LoRA	Kendall is nervous about her trial. Bianca tells her that she had a dream that she was at Michael’s condo the night he was killed. Kendall tells her that it was impossible. Bianca tells her that she is worried about her. Kendall tells her that she is worried about her. . .
Ours	Kendall is in court for her trial. Bianca tells Kendall that she had a dream that she was at Michael’s condo the night that he died. Kendall tells Bianca that she cannot believe that she slept the whole day. . .

Table 6: Comparison of generated summary between LoRA and our Hierarchical Attention Adapter for two examples in the SamSum (top 2) and SummScreen (bottom 2) dataset. LoRA tends to produce overly brief summaries on SAMSum and repetitive content on ScreenSumm, indicating a limitation in generating longer outputs.

rectional contextualization to model the hierarchical structure of the input sequence for dialogue summarization task. Experiments on two datasets show that our proposed method outperforms other baselines for summaries with long context and achieves the best overall performance when combined with LoRA. We perform analysis on the average gate value to assess the relative importance of our adapter at each layer of the model and find that while the adapters of upper layers have higher

importance, the model learns to not use the final layer since it is used for computing the next-token probability.

For future work, we wish to perform additional analysis to study the usefulness of the different components in our proposed architecture and perform further experiments on additional datasets. We will also perform further evaluation to compare the faithfulness and factuality of summaries generated by different models. Lastly, while our

current proposal requires segment annotation of the input sequence (i.e., dialogue turns), we intend to extend our approach to implicitly learn the segment boundaries during training and generalize to other summarization tasks such as scientific paper summarization.

Limitations

Automatically assessing the quality of dialog summaries is a huge challenge. We recognize the importance of manual annotators for results comparison. However, human evaluation is costly and inefficient. While most of the dialogue summarization work relies heavily on the ROUGE score, we also report on the GPTScore with instruction prompts, which is an automatic metric that gives multi-faceted evaluation and is closely related to human judgment.

Other types of dialogue summarization tasks such as meeting summarization (e.g., AMI (Kraaij et al., 2005) ICSI meeting corpus (Shriberg et al., 2004)) often do not have enough examples for sufficiently fine-tuning an LLM. Real-world meeting dialogues often span multiple topics and include disfluencies, interruptions, and other artifacts. These characteristics make summarization more realistic but also more challenging. We plan to address these practical issues in future work.

Lastly, while PEFT methods such as LoRA are well optimized in existing libraries such as Ollama³, adapter-based methods often lack in inference speed due to the computation of additional modules. We hope our work can motivate future studies to efficiently integrate adapter modules into LLMs.

Ethical Considerations

We have taken proactive steps to address ethical concerns related to our research. Our testing datasets were carefully selected to minimize potential issues with biased or hateful content. If this method is applied to new datasets that involve recording multi-party dialogues, informed consent should be obtained from all participants. Because dialogue data may contain sensitive personal information, we urge caution in such applications, especially in summarization.

³<https://github.com/ollama/ollama>

Acknowledgments

The authors thank the anonymous reviewers for their valuable feedback and suggestions. The authors acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC). Nous remercions le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG) de son soutien.

References

- Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hessel, Julien Launay, Quentin Malartic, et al. 2023. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.
- Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenzhiang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. *A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity*. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Matias Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. *Language models are few-shot learners*. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. *Sparks of artificial general intelligence: Early experiments with gpt-4*. *arXiv preprint arXiv:2303.12712*.
- Jiaao Chen and Diyi Yang. 2021a. Structure-aware abstractive conversation summarization via discourse

- and action graphs. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1380–1391.
- Jiaao Chen and Diyi Yang. 2021b. **Structure-aware abstractive conversation summarization via discourse and action graphs**. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1380–1391. Association for Computational Linguistics.
- Mingda Chen, Zewei Chu, Sam Wiseman, and Kevin Gimpel. 2021a. Summscreen: A dataset for abstractive screenplay summarization. *arXiv preprint arXiv:2104.07091*.
- 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. Association for Computational Linguistics.
- Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021b. Dialogsum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074.
- Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*.
- Sangwoo Cho, Kaiqiang Song, Chao Zhao, Xiaoyang Wang, and Dong Yu. 2024. Spectrum: Speaker-enhanced pre-training for long dialogue summarization. *arXiv preprint arXiv:2401.17597*.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayanan Pilla, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. **PalM: Scaling language modeling with pathways**. *Journal of Machine Learning Research*, 24(240):1–113.
- Tanya Chowdhury and Tanmoy Chakraborty. 2019. Cqa-summ: Building references for community question answering summarization corpora. In *Proceedings of the ACM india joint international conference on data science and management of data*, pages 18–26.
- 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. Association for Computational Linguistics.
- Yue Dong, Andrei Mircea, and Jackie Chi Kit Cheung. 2021. **Discourse-aware unsupervised summarization for long scientific documents**. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1089–1102. Association for Computational Linguistics.
- Haowei Du, Yansong Feng, Chen Li, Yang Li, Yunshi Lan, and Dongyan Zhao. 2023. **Structure-discourse hierarchical graph for conditional question answering on long documents**. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6282–6293. Association for Computational Linguistics.
- Xinyu Duan, Yating Zhang, Lin Yuan, Xin Zhou, Xiaozhong Liu, Tianyi Wang, Ruocheng Wang, Qiong Zhang, Changlong Sun, and Fei Wu. 2019. Legal summarization for multi-role debate dialogue via controversy focus mining and multi-task learning. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 1361–1370.
- Stefan Elfving, Eiji Uchibe, and Kenji Doya. 2018. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural networks*, 107:3–11.
- Guy Feigenblat, Chulaka Gunasekara, Benjamin Szneider, Sachindra Joshi, David Konopnicki, and Ranit Aharonov. 2021. Tweetsumm-a dialog summarization dataset for customer service. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 245–260.
- Xiachong Feng, Xiaocheng Feng, Bing Qin, and Xinwei Geng. 2020. Dialogue discourse-aware graph model and data augmentation for meeting summarization. *arXiv preprint arXiv:2012.03502*.
- Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.
- Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. **SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization**. In *Proceedings of the 2nd Workshop on New*

- Frontiers in Summarization*, pages 70–79. Association for Computational Linguistics.
- Max Grusky. 2023. **Rogue scores**. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1914–1934. Association for Computational Linguistics.
- Jia-Chen Gu, Chongyang Tao, Zhenhua Ling, Can Xu, Xiubo Geng, and Daxin Jiang. 2021. **MPC-BERT: A pre-trained language model for multi-party conversation understanding**. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3682–3692. Association for Computational Linguistics.
- Jie Hao, Xing Wang, Shuming Shi, Jinfeng Zhang, and Zhaopeng Tu. 2019. **Multi-granularity self-attention for neural machine translation**. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 887–897. Association for Computational Linguistics.
- Hardy Hardy and Andreas Vlachos. 2018. **Guided neural language generation for abstractive summarization using Abstract Meaning Representation**. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 768–773. Association for Computational Linguistics.
- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack Rae, and Laurent Sifre. 2022. **An empirical analysis of compute-optimal large language model training**. In *Advances in Neural Information Processing Systems*, volume 35, pages 30016–30030. Curran Associates, Inc.
- Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. **Parameter-efficient transfer learning for NLP**. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.
- Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. **LoRA: Low-rank adaptation of large language models**. In *International Conference on Learning Representations*.
- Yilun Hua, Zhaoyuan Deng, and Kathleen McKeown. 2023. **Improving long dialogue summarization with semantic graph representation**. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13851–13883. Association for Computational Linguistics.
- Md Shamim Hussain, Mohammed J. Zaki, and Dharmashankar Subramanian. 2022. **Global self-attention as a replacement for graph convolution**. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '22*, page 655–665, New York, NY, USA. Association for Computing Machinery.
- Sebastian Jaszczur, Aakanksha Chowdhery, Afroz Mohiuddin, Lukasz Kaiser, Wojciech Gajewski, Henryk Michalewski, and Jonni Kanerva. 2021. **Sparse is enough in scaling transformers**. In *Advances in Neural Information Processing Systems*.
- Qi Jia, Yizhu Liu, Siyu Ren, and Kenny Q. Zhu. 2023. **Taxonomy of abstractive dialogue summarization: Scenarios, approaches, and future directions**. *ACM Comput. Surv.*, 56(3).
- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.
- Anirudh Joshi, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2020. Dr. summarize: Global summarization of medical dialogue by exploiting local structures. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3755–3763.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.
- Wessel Kraaij, Thomas Hain, Mike Lincoln, and Wilfried Post. 2005. The ami meeting corpus. In *Proc. International Conference on Methods and Techniques in Behavioral Research*.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. **BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880. Association for Computational Linguistics.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020b. **Retrieval-augmented generation for knowledge-intensive nlp tasks**. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

- Raymond Li, Gabriel Murray, and Giuseppe Carenini. 2023. [Mixture-of-linguistic-experts adapters for improving and interpreting pre-trained language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9456–9469. Association for Computational Linguistics.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81. Association for Computational Linguistics.
- Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohhta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. 2022. [Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 1950–1965. Curran Associates, Inc.
- Junpeng Liu, Yanyan Zou, Hainan Zhang, Hongshen Chen, Zhuoye Ding, Caixia Yuan, and Xiaojie Wang. 2021a. [Topic-aware contrastive learning for abstractive dialogue summarization](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1229–1243.
- Zhengyuan Liu, Ke Shi, and Nancy Chen. 2021b. [Coreference-aware dialogue summarization](#). In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 509–519.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *International Conference on Learning Representations*.
- Lovish Madaan, Srinadh Bhojanapalli, Himanshu Jain, and Prateek Jain. 2023. [Treeformer: Dense gradient trees for efficient attention computation](#). In *The Eleventh International Conference on Learning Representations*.
- Yuning Mao, Lambert Mathias, Rui Hou, Amjad Almahairi, Hao Ma, Jiawei Han, Scott Yih, and Madian Khabsa. 2022. [UniPELT: A unified framework for parameter-efficient language model tuning](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6253–6264. Association for Computational Linguistics.
- Marius Mosbach, Tiago Pimentel, Shauli Ravfogel, Dietrich Klakow, and Yanai Elazar. 2023. [Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12284–12314. Association for Computational Linguistics.
- Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. [Abstractive text summarization using sequence-to-sequence RNNs and beyond](#). In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290. Association for Computational Linguistics.
- Xuan-Phi Nguyen, Shafiq Joty, Steven Hoi, and Richard Socher. 2020. [Tree-structured attention with hierarchical accumulation](#). In *International Conference on Learning Representations*.
- Erik Nijkamp, Tian Xie, Hiroaki Hayashi, Bo Pang, Congying Xia, Chen Xing, Jesse Vig, Semih Yavuz, Philippe Laban, Ben Krause, et al. 2023. Xgen-7b technical report. *arXiv preprint arXiv:2309.03450*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.
- Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. [Deep contextualized word representations](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.
- Ofir Press, Noah Smith, and Mike Lewis. 2022. [Train short, test long: Attention with linear biases enables input length extrapolation](#). In *International Conference on Learning Representations*.
- Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. [Improving language understanding by generative pre-training](#). *OpenAI*.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2020. [Language models are unsupervised multitask learners](#). *OpenAI*.
- 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.
- Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. [Learning to retrieve prompts for in-context learning](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671. Association for Computational Linguistics.
- Andreas Rücklé, Gregor Geigle, Max Glockner, Tilman Beck, Jonas Pfeiffer, Nils Reimers, and Iryna Gurevych. 2021. [AdapterDrop: On the efficiency of adapters in transformers](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7930–7946. Association for Computational Linguistics.

- Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. [A neural attention model for abstractive sentence summarization](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389. Association for Computational Linguistics.
- Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. [Get to the point: Summarization with pointer-generator networks](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083. Association for Computational Linguistics.
- Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004. [The ICSI meeting recorder dialog act \(MRDA\) corpus](#). In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pages 97–100, Cambridge, Massachusetts, USA. Association for Computational Linguistics.
- Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. 2018. [Linguistically-informed self-attention for semantic role labeling](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5027–5038. Association for Computational Linguistics.
- Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. [Roformer: Enhanced transformer with rotary position embedding](#). *Neurocomputing*, 568:127063.
- Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. [Sequence to sequence learning with neural networks](#). In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. [Llama: Open and efficient foundation language models](#). *arXiv preprint arXiv:2302.13971*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shrutu Bhosale, et al. 2023b. [Llama 2: Open foundation and fine-tuned chat models](#). *arXiv preprint arXiv:2307.09288*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Bin Wang, Zhengyuan Liu, and Nancy Chen. 2023. [Instructive dialogue summarization with query aggregations](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7630–7653, Singapore. Association for Computational Linguistics.
- Xing Wang, Zhaopeng Tu, Longyue Wang, and Shuming Shi. 2019a. [Self-attention with structural position representations](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1403–1409. Association for Computational Linguistics.
- Yaoshian Wang, Hung-Yi Lee, and Yun-Nung Chen. 2019b. [Tree transformer: Integrating tree structures into self-attention](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1061–1070. Association for Computational Linguistics.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.
- BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Lucicioni, François Yvon, et al. 2022. [Bloom: A 176b-parameter open-access multilingual language model](#). *arXiv preprint arXiv:2211.05100*.
- Qitian Wu, Wentao Zhao, Zenan Li, David Wipf, and Junchi Yan. 2022. [Nodeformer: A scalable graph structure learning transformer for node classification](#). In *Advances in Neural Information Processing Systems*.
- Qitian Wu, Wentao Zhao, Chenxiao Yang, Hengrui Zhang, Fan Nie, Haitian Jiang, Yatao Bian, and Junchi Yan. 2023. [Simplifying and empowering transformers for large-graph representations](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Wei Wu, Houfeng Wang, Tianyu Liu, and Shuming Ma. 2018. [Phrase-level self-attention networks for universal sentence encoding](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3729–3738. Association for Computational Linguistics.
- Wen Xiao, Patrick Huber, and Giuseppe Carenini. 2020. [Do we really need that many parameters in transformer for extractive summarization? discourse can help !](#) In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 124–134. Association for Computational Linguistics.
- Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, et al. 2023. [Effective long-context scaling of foundation models](#). *arXiv preprint arXiv:2309.16039*.

- Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. [Discourse-aware neural extractive text summarization](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5021–5031. Association for Computational Linguistics.
- Yiheng Xu, Hongjin Su, Chen Xing, Boyu Mi, Qian Liu, Weijia Shi, Binyuan Hui, Fan Zhou, Yitao Liu, Tianbao Xie, et al. 2023. Lemur: Harmonizing natural language and code for language agents. *arXiv preprint arXiv:2310.06830*.
- Bruce Yu, Jianlong Chang, Lingbo Liu, Qi Tian, and Chang Wen Chen. 2023. [Towards a unified view on visual parameter-efficient transfer learning](#).
- Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. [PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization](#). In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.
- Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.
- Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2022. Dialoglm: Pre-trained model for long dialogue understanding and summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11765–11773.
- Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. 2023. [Least-to-most prompting enables complex reasoning in large language models](#). In *The Eleventh International Conference on Learning Representations*.

CS-Sum: A Benchmark for Code-Switching Dialogue Summarization and the Limits of Large Language Models

Sathya Krishnan Suresh, Tanmay Surana, Lim Zhi Hao, Eng Siong Chng

Nanyang Technological University

Correspondence: sathyakr001@e.ntu.edu.sg

Abstract

Code-switching (CS) poses a significant challenge for Large Language Models (LLMs), yet its comprehensibility remains underexplored in LLMs. We introduce **CS-Sum**, to evaluate the comprehensibility of CS by the LLMs through CS dialogue to English summarization. CS-Sum is the first benchmark for CS dialogue summarization across Mandarin-English (EN-ZH), Tamil-English (EN-TA), and Malay-English (EN-MS), with 900-1300 human-annotated dialogues per language pair. Evaluating ten LLMs, including open and closed-source models, we analyze performance across few-shot, translate-summarize, and fine-tuning (LoRA, QLoRA on synthetic data) approaches. Our findings show that though the scores on automated metrics are high, LLMs make subtle mistakes that alter the complete meaning of the dialogue. To this end, we introduce 3 most common type of errors that LLMs make when handling CS input. Error rates vary across CS pairs and LLMs, with some LLMs showing more frequent errors on certain language pairs, underscoring the need for specialized training on code-switched data.

1 Introduction

Code-switching (CS) is the practice of alternating between two or more languages within a single conversation or utterance. Bilingual and multilingual speakers frequently engage in CS. This form of communication has also become prevalent in social media and is the norm in multilingual societies. Large Language Models (LLMs) (Hurst et al., 2024; Team et al., 2023; Dubey et al., 2024), are trained on huge volumes of data majority of which is written in English. Despite the prevalence of CS in real-world communication, existing benchmarks (Huzaifah et al., 2024; Zhang et al., 2023; Yong et al., 2023) fail to assess how well LLMs process CS dialogues. Improved CS comprehension would enable LLMs to interpret multilingual

CS-Dialogue Adele: Naa innum antha song ah ketuttu irukan. Kode: Ethu Adele: Nee recommended pannathu Kode: Ama athu nalla irukkum Adele: Ama nalla irukku Kode: Atha ennoda music library la 3 years ah vechirkan Adele: Omg athu avlo old ah? Kode: hmm Adele: Athey singer kitta irunthu vera ethavthu suggest panriya Kode: Naa unakku whole album ah USB la kudukkren Adele: Eppo nee en veetukku varuva? Kode: Naa ippo free ah irukken Adele: Apo naa wait panren Kode: Varen.
EN-Summary Adele is still listening to the song which Kode recommended. Kode will come to Adele's home today and will give her the whole album of the same singer.

Figure 1: An instance from the benchmark

prompts more effectively, enhancing accessibility for CS speakers (Bawa et al., 2020).

Summarization is an ideal task to evaluate CS understanding, as it requires both grasping the dialogue’s main idea and generating a concise, coherent summary across languages. The only existing CS dialogue summarization dataset, Gupshup (Mehnaz et al., 2021), covers only one language pair (English-Hindi). Other CS benchmarks like LinCE (Aguilar et al., 2020) and GLUECos (Khanuja et al., 2020) focus on tasks such as language identification and POS-tagging, which assess word- or sentence-level understanding, whereas summarization requires discourse-level comprehension, a much harder test of CS ability.

To truly evaluate the ability of LLMs on CS, we propose the **CS-Sum** benchmark, which contains 900-1300 CS dialogue-summary pairs for three language pairs: Mandarin-English (EN-ZH), Tamil-English (EN-TA) and Malay-English (EN-MS). The CS dialogues were created by native

speakers of the respective languages. Instead of generating a dialogue from scratch, the speakers were asked to translate the English dialogues in test sets of DialogSum (Chen et al., 2021) and SAM-Sum (Gliwa et al., 2019) - two popular English dialogue summarization datasets - to CS dialogues in their respective languages.

We evaluate 9 open source and 1 proprietary LLMs under few-shot, translate-summarize, LoRA (Hu et al., 2022) and QLoRA (Detmeters et al., 2023) on synthetic data and provide a detailed analysis on the struggles of current SOTA LLMs($\leq 9B$). For LoRA and QLoRA fine-tuning, the training data was generated using Gemini-2-flash (Team et al., 2023), as it is a large-scale LLM with a free-to-use API, making it a practical choice for our experiments.

Our qualitative analysis of the summaries generated by LLMs showed that traditional summarization metrics like ROUGE (Lin, 2004), BERTScore (Zhang et al., 2019) are not sufficient for measuring the quality of the summaries since the LLMs are capable of generating summaries that contain subtle errors that change the entire meaning of the CS dialogues. To this end, we identify the three most common errors (CSL, MST, SMA Section 5) that the LLMs make when summarizing CS dialogues and suggest an LLM-driven approach to analyse their summaries with respect to the errors.

Our main contributions are:

- We introduce CS-Sum, the first CS dialogue summarization benchmark for EN-ZH, EN-TA and EN-MS.
- We conduct an in-depth evaluation of state-of-the-art LLMs, identifying critical failure patterns in their ability to process and summarize CS dialogues.
- We release CS-Sum¹, along with the full synthetic training dataset.

2 Related Work

While LLMs have demonstrated strong performance in many multilingual tasks, they still face significant challenges when dealing with CS. Studies (Zhang et al., 2023; Yong et al., 2023) have shown that LLMs, including GPT-4 and GPT-3.5, struggle with generating code-switched text, often

resulting in language collapse, where the model fails to mix languages properly, and in task-specific failures such as poor performance in summarization and machine translation for CS data. These issues arise because LLMs are typically trained on monolingual data and lack the capacity to handle the intricacies of CS discourse. Additionally (Huzaifah et al., 2024), their performance is inconsistent, particularly when translating low-resource language pairs, with models showing better results for high-resource languages but underperforming when confronted with languages that are under-represented in training datasets. These findings highlight a big problem as people prefer to interact with LLMs in code-switch (Bawa et al., 2020).

Recent benchmarks such as LinCE (Aguilar et al., 2020) and GLUECoS (Khanuja et al., 2020) have been instrumental in advancing the understanding of CS phenomena by focusing on tasks like LID, NER, and POS tagging. While these datasets have provided valuable insights, it’s important to recognize that LLMs continue to encounter challenges when dealing with the complexities inherent in code-switched discourse. To truly evaluate the CS comprehension of LLMs, other benchmarks are necessary—ones that go beyond word-level tasks and assess deeper linguistic understanding. The first step in this direction is CS-Sum, a benchmark that evaluates LLMs on a more complex task, such as summarization, which requires a nuanced comprehension of code-switched text.

3 CS-Sum Benchmark

3.1 Dataset Construction

The goal of building the CS-Sum benchmark, was to address the lack of a benchmark that evaluates the CS comprehension of LLMs. We selected Mandarin (zh), Bahasa Melayu (ms), and Tamil (ta) because these languages have large bilingual speaker communities that frequently code-switch with English. To efficiently create a diverse CS benchmark within resource constraints, we translated a subset of the combined test sets from DialogSum and SAMSum. We recruited 7 native speakers for each language to translate the English dialogues to CS dialogues in their respective languages. The translators were native speakers, all university students (bachelor’s or master’s level). They were instructed to translate the dialogues depending on how they would have conversed with their peers in CS. The entire translation process lasted for about 5 months.

¹<https://huggingface.co/datasets/SkAndM1/cs-sum>

Language	M-Index	I-Index	Burstiness	Span Entropy	Memory
EN-ZH	0.40	0.36	-0.79	0.13	-0.08
EN-MS	0.41	0.38	-0.65	0.64	-0.24
EN-TA	0.42	0.40	-0.68	0.58	-0.22

Table 1: CS metrics measured on the CS-Sum dataset

	EN-ZH	EN-TA	EN-MS
Number of instances	1320	1000	918
EN utterances	1140	344	517
Lang utterances	2079	983	265
EN as matrix language	4699	2340	2984
Lang as matrix language	4779	6790	1919
Avg. monolingual utterance length	2.344	3.616	5.515
Avg. CS utterance length	5.816	9.642	13.885

Table 2: CS statistics of CS-Sum. ‘Lang’ refers to the language other than English

3.2 Corpus Overview and Analysis

Table 2 presents key statistics of the CS-Sum benchmark across three language pairs. We observe that non-English utterances outnumber English ones in EN-ZH and EN-TA, while the reverse is true for EN-MS, suggesting stronger local language dominance in the former two. The matrix language distribution further highlights this variation: Tamil dominates as the matrix language in EN-TA, whereas EN-ZH and EN-MS are more balanced. This indicates differing CS behaviors—embedding in EN-TA versus alternation in EN-ZH and EN-MS.

Utterance lengths also vary significantly. EN-MS exhibits the longest average CS utterances (13.88 tokens), compared to 5.81 in EN-ZH, suggesting more complex span-level dependencies in Malay-English dialogues. The longer monolingual spans and varied matrix language roles imply that LLMs must manage different CS dynamics across language pairs.

3.3 Quantifying Code-Switching

We quantify the structural properties of code-switching in CS-Sum using five established metrics (Guzmán et al., 2017), reported in Table 1. The **M-Index** measures the distributional balance between languages, with values near 1 indicating equal usage. The **I-Index** captures the token-level likelihood of language alternation. **Burstiness** characterizes the temporal irregularity of switching, where negative values imply regular alternation and positive values indicate clustering. **Span Entropy** quantifies the unpredictability in monolin-

CS pair	KL Div.	JS Div.
EN-ZH	2.4786	0.4918
EN-MS	0.5481	0.1454
EN-TA	0.5089	0.1017

Table 3: Divergence between CS-Sum and CS-Sum-Syn

gual segment lengths, while **Memory** reflects the autocorrelation of consecutive spans—positive values suggest persistence, negative values suggest alternation.

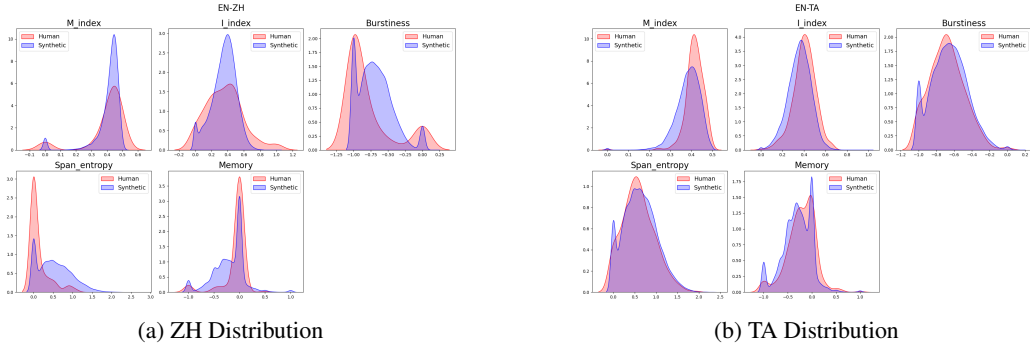
Across the three CS pairs, CS-Sum exhibits moderate switching behavior. The M-Index values (0.40–0.42) suggest relatively balanced bilingual distributions, and I-Index scores (0.36–0.40) indicate frequent intra-sentential switching, most prominently in EN-TA. Burstiness values are consistently negative, reflecting structured rather than chaotic switching. Span Entropy is lowest for EN-ZH (0.13), indicating predictable switch boundaries, while EN-TA (0.58) and EN-MS (0.64) exhibit higher variability. Memory values are near zero or negative across all pairs, implying weak temporal consistency in span lengths, and highlighting the need for LLMs to dynamically track and adapt to switching boundaries during summarization.

3.4 Gemini-Generated Training Data

To support fine-tuning for evaluating LLMs on CS-Sum, we construct a synthetic training set, **CS-Sum-Syn**, by converting 19,014 English dialogue-summary pairs from DialogSum and SAMSum into code-switched dialogues using Gemini-2² (Team et al., 2023). This augmentation is not a contribution in itself, but a means to enable parameter-efficient adaptation (e.g., LoRA, QLoRA). To validate the distributional alignment of synthetic and human CS data, we compute CS metrics across both sets and visualize them for each language pair in Figures 2a, 2b, and 2c.

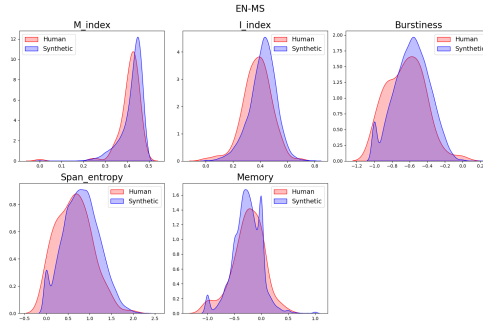
We find that Gemini-generated data broadly approximates human CS patterns for EN-TA and EN-

²<https://ai.google.dev/>, gemini-2-flash-exp



(a) ZH Distribution

(b) TA Distribution



(c) MS Distribution

MS. In EN-TA, distributions over I-Index, burstiness, and span entropy align closely, suggesting that Gemini captures intra-sentential switching well. EN-MS also shows good overlap, though with slightly more bursty and irregular switching in the synthetic set. In contrast, EN-ZH exhibits substantial divergence: synthetic dialogues show flatter memory and burstiness curves, as well as a long-tailed span entropy distribution, indicating reduced switching regularity and predictability. This mismatch is quantified in Table 3, where the KL and JS divergence for EN-ZH are $4\text{--}5\times$ higher than for the other pairs. These discrepancies may affect the quality of fine-tuning and are revisited in Section 5 during error diagnosis.

4 LLM performance on CS-Sum

In this section, we present the performance of ten LLMs—nine open-source³ and one proprietary—on CS-Sum across four settings: Few-Shot, Translate-Summarize, LoRA, and QLoRA. In LoRA and QLoRA, open-source models are fine-tuned on CS-Sum-Syn. The evaluated open-source LLMs include LLAMA-3-8B (Dubey et al., 2024), MISTRAL-7B (Jiang et al., 2023), MINISTRAL-8B⁴, GEMMA-2-2B and GEMMA-2-9B (Team

et al., 2024), QWEN2.5-2B and QWEN2.5-7B (Yang et al., 2024), SEA-LION-GEMMA-2-9B⁵, and SEALLM-7B (Nguyen et al., 2024). The proprietary model used is GPT-4o (Hurst et al., 2024). These models were selected based on their multilingual pretraining exposure and computational feasibility, with sizes ranging from 2B to 9B parameters. Models like LLAMA-3-8B, MISTRAL-7B, and GEMMA-2-9B have shown strong generalization across NLP tasks, while QWEN2.5-7B, SEA-LION-GEMMA-2-9B, and SEALLM-7B were chosen for their targeted multilingual capabilities, especially in languages like Bahasa Melayu, Tamil, and Mandarin. GPT-4o serves as a strong proprietary baseline due to its SOTA performance in cross-lingual comprehension.

By evaluating these models across different parameter scales and varying degrees of multilingual exposure, we investigate how model size and pretraining diversity affect performance. LoRA and QLoRA fine-tuning on CS-Sum-Syn result in higher scores on automated metrics, suggesting improved alignment with reference summaries. However, it remains unclear whether these gains reflect actual comprehension of code-switched dialogue or simply the ability to replicate the distributional patterns of Gemini-generated synthetic data. To explore this, we conduct an in-depth error anal-

³Instruction-tuned and latest versions

⁴<https://huggingface.co/mistralai/Ministral-8B-Instruct-2410>

⁵<https://sea-lion.ai/>

Model	Lang	ROUGE-L	BERTScore	SBERT-COSINE	JACCARD	METEOR
Gemma-2-2B	EN-ZH	0.2330 / -4	0.8876 / -0.35	0.7268 / -1.40	0.1658 / -7.64	0.2699 / -0.9
	EN-TA	0.2202 / -7	0.8812 / -0.38	0.6755 / -0.85	0.1523 / -9.19	0.2453 / -5.23
	EN-MS	0.2421 / -7.18	0.8898 / -0.98	0.7352 / -4.35	0.1739 / -13.52	0.2902 / -12.43
Qwen2.5-3B	EN-ZH	0.2393 / 9.88	0.8824 / 0.93	0.7143 / 1.70	0.1643 / 5.82	0.2974 / 2.05
	EN-TA	0.2227 / -1.46	0.8805 / 0.44	0.6880 / -2.35	0.1468 / -0.13	0.2591 / -4.82
	EN-MS	0.2408 / 6.21	0.8868 / 0.60	0.7308 / -2.25	0.1694 / 1.14	0.3100 / -4.57
Qwen2.5-7B	EN-ZH	0.2482 / 27.10	0.8106 / 11.20	0.4927 / 54.12	0.0783 / 162.62	0.1118 / 188.89
	EN-TA	0.2800 / -5.11	0.8937 / -0.55	0.7081 / -2.14	0.1837 / -9.62	0.2723 / -9.20
	EN-MS	0.3117 / -1.65	0.9041 / -0.46	0.7520 / -0.67	0.2122 / -8.37	0.3372 / -10.26
SEALLM-7B	EN-ZH	0.2115 / 19.25	0.8122 / 9.15	0.4836 / 45.48	0.0834 / 89.27	0.1297 / 104.29
	EN-TA	0.2270 / 0.99	0.8686 / 1.21	0.6470 / 2.53	0.1336 / 3.44	0.2287 / 1.55
	EN-MS	0.2575 / 8.41	0.8801 / 1.28	0.6956 / 4.91	0.1633 / 9.04	0.2800 / 5.95
Mistral-7B	EN-ZH	0.2568 / -3.10	0.8902 / -0.42	0.7208 / -0.34	0.1754 / -8.01	0.3010 / -7.82
	EN-TA	0.2198 / -2.15	0.8792 / 0.04	0.6847 / -0.95	0.1470 / -6.91	0.2657 / -10.45
	EN-MS	0.2421 / 1.59	0.8897 / -0.46	0.7293 / -2.15	0.1796 / -11.49	0.3084 / -8.40
Ministral-8B	EN-ZH	0.2560 / 4.23	0.8875 / 0.52	0.7037 / 1.06	0.1588 / 6.13	0.2517 / 1.15
	EN-TA	0.2468 / -5.61	0.8858 / -0.18	0.6752 / -2.95	0.1561 / -3.66	0.2387 / -8.53
	EN-MS	0.2700 / 6.97	0.8917 / 0.37	0.7262 / 0.03	0.1794 / 2.93	0.2843 / -0.09
LLaMA-3-8B	EN-ZH	0.2868 / -2.29	0.8845 / 0.76	0.7090 / 4.09	0.1918 / -6.99	0.2916 / 0.88
	EN-TA	0.2453 / 0.47	0.8769 / 0.99	0.6707 / 3.28	0.1598 / 1.07	0.2570 / -2.13
	EN-MS	0.2599 / 8.06	0.8724 / 2.04	0.6538 / 11.21	0.1804 / 0.02	0.2833 / 5.50
Gemma-2-9B	EN-ZH	0.2995 / -5.62	0.8987 / -0.27	0.7603 / -0.91	0.2054 / -5.83	0.3246 / -3.06
	EN-TA	0.2761 / -3.85	0.8911 / 0.57	0.7203 / 4.42	0.1832 / 4.63	0.3107 / 1.25
	EN-MS	0.3009 / -3.00	0.8975 / 0.01	0.7573 / -0.94	0.2096 / -3.96	0.3279 / -0.07
SEA-LION-Gemma-2-9B	EN-ZH	0.2799 / -3.96	0.8968 / -0.34	0.7542 / -0.41	0.1820 / -4.29	0.2944 / -4.46
	EN-TA	0.2703 / 0.26	0.8898 / 0.48	0.7138 / 3.79	0.1727 / -0.12	0.2884 / -2.38
	EN-MS	0.2860 / -1.56	0.8963 / -0.12	0.7404 / 1.21	0.1837 / -4.16	0.2983 / -4.75
GPT-4o	EN-ZH	0.2965 / -1.03	0.8816 / 2.21	0.7082 / 9.63	0.1792 / 11.66	0.2853 / 16.52
	EN-TA	0.3157 / -4.92	0.8981 / 0.39	0.7570 / 2.19	0.2041 / -0.55	0.3419 / 0.75
	EN-MS	0.3126 / -1.70	0.9037 / -0.15	0.7750 / 0.12	0.2045 / 0.33	0.3365 / 1.09

Table 4: Few-Shot Performance/Translate-Summarize % improvement on CS-Sum

ysis in Section 5, which reveals that fine-tuning on synthetic data does not improve—and often degrades—the model’s ability to accurately interpret and summarize code-switched content.

4.1 Result Analysis

In this subsection, we present the results of the LLMs’ performance on the CS-Sum in Few-Shot, Translate-Summarize, LoRA and QLoRA settings. Appendix B provides an explanation on the different settings. We use standard summarization metrics like ROUGE (Lin, 2004), BERTScore (Zhang et al., 2019), SBERT-Cosine (Reimers and Gurevych, 2019), Jaccard and METEOR (Banerjee and Lavie, 2005). This comprehensive set of metrics evaluate both word-level matching and semantic similarity.

The few-shot performance of LLMs across the three CS language pairs is reported in Table 4. While semantic similarity metrics such as BERTScore and SBERT-Cosine report high val-

ues, these can be misleading. As illustrated in Figure 3, the summaries often fail to accurately capture the core content or intent of the original dialogue, despite appearing semantically plausible. This reflects a common failure mode of LLMs in CS summarization: generating fluent, high-overlap summaries that omit or distort key factual elements. In this example, Gemma-2-9B misrepresents who confirms the location of the conference and fabricates speaker opinions, despite achieving a high semantic similarity score. ROUGE-L scores further corroborate this gap, with LLMs scoring nearly 50% lower than the best-performing models on the monolingual DialogSum⁶ and SAMSum⁷ benchmarks.

From Table 4, we observe that GPT-4o is the best-performing model overall. However, due to

⁶<https://paperswithcode.com/sota/text-summarization-on-dialogsum>

⁷<https://paperswithcode.com/sota/text-summarization-on-samsum-corpus>

the lack of transparency in its training and architecture, we cannot provide deeper analysis on its performance characteristics. Among open-source models, Gemma-2-9B consistently outperforms others across most metrics and language pairs. This aligns with findings from the Gemma-2 paper (Team et al., 2024), which attributes its strong multilingual performance to extensive exposure to diverse languages and an optimized tokenizer. Interestingly, SEA-Lion-Gemma-2-9B lags slightly behind, despite its regional specialization in Southeast Asian languages, suggesting that broader multilingual exposure may be more beneficial than region-specific pretraining for CS comprehension.

One notable observation is the underperformance of most models on the EN-ZH pair, which is surprising given that all models were exposed to Mandarin during pretraining. Manual inspection reveals that some LLMs generate summaries partially or entirely in Chinese, despite being explicitly prompted to produce English outputs. This mismatch leads to lower scores on automated metrics, particularly those that penalize non-English outputs. However, as discussed in our fine-grained evaluation (Section 5), these summaries often contain fewer semantic or structural errors than those in EN-TA or EN-MS, indicating that the low metric scores may not reflect actual comprehension quality.

We also find that smaller models, such as Gemma-2-2B and Qwen2.5-3B, perform significantly worse than their larger counterparts, reinforcing the impact of model scale on CS summarization. Surprisingly, SEALLM-7B performs comparably to these smaller models, despite being explicitly trained on Southeast Asian languages. Qualitative analysis of its outputs reveals several failure modes in the few-shot setting: **(a)** copying utterances directly from the dialogue, **(b)** generating descriptive paraphrases rather than true summaries, and **(c)** producing outputs in the wrong language. These behaviors suggest that SEALLM struggles with in-context learning and generalization, even within its intended linguistic domain.

Translate-Summarize % Improvement Table 4 reports percentage improvements under the Translate-Summarize setting. Contrary to expectations, most models show little to no improvement, and in many cases, performance declines across all metrics. This suggests that translation introduces additional challenges, likely due to LLMs’ inability

LLM	Human
Emily is asking where the ASEES conference is held. Chloe thinks it's probably in San Francisco or in Hawaii, but James thinks it's held in California. Chloe thinks that's sad.	James informs that the next ASEES conference will be held in San Francisco. Chloe was hoping for Hawaii. James doesn't think there will be a conference in Hawaii soon due to the high cost of travel.

Figure 3: Wrong summary with high BERTScore 0.903

to preserve the structural and discourse-level nuances of code-switched input. Notably, Qwen2.5-7B and SEALLM-7B improve significantly for EN-ZH (27.1% and 19.25% in ROUGE-L), reflecting benefits from Mandarin or SEA-specific pretraining. However, similar gains are absent for EN-TA and EN-MS, indicating that CS translation remains unreliable outside of high-resource language pairs.

LoRA and QLoRA Results We fine-tune LLMs on the synthetic CS-Sum-Syn dataset using LoRA and QLoRA to evaluate their ability to handle CS summarization beyond just metric gains. As shown in Table 5⁸, EN-ZH sees the largest improvements (e.g., SEALLM-7B and Qwen2.5-7B exceed 200% in METEOR), EN-TA shows consistent gains, while EN-MS exhibits mixed results across metrics.

These improvements are surprising given the distributional mismatch between CS-Sum and CS-Sum-Syn (Figure 2a, Table 3). A qualitative review reveals that models often produce summaries with subtle semantic errors, indicating they may have learned surface-level summary structure rather than true code-switching comprehension. This motivates the deeper analysis in Section 5.

5 Error Analysis

In this section, we present an in-depth analysis of the common error types exhibited by LLMs when summarizing CS dialogues. Our qualitative analysis across the EN-ZH, EN-TA, and EN-MS language pairs revealed three major failure modes:

- **Code-Switching Loss (CSL):** The summary primarily utilizes the English parts of the dialogue, ignoring or missing critical information from non-English segments.
- **Meaning Shift from Poor Translation**

⁸Due to space constraints, rest of the scores are presented in Table 8

Model	Lang	ROUGE-L	BERTScore	SBERT-COSINE	JACCARD	METEOR
SEALLM-7B	EN-ZH	50.45 / 51.94	11.09 / 11.21	60.22 / 60.69	170.47 / 160.94	163.13 / 142.51
	EN-TA	47.99 / 43.82	4.49 / 4.52	19.84 / 19.86	75.18 / 69.14	47.14 / 35.42
	EN-MS	34.20 / 17.03	3.29 / 2.53	12.01 / 27.78	47.49 / 4.22	13.08 / 27.78
LLaMA-3-8B	EN-ZH	23.51 / 25.21	3.14 / 3.12	12.88 / 12.79	29.85 / 34.26	19.70 / 29.68
	EN-TA	51.45 / 47.07	4.32 / 4.00	18.91 / 18.28	66.19 / 60.27	39.72 / 42.35
	EN-MS	46.81 / 21.73	4.85 / 3.27	22.19 / 16.94	50.56 / 23.01	39.34 / 17.83
Gemma-2-9B	EN-ZH	-1.45 / 10.66	0.12 / 0.82	4.10 / 2.32	8.56 / 11.61	1.68 / 7.95
	EN-TA	6.54 / 19.37	0.63 / 1.56	7.14 / 9.12	18.12 / 27.89	7.28 / 13.01
	EN-MS	7.81 / -6.27	0.71 / -0.61	3.60 / -1.41	8.48 / -3.45	5.31 / -7.06
SEA-LION-Gemma-2-9B	EN-ZH	-22.93 / -6.64	-3.10 / -0.71	-3.27 / 1.73	-4.13 / 8.67	6.29 / 14.39
	EN-TA	-21.48 / 3.59	-2.61 / 0.24	1.74 / 7.08	-3.49 / 20.38	7.38 / 17.21
	EN-MS	-21.98 / -8.08	-2.95 / -0.99	-3.53 / -0.77	-8.16 / 3.18	9.09 / 3.03

Table 5: LoRA / QLoRA % Improvement over Few-Shot on CS-Sum

(MST): The model misunderstands the code-switched segments, resulting in summaries that deviate from the dialogue’s true meaning.

- **Speaker Misattribution (SMA):** Summaries incorrectly assign statements to speakers, thus distorting intended meanings or misrepresenting participants’ views.

Figure 4 illustrates CSL, demonstrating how the summary incorrectly emphasizes English segments and misinterprets speaker intentions. Further examples across categories are provided in Tables 9, 10, and 11. Errors were identified automatically via GPT-4o, detailed further in Appendix C.

To quantify these errors and compare model behaviors, we report the percentage of summaries exhibiting each error type across 9 LLMs under both Few-Shot and LoRA settings (Table 6). The analysis reveals the following:

CSL persists as the dominant failure mode in Few-Shot evaluation. Across *all* 9 models and the three language pairs, CSL exceeds 50% in the Few-Shot setting (Table 6). Even the best performer, SEA-LION-GEMMA-2-9B, attains a minimum of 53.73% CSL on **EN-ZH**, confirming that LLMs systematically ignore non-English spans when summarizing code-switched dialogues, regardless of architecture or pre-training recipe.

Scaling within a model family does not reliably reduce errors. Comparing parameter-matched variants shows no monotonic gains: GEMMA-2-9B lowers CSL on **EN-ZH** (77.61→59.70) yet *raises* Speaker Misattribution (SMA) on **EN-TA** by 7 pp over its 2B sibling, while QWEN2.5-7B improves CSL on **EN-MS** but leaves MST virtually unchanged. These inconsistencies indicate that

sheer scale is not a substitute for task-specific multilingual evaluation.

EN-TA exhibits consistently higher error rates across models. All evaluated models show their highest rates of CSL and SMA on EN-TA dialogues (e.g., QWEN2.5-7B: 91.81% CSL, 57.31% SMA). This suggests that certain language-specific properties, such as morphological complexity and syntactic divergence from English, may contribute to increased summarization difficulty in this setting.

SEA-oriented models excel in error analysis despite mediocre automatic scores. Regional models such as SEALLM-7B and SEA-LION-GEMMA-2-9B achieve the lowest CSL and SMA within their parameter class (e.g., 53.73% CSL on **EN-ZH**), yet trail larger English-specific models on ROUGE and BERTScore. This divergence reinforces that conventional metrics do not capture CS comprehension, underscoring the need for better metrics.

Synthetic fine-tuning amplifies errors under distribution shift. Fine-tuning on the Gemini-generated CS-Sum-Syn corpus degrades performance when the training distribution diverges from CS-Sum, most notably for **EN-ZH**, whose KL divergence is 2.48 versus ≤ 0.55 for the other pairs (Table 3). After adaptation, SEA-LION-GEMMA-2-9B’s CSL jumps from 53.73% to 83.94% and MST from 11.19% to 76.64% (Table 6). These results illustrate that even synthetic data generated using a high-performing LLM does not improve the model’s ability to comprehend code-switched input, and the failure to produce data that enhances CS understanding also underscores the limitations of current LLMs comprehensibility of CS.

Model	Lang	CSL		MST		SMA	
		Few-Shot	LoRA	Few-Shot	LoRA	Few-Shot	LoRA
Gemma-2-2B	EN-ZH	77.61	88.32	60.45	90.51	40.30	77.37
	EN-TA	94.15	92.69	86.26	90.06	69.88	74.85
	EN-MS	77.19	94.87	62.28	84.62	35.09	67.52
Qwen2.5-3B	EN-ZH	55.22	83.94	52.24	69.34	36.57	39.42
	EN-TA	91.81	88.01	89.47	75.15	76.02	50.58
	EN-MS	60.53	84.62	49.12	60.68	31.58	40.17
Qwen2.5-7B	EN-ZH	71.64	81.75	32.84	48.91	9.70	26.28
	EN-TA	91.81	79.23	76.32	58.64	57.31	32.58
	EN-MS	78.95	82.05	35.96	42.74	16.67	24.79
SEALLM-7B	EN-ZH	64.18	84.67	41.79	70.80	21.64	33.58
	EN-TA	91.23	93.57	81.87	71.64	63.16	47.95
	EN-MS	68.42	85.47	45.61	69.23	23.68	40.17
Mistral-7B	EN-ZH	55.97	83.21	37.31	69.34	22.39	40.88
	EN-TA	88.30	89.47	81.58	79.82	66.37	53.22
	EN-MS	67.54	77.78	44.74	70.09	28.07	44.44
Ministral-8B	EN-ZH	76.12	85.40	52.24	74.45	31.34	48.91
	EN-TA	95.03	90.94	88.80	79.82	66.37	55.56
	EN-MS	78.95	95.73	49.12	82.91	20.18	57.26
LLaMA-3-8B	EN-ZH	74.63	71.17	35.82	45.07	16.41	21.35
	EN-TA	90.94	56.92	80.70	37.89	55.85	19.93
	EN-MS	73.68	62.91	40.35	33.38	14.04	17.33
Gemma-2-9B	EN-ZH	59.70	85.40	18.66	57.66	8.21	35.77
	EN-TA	71.35	82.75	35.09	67.84	22.22	46.20
	EN-MS	59.65	79.60	18.42	52.53	8.77	29.45
SEA-LION-Gemma-2-9B	EN-ZH	53.73	83.94	11.19	76.64	3.73	52.55
	EN-TA	60.53	84.50	27.78	76.61	17.54	49.71
	EN-MS	48.25	86.32	14.04	64.96	7.02	46.15

Table 6: Comparison of error rates (%) in Few-Shot vs. LoRA fine-tuned models across different CS language pairs.

<p>Dialogue: Lauren: Hi kau masih perlukan aku tak esok Pam: Yes please!! Lauren: Kau ada lagi rota? Pam: No, tapi Manager aku balik esok so she may do some more then. Aku akan call kau in the morning and let you know. Lauren: ok that's great Pam: Did you have a good holiday? Lauren: Yes, esok aku akan share ngan kau Pam: Look forward to it!</p> <p>Qwen2.5-7B summary: Pam needs Lauren's help tomorrow. Lauren will be contacted by Pam's manager in the morning to confirm the arrangements. Pam asks Lauren if her holiday was good, and Lauren confirms it was, and she will share the details with Pam tomorrow.</p>
--

Figure 4: CSL error example

6 Conclusion

In this paper, we introduced CS-Sum, the first CS dialogue-to-English summarization benchmark across three language pairs: EN-TA, EN-ZH, and EN-MS. Our experiments revealed that LLMs perform inconsistently across CS pairs, with some open-source models outperforming SEA-specific models, highlighting that CS data must be present in pretraining, not just added through fine-tuning. Furthermore, our analysis identified three major error types—CSL, MST, and SMA—persisting across all models, demonstrating that current LLMs lack intrinsic CS comprehension. The quantitative results reinforce the severity of these failures, showing that fine-tuning alone does not fully resolve CS-specific challenges. To drive further research in multilingual NLP, we will release CS-Sum along with our codebase and evaluation framework.

7 Limitations

While this work provides a comprehensive evaluation of CS dialogue summarization through the

CS-Sum benchmark, there are some limitations that should be addressed in future work.

- **Focus on Summarization:** Our experiments primarily focus on the task of summarization. However, code-switching also poses significant challenges for other long-context NLP tasks, such as machine translation, dialogue generation, and question answering. Extending the CS-Sum benchmark to these tasks would provide a broader understanding of LLMs' capabilities in handling CS data across different domains.
- **Finetuning with Synthetic Data:** The finetuning experiments in this paper are limited to synthetic code-switched data generated using Gemini-2. While this approach allowed us to assess the potential of fine-tuning for CS comprehension, it does not capture the complexities of real-world CS data.

References

- Gustavo Aguilar, Sudipta Kar, and Thamar Solorio. 2020. [LinCE: A centralized benchmark for linguistic code-switching evaluation](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1803–1813, Marseille, France. European Language Resources Association.
- Satanjeev Banerjee and Alon Lavie. 2005. [METEOR: An automatic metric for MT evaluation with improved correlation with human judgments](#). In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.
- Anshul Bawa, Pranav Khadpe, Pratik Joshi, Kalika Bali, and Monojit Choudhury. 2020. [Do multilingual users prefer chat-bots that code-mix? let's nudge and find out!](#) *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW1).
- Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. [DialogSum: A real-life scenario dialogue summarization dataset](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074, Online. Association for Computational Linguistics.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: efficient finetuning of quantized llms. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23*, Red Hook, NY, USA. Curran Associates Inc.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. [SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization](#). In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.
- Gualberto Guzmán, Joseph Ricard, Jacqueline Serigos, Barbara E. Bullock, and Almeida Jacqueline Toribio. 2017. [Metrics for modeling code-switching across corpora](#). In *Interspeech 2017*, pages 67–71.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Muhammad Huzaifah, Weihua Zheng, Nattapol Chanpaisit, and Kui Wu. 2024. [Evaluating code-switching translation with large language models](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6381–6394, Torino, Italia. ELRA and ICCL.
- Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. [Mistral 7b](#). *ArXiv*, abs/2310.06825.
- Simran Khanuja, Sandipan Dandapat, Anirudh Srivasan, Sunayana Sitaram, and Monojit Choudhury. 2020. [GLUECoS: An evaluation benchmark for code-switched NLP](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3575–3585, Online. Association for Computational Linguistics.
- 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.
- Laiba Mehnaz, Debanjan Mahata, Rakesh Gosangi, Uma Sushmitha Gunturi, Riya Jain, Gauri Gupta, Amardeep Kumar, Isabelle G. Lee, Anish Acharya, and Rajiv Ratn Shah. 2021. [GupShup: Summarizing](#)

- open-domain code-switched conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6177–6192, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Zhiqiang Hu, Chenhui Shen, Yew Ken Chia, Xingxuan Li, Jianyu Wang, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, and Lidong Bing. 2024. [SeaLLMs - large language models for Southeast Asia](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 294–304, Bangkok, Thailand. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.
- Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.
- Zheng Xin Yong, Ruochen Zhang, Jessica Forde, Skyler Wang, Arjun Subramonian, Holy Lovenia, Samuel Cahyawijaya, Genta Winata, Lintang Sutawika, Jan Christian Blaise Cruz, Yin Lin Tan, Long Phan, Long Phan, Rowena Garcia, Thamar Solorio, and Alham Fikri Aji. 2023. [Prompting multilingual large language models to generate code-mixed texts: The case of south East Asian languages](#). In *Proceedings of the 6th Workshop on Computational Approaches to Linguistic Code-Switching*, pages 43–63, Singapore. Association for Computational Linguistics.
- Ruochen Zhang, Samuel Cahyawijaya, Jan Christian Blaise Cruz, Genta Winata, and Alham Fikri Aji. 2023. [Multilingual large language models are not \(yet\) code-switchers](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12567–12582, Singapore. Association for Computational Linguistics.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Synthetic Data Generation Process

To create a synthetic CS-dialogue summarization dataset, we combined the training sets of DialogSum and SAMSum, resulting in 19,014 English dialogue-summary pairs. We used Gemini-2-flash to translate the English dialogues to code-switched dialogues in three CS pairs, as it is a large-scale SOTA LLM with a free-to-use API, making it a practical choice for our experiments.

We used the following prompt to convert English dialogues into CS dialogues. Each prompt was augmented with few-shot examples that were human generated.

Example Prompt for Malay-English Code-Switching

You are a {Malay/Tamil/Chinese} person in your 20s. You are recruited for translating English dialogues to English-{Malay/Tamil/Chinese} code-switched dialogues. The code-switched dialogues should follow the same structure as the English dialogue. This is for educational purposes, so DO NOT include swear words in your translation. Return just the code-switched dialogue.

B Experiment settings

B.1 Few-shot and Translate-Summarize

The Few-Shot and Translate-Summarize settings evaluate the inherent CS comprehension of LLMs. Evaluating the LLMs in this setting gives us an idea of their performance without any additional training.

Few-shot In the Few-Shot setup, we provide each LLM with three human-crafted examples containing code-switched dialogues and their corresponding summaries. The results from this setup, will quantify the inherent CS understanding of the LLMs for the three language pairs.

Translate-Summarize In the Translate-Summarize approach, each CS dialogue is first translated into English before being summarized. This serves as a baseline to determine whether the presence of code-switching affects summary quality. The translation step is performed by the same LLM, ensuring that the summarization model receives only monolingual input.

B.2 LoRA and QLoRA

In this subsection, we describe the fine-tuning process using *LoRA* and *QLoRA* to adapt open-source

LLMs for CS dialogue summarization. Given that full fine-tuning is computationally expensive, we use parameter-efficient fine-tuning (PEFT) techniques to assess whether task-specific adaptation improves CS comprehension. The models are fine-tuned on CS-Sum-Syn.

Model	r	α	LR
Ministral-8B	32	32	5e-5
LLaMA-3-8B	16	16	5e-5
Qwen2.5-3B	64	64	3e-5
Gemma-2-9B	32	32	5e-5
SEA-Lion-Gemma-9B	32	32	5e-5
Gemma-2-2B	8	8	5e-5
Qwen-2.5-7B	64	64	3e-5
Mistral-7B	32	32	5e-5
SEALLM-7B	64	64	3e-5

Table 7: Hyperparameters used for LoRA and QLoRA fine-tuning.

The fine-tuning process follows model-specific hyperparameter configurations, as shown in Table 7. The main hyperparameters include the rank parameter (r), scaling factor (α), and learning rate (lr). These configurations were chosen based on empirical observations to balance training stability and performance.

All the models were trained for 4 epochs with a batch size of 8 and gradient accumulation for 8 steps. We followed the cosine learning rate scheduling after warming up for the first 3% of steps. We used bf16 and gradient checkpointing for memory efficient training. All the LLMs were trained on an H100 GPU and the training lasted between 1 to 4 hours depending on the architecture and size of the LLM.

C Error Analysis Process

The summaries generated by the LLMs typically exhibit the three major errors listed in Section 5. In this section, we describe the process we followed to detect these errors in the generated summaries.

We used GPT-4o to analyse the summary given the CS-dialogue for different errors. For each error type, we used a separate prompt to guide the classification process. The LLM was asked to analyse the summary with respect to the particular error type before providing a verdict on presence of the error.

Below, we present the prompts used for each category.

Prompt for CSL

You will be given a code-switched (CS) dialogue between two speakers and its generated summary. Identify whether the summary has ignored or removed key information from the non-English parts of the dialogue.

The summary has an error if it meets one or more of the following criteria:

- Does not include important content from the CS part of the dialogue.
- The summary only relies on the English part of the dialogue.

Prompt for SMA

You will be given a code-switched (CS) dialogue and its generated summary. Analyze whether the summary has mixed up who said what.

The summary has an error if it meets one or more of the following criteria:

- Has swapped roles or perspectives between speakers.
- Statements are wrongly assigned.

Prompt for MST

You will be given a code-switched (CS) dialogue and its generated summary. Analyze whether the summary distorts the meaning of the original conversation due to incorrect understanding of the code-switching. The summary has an error if it meets one or more of the following criteria:

- The summary misinterprets the dialogue due to poor translation.
- Intent of the speakers has changed because of poor understanding of a CS phrase.
- Key points of the dialogue are misrepresented.

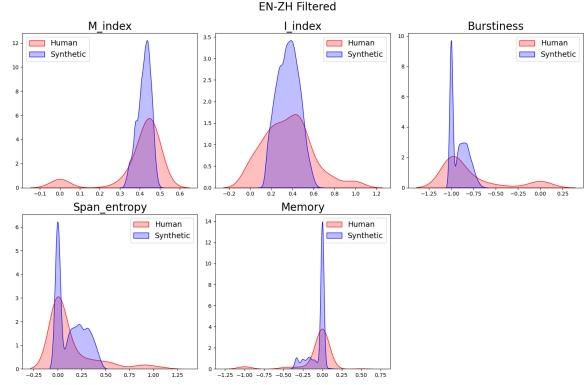


Figure 5: Distribution b/w filtered CS-Sum-Syn and CS-Sum for EN-ZH

D Training on filtered data

The difference in the distribution of CS-Sum and CS-Sum-Syn for EN-ZH (see Figure 2a) is much more than the differences in distribution for EN-MS and EN-TA. We hypothesize that this might be due to the tokenization strategy of the LLMs which is more robust for Roman script compared to Chinese script. The % improvement gains for EN-ZH after LoRA and QLoRA finetuning on EN-ZH is also less compared to the other two CS pairs (Table 5).

To check if filtering CS-Sum-Syn to match CS-Sum’s distribution would improve the finetuning results, we treat the CS metrics shown in Table 1 as a multivariate gaussian. We calculated the mean μ and the covariance matrix Σ of CS-Sum and used the Mahalanobis distance (eqn. 1) to calculate the distance between the instances in the training set and the test set.

$$D_M(\mathbf{x}) = \sqrt{(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)} \quad (1)$$

We filtered the training data aggressively by selecting only the top 20 percentile of the training instances with the shortest distances. The filtered dataset had 3801 instances. The resulting distribution is shown in Figure 5 which matches CS-Sum’s distribution better than the unfiltered one (see Figure 2a).

We finetuned (QLoRA) all the 9 open-source LLMs on the filtered data for 3 epochs (4 for the unfiltered one) with all the other parameters staying the same. Figure 6 shows the % improvement of LLMs finetuned on filtered data over the LLMs finetuned on the unfiltered one.

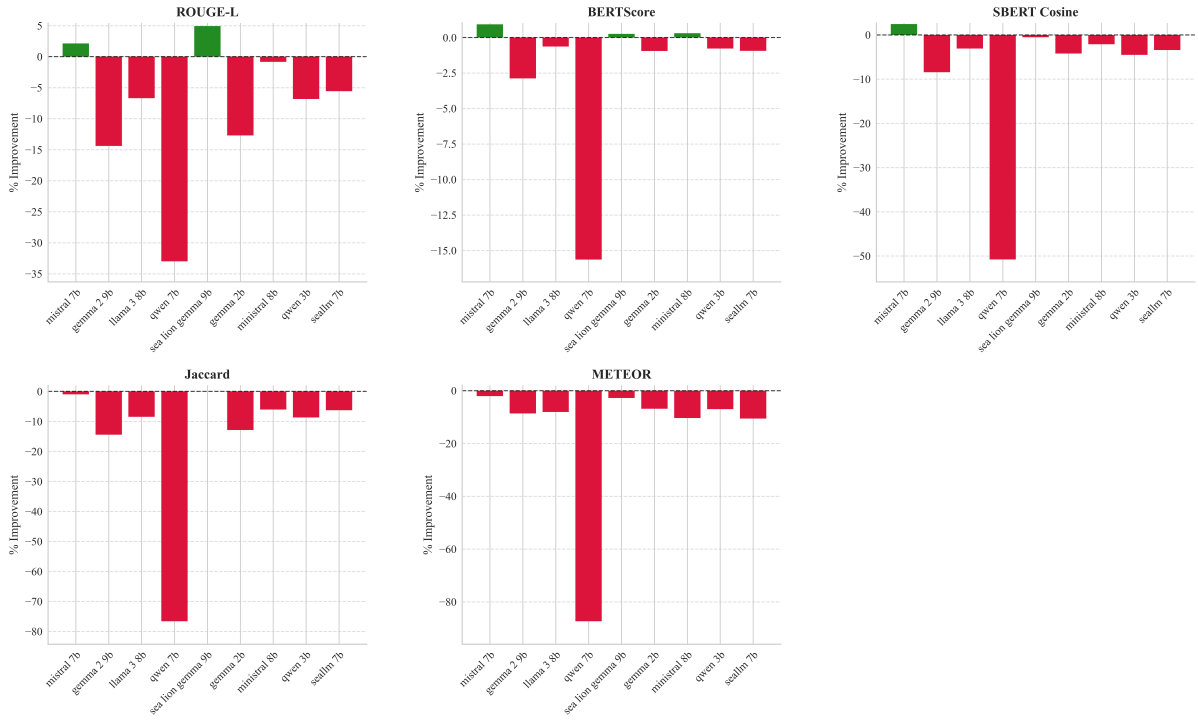


Figure 6: Filtered data % improvement over CS-Sum-Syn for EN-ZH

It can be seen that only 2 LLMs - Mistral 7B and SEA-Lion-Gemma-9B show actual improvement while other LLMs either don't improve much or have worse performances. Qwen2.5-7B suffers the worst when finetuned on filtered data. This suggests that filtering using Mahalanobis distance may only be beneficial for certain architectures while being detrimental to others. Further investigation is required to determine if alternative filtering thresholds or strategies could yield better generalization across LLMs.

Model	Lang	ROUGE-L	BERTScore	SBERT-COSINE	JACCARD	METEOR
Gemma-2-2B	EN-ZH	9.13 / 22.26	0.15 / 1.21	3.52 / 6.23	19.28 / 21.69	15.47 / 15.78
	EN-TA	21.52 / 23.29	1.62 / 1.79	12.10 / 11.86	23.88 / 29.04	19.16 / 20.98
	EN-MS	16.59 / 7.98	0.76 / 0.34	4.82 / 1.26	14.51 / 2.31	8.16 / -4.92
Qwen2.5-3B	EN-ZH	38.32 / 36.44	2.72 / 2.80	10.99 / 11.20	44.42 / 41.06	21.43 / 21.29
	EN-TA	39.08 / 35.50	2.73 / 2.55	12.30 / 11.96	47.58 / 45.49	30.90 / 33.84
	EN-MS	34.00 / 14.60	2.11 / 1.03	5.97 / 3.56	33.91 / 17.70	16.09 / 10.70
Qwen2.5-7B	EN-ZH	40.89 / 39.07	11.95 / 12.20	59.88 / 60.92	217.69 / 208.36	234.54 / 239.63
	EN-TA	22.86 / 20.85	1.83 / 1.72	11.50 / 11.10	32.72 / 30.17	35.66 / 36.26
	EN-MS	9.71 / 2.06	0.74 / -0.10	6.27 / 3.64	12.16 / 1.82	11.38 / 5.36
Mistral-7B	EN-ZH	17.91 / 2.18	0.73 / -0.80	7.78 / 2.43	28.10 / 15.19	20.10 / 9.41
	EN-TA	-1.79 / -3.62	-2.20 / -2.85	3.90 / 1.90	8.93 / 6.68	13.76 / 11.95
	EN-MS	-11.88 / 15.12	-2.93 / 0.56	-1.21 / 2.93	-8.56 / 8.96	3.83 / -2.09
Ministral-8B	EN-ZH	9.92 / 13.16	1.15 / 0.69	9.58 / 9.85	30.37 / 34.98	29.41 / 33.61
	EN-TA	18.38 / 20.88	1.41 / 1.41	13.73 / 14.13	34.19 / 36.83	35.43 / 39.83
	EN-MS	-24.12 / 7.51	-2.73 / 0.63	-1.64 / 4.51	-8.07 / 9.89	7.89 / 2.76

Table 8: LoRA / QLoRA % Improvement over Few-Shot on CS-Sum

EN Dialogue	CS Dialogue	LLM Summary	Error Explanation
<p>Matthew: I have a cold. Need y'all support. It's painful Lisa: Fight it. Drink orange juice. Make some soup Matthew: I'll try Lisa: I believe in u ...</p>	<p>Matthew: Enakku cold irukku. Unga ellaroda support um vennum. Athu romba painful Lisa: Atha fight pannu. Orange juice kudi. Soup pannu Matthew: Naa try panren Lisa: Naa unna believe panran ...</p>	<p>Matthew is feeling unwell due to a cold and an allergy. Lisa suggests that he drink orange juice and soup to feel better. Matthew tries it and feels a bit better.</p>	<p>Matthew says that he will try the pill but the summary mentions that he had tried it and feels better</p>
<p>Anna: is anyone going to pick Mark from the airport? Marcus: i could but when and where from? Anna: Sydney, Thursday at 3 Marcus: am or pm? :D Leslie: haha fortunately pm:D Marcus: hmm i have a meeting at 1. I don't think i can make it Leslie: well i guess it will take him some time after landing, reclaiming luggage etc Anna: yeah I reckon it's fine if you're there at 4 Marcus: oh well ok then ...</p>	<p>Anna: 有人去机场接Mark吗? Marcus: 我可以，但是什么时候去哪里呢? Anna: Sydney, 星期四3点 Marcus: am 还是pm? :D Leslie: haha, 幸运的是下午:D Marcus: hmm 我1点有个会议。我想我去不了了 Leslie: well 我想after his landing 他需要一些时间，reclaiming luggage etc Anna: 是的，我认为如果你4点到那里没问题了 Marcus: oh 那好吧 ...</p>	<p>Marcus originally plans to pick up Mark at the airport but has a meeting at 1 PM and can't make it. Leslie suggests Marcus arrive after 4 PM to give Mark enough time to reclaim luggage. Anna will inform Mark about Marcus's new plan.</p>	<p>Anna was the one who suggest Marcus to go at 4 PM which is wrongly attributed to Leslie</p>
<p>Ken: You are an idiot. Greg: What? Ken: Yeah you are, I want my money back Greg: Was your account hacked by some prankster? Ken: No, I'm the prankster, just having a laugh at your expense Greg: Well, you are an idiot too then XD</p>	<p>Ken: 操你，你pimp Greg: 什么? Ken: 操你，我要回我的钱 Greg: 你的账户被恶作剧者hacked了吗? Ken: 不，我才是恶作剧者，只是开个玩笑以你为代价 Greg: Well，那你也操XD</p>	<p>Ken and Greg were joking with each other. Ken played a prank on Greg but now he wants his money back. Greg was not happy about it</p>	<p>The dialogue is just a friendly banter and there's actually no transaction of money.</p>

Table 9: Error examples

EN Dialogue	CS Dialogue	LLM Summary	Error Explanation
<p>...</p> <p>Person2: Large deposits of coal were found nearby and so many industries located themselves here. The village quickly grew into a key industrial centre.</p> <p>Person1: As the city grew, it must have absorbed many village nearby.</p> <p>Person2: Yes, it did. The names of those village survive as the names of parts of the city.</p> <p>Person1: I see. Are there any building more than 200 years old in your city?</p> <p>Person2: Oh, yes. Several of the buildings from the villages still survive. Many of them were inns for travelers and today survive as pubs. There was a castle near one village, so our city has a castle too.</p> <p>Person1: Really? So your city does have some old history after all.</p>	<p>...</p> <p>Person2: 附近发现了大量coal. 许多工业都设在这里。这个村庄迅速发展成为一个重要的industrial centre。</p> <p>Person1: 随着城市发展，它一定absorbed了附近的许多村庄。</p> <p>Person2: 是的。这些村庄的名字作为城市部分地区的名字而留存下来。</p> <p>Person1: I see. 你的城市里有超过200年历史的building吗?</p> <p>Person2: 村庄里的一些buildings仍然保存完好。其中许多都是旅客的inns，如今作为pubs保留了下来。一个村庄附近有一座castle，所以我们的城市也有一座castle。</p> <p>Person1: 真的吗？原来你们城市也有悠久的历史啊。</p>	<p>Person2 tells Person1 how their city grew and has some buildings over 200 years old. Person2 also introduces a castle nearby</p>	<p>The summary is based on only the English words of the dialogue and also Person2 talks about a castle, does not introduce it</p>
<p>Alan: <file_photo></p> <p>Alan: look what I just found</p> <p>Robert: dude, that's just nasty and you know it</p> <p>Robert: it has no sugar, no taste, and additional cinnamon flavoring</p> <p>Alan: yeah, I know - that's awesome</p> <p>Robert: you sir have a very strange tastes :P</p> <p>Alan: well, and I found a perfect company for it <file_photo></p> <p>Robert: oh, that's more like it!</p> <p>Robert: but does the whiskey go well with the cinnamon? flavored whiskey is the worst...</p> <p>Alan: Actually it does taste surprisingly well. The cinnamon is not overpowering. If you put enough whiskey that is</p> <p>...</p>	<p>Alan: <file_photo></p> <p>Alan: tengok ape aku jumpe</p> <p>Robert: dude, that's just nasty and you know it</p> <p>Robert: it has no sugar, no taste, and additional cinnamon flavoring</p> <p>Alan: yeah, I know - itulah yang awesome</p> <p>Robert: you sir ada tastes pelik :P</p> <p>Alan: well, and I found a perfect company for it <file_photo></p> <p>Robert: oh, that's more like it!</p> <p>Robert: Tapi does the whiskey go well with the cinnamon? flavored whiskey rabak...</p> <p>Alan: Actually it does taste surprisingly well. The cinnamon tak terlalu kuat. If you put enough whiskey that is</p> <p>...</p>	<p>Alan shares a photo of a cinnamon-flavored whiskey with Robert, who initially finds it disgusting. Alan defends the unique taste and shares another photo of him enjoying it with a companion. Robert jokingly points out Alan's unusual taste preferences but admits the cinnamon taste isn't overpowering when paired with enough whiskey. They reminisce about a previous cherry flavor and agree it was better than the current sugar-free version.</p>	<p>The summary wrongly states that Alan has a companion when he was actually referring to eating cinnamon along with it</p>

Table 10: Error examples

EN Dialogue	CS Dialogue	LLM Summary	Error Explanation
<p>Callan: Something's wrong with my Samsung S8. Wade: What, exactly? Callan: All I get is a black screen. Wade: Did you drop it or something? Callan: I think I know, this has happened before. Wade: It did? What happened before? Callan: It overheats and then it takes a while for it to restart again.. Wade: That sucks.. Be careful with the battery, it can be dangerous. ...</p>	<p>Callan: Samsung S8 aku ada problem. Wade: What, exactly? Callan: Aku asyik dapat black screen. Wade: Kau terjatuhkan ke apa? Callan: Aku rasa aku tahu kot, ni pernah happened kat aku. Wade: It did? What happened before? Callan: Dia overheat pastu it takes a while for it to restart again.. Wade: That sucks.. Be careful with the battery, boleh jadi bahaya. ...</p>	<p>The speaker, Callan, is having issues with his Samsung S8 as it keeps getting black screens. Wade suggests it might be overheating. Callan intends to take it to a store as it's still under warranty</p>	<p>Callan was the one who suggested the phone might be overheating</p>
<p>Person1: What do you know about our company? Person2: Well, as I know this company is one of the largest and best in this field of business. It mainly produces toys for children. It employs more than 10,000 people throughout the world. The president now is Mr. Jackson. The Shanghai branch was founded five years ago with a staff of more than 2,000. Person1: Very good. If you enter this company, what department would you like to work in? Person2: May I work in sales department? Person1: OK. Oh, how much do you make at your current job, if you don't mind my asking? Person2: I'm paid 2,000 yuan per month plus some allowances and bonus as that come to about 3,000 yuan a month. How much will you pay me? Person1: Well, the starting salary for a clerk in our company is 2,500 yuan monthly and after three months he would get a raise if his work is satisfactory. Person2: Do you have any fringe benefits? ...</p>	<p>Person1: Enga company pathi enna therium ungaluku? Person2: Indha field indha company dhaan largest and best nu therium. Main focus vandhu children toys produce panradhu. Current president Mr. Jackson. Shanghai la 5 year munnadi oru branch start panirkeenga. Anga 2000 employees irukanga. Total ah 10000 employees. Person1: Very good. Inga job kedaicha, endha department la work pananum ungaluku? Person2: Sales department la? Person1: OK. Current job la evlo earn panreenga? If you don't mind my asking? Person2: Current salary 2000 yuan plus konjam allowances and bonus ku apram 3000 yuan per month. Neenga evlo pay panuveenga? Person1: Clerks ku starting salary 2500 yuan. 3 months ku apram raise varum. Person2: Fringe benefits edhavadhu iruka? ...</p>	<p>Person2 is describing their current company which is the largest and best in the field, focusing on children's toys. They have a branch in Shanghai with 2000 employees, totaling 10,000 employees worldwide. Person1 is interested in a sales position and asks about the salary and benefits, learning that Person2 currently earns 3000 yuan per month with allowances and bonuses. Person1 explains that the starting salary for clerks is 2500 yuan, with a raise after three months, plus benefits like semi-annual bonuses, a small Spring Festival bonus, and 4 weeks of</p>	<p>LLM has completely switched Person1 and Person2 in the first half of the summary</p>

Table 11: Error examples

Beyond Paraphrasing: Analyzing Summarization Abtractiveness and Reasoning

Nathan Zeweniuk^{1,2}, Ori Ernst^{1,2}, and Jackie Chi Kit Cheung^{1,2,3}

¹Mila – Quebec Artificial Intelligence Institute

²McGill University

³Canada CIFAR AI Chair, Mila

{oriern}@gmail.com

{nathan.zeweniuk@mail., jackie.cheung@}mcgill.ca

Abstract

While there have been many studies analyzing the ability of LLMs to solve problems through reasoning, their application of reasoning in summarization remains largely unexamined. This study explores whether reasoning is essential to summarization by investigating three questions: (1) Do humans frequently use reasoning to generate new summary content? (2) Do summarization models exhibit the same reasoning patterns as humans? (3) Should summarization models integrate more complex reasoning abilities? Our findings reveal that while human summaries often contain reasoning-based information, system-generated summaries rarely contain this same information. This suggests that models struggle to effectively apply reasoning, even when it could improve summary quality. We advocate for the development of models that incorporate deeper reasoning and abtractiveness, and we release our annotated data to support future research.

1 Introduction

In recent decades, the amount of textual information available has grown exponentially, creating a pressing need for automatic systems that can process this information and derive meaningful conclusions from it. Recent advances in large language models (LLMs) have shown remarkable progress in handling tasks that appear to require reasoning—namely, deriving conclusions not explicitly stated in the text. For instance, LLMs have demonstrated strong performance in question answering tasks that involve background knowledge and inference (Zhao et al., 2023; Liu et al., 2025). Yet, despite these advances, the role of reasoning in generic summarization remains largely underexplored. A key question arises: Can and should automatic summaries incorporate new conclusions that go beyond the information explicitly present in the source?

Traditionally, research in automatic summarization has focused on information selection and

paraphrasing (Zhang et al., 2018; Lebanoff et al., 2019b; Ernst et al., 2022). One widely used quality measure to human-like summaries has been abtractiveness—the degree to which a summary uses its “own words” rather than copying source text. With the emergence of LLMs, summarization systems have achieved substantial gains not only in content selection but also in producing highly fluent and abtractive outputs comparable to human-written summaries (Goyal et al., 2022). These advances invite a deeper investigation into the next frontier: Can automatic summaries perform reasoning, deriving conclusions like humans do?

In principle, the ability to reason during summarization could enhance content focus and informativeness, enabling the generation of summaries that emphasize the most salient insights rather than merely restating information. Table 1 illustrates how different summarization processes can yield outputs with varying levels of focus, conciseness, and reasoning.

To investigate this, we outline several research questions: (1) Do *humans* rely on reasoning to create summary content, and if so, how often? (2) Do *models* employ the same reasoning as humans, or do they bypass it? (3) Should we aim to incorporate reasoning abilities in summarization modeling?

To address these questions, we began with a manual annotation of human-generated summaries. We identified three common operations that humans perform when rewriting selected text for summaries (will be defined clearly in Section 3): paraphrasing, generalization, and drawing conclusions. The latter two operations, generalization and conclusion, change the semantic meaning from the source to the summary and are considered to require reasoning.¹ We manually classified matching text spans between summaries and source texts according to

¹We acknowledge that reasoning may be used in paraphrasing, but such reasoning does not fall within the scope of our work as it does not lead to new semantic content.

Type	Text
Source	Investigators are trying to piece together what led to the deaths
Paraphrase	Investigators are attempting to figure out how the deaths occurred
Generalization	Investigators are working on a case
Conclusion	Mystery surrounds the death of two brothers

Table 1: An example of abstractiveness levels that can be applied to a source sentence

these levels of abstractiveness and found that approximately 25% of human summary spans involve such reasoning.

However, our manual evaluation of system-generated summaries revealed a different pattern. Despite high overall evaluation scores with respect to reference summaries, these systems predominantly matched the reference with paraphrased text spans. Crucially, reference spans that require reasoning to extract important information were under-represented in systems summaries. In other words, while these models seem to display reasoning abilities in other tasks, they tend to avoid using them correctly in summarization, where output is often deemed acceptable without it, despite the fact that this omission can reduce the quality of the summary.

This analysis highlights the importance of reasoning in summarization and calls for the development of new models that better integrate reasoning and different levels of abstraction. To facilitate further research, we are releasing the manual annotations and data.²

2 Related Work

2.1 Abstractiveness Analysis

Prior research on abstractiveness has focused on how multiple source sentences are fused into one summary sentence (Barzilay and McKeown, 2005). Early work analyzed syntactic aspects of fusion, investigating whether sentences are merged or concatenated (Lebanoff et al., 2019a), and explored points of correspondence between sentences (Lebanoff et al., 2020). These studies were con-

ducted on nearly-extractive data and based mostly on paraphrasing.

More recent research has expanded to include specific cases of entity generalization in summaries that went beyond paraphrasing (González et al., 2022; Jumel et al., 2020). A new aggregation metric (He et al., 2023) has also been introduced to capture even non-paraphrasing forms of fusion. Overall, most existing studies focus on sentence fusion and lack detailed manual annotations of abstractiveness across varying levels of sentence complexity.

The work most closely related to our approach is that of Jing (2002), who conducted a seminal study examining the operations employed by human summarizers when composing abstractive summaries. Specifically, they identified six key actions: sentence reduction, sentence combination, syntactic transformation, lexical paraphrasing, generalization/specification, and content reordering. Based on the hypothesis that abstractive summarization relies on these operations, one can characterize a summary’s level of abstractiveness by measuring the extent to which each action is applied.

As mentioned, while modern summarization systems are capable of performing many of these operations, especially those related to paraphrasing, our work focuses on a different dimension: reasoning. We argue that, despite advancements in paraphrasing and surface-level transformations, reasoning-based abstraction, the ability to derive new conclusions or implicit insights, remains underexplored and is largely absent from prior work.

2.2 LLMs Reasoning Ability

Reasoning benchmarks, such as commonsense (Talmor et al., 2019), logical (Sinha et al., 2019), math (Saxton et al., 2019), or multi-hop (Tu et al., 2019), were considered a difficult task for language models to solve. Most of these benchmarks were designed in a Question-Answering (QA) setup, where a query or a question is given, and the answer can be found in a defined set of sources. In order to find the answer, the system is expected to use reasoning. Recent advancements in LLMs showed significant improvement on these benchmarks (Driess et al., 2023; Touvron et al., 2023; Espejel et al., 2023), especially while using chain of thought (Kojima et al., 2022; Wei et al., 2022). However, all these benchmarks are designed to elicit reasoning. In contrast, this paper aims to evaluate the reasoning abilities of models in summarization tasks, where acceptable outputs can still be achieved with-

²Annotated data is publicly available at <https://github.com/orienn/ReasoningSummarization>.

out reasoning.

3 Abtractiveness Levels

There are a few ways in which source information can be utilized in the process of summarization. The simplest approach is to copy the information directly and reword or restructure it to fit the rest of the summary. Generalizing certain parts of the source material can help compress information even further by reducing the amount of detail and level of specificity. Sometimes, it is necessary to add entirely new information drawn from the source through reasonable conclusions to avoid relying on reader inference. We use these observations regarding the various uses of source information to define abstractiveness levels.

First, we define a span-level matching between the information in the summary and its corresponding evidence from the source. Having these matching pairs allow to analyze the abstractiveness level performed in the summary. Following Ernst et al. (2021, 2024) the spans are standalone facts that are usually formed into a proposition, where the source span entails the summary span. We also required *tight* matching, where each source token that adds additional information that the summary is not based on, is omitted.

Given these pairs, the abstractiveness levels are defined as follows:

Paraphrase. Bi-directional entailment between the summary span and the document span. That is, both sides share the same information

Generalization. The summary and document spans are event-coreferred³. The summary span does not explicitly mention specific details but instead uses broader terms that encompass those details. As a result, while the source span entails the summary span, the summary span does not fully entail the source span.

Conclusion. The summary span adds new information that is not mentioned explicitly in the source but derived from it. Accordingly, while the source span entails the summary span, the summary does not entail the source in full, and they are not event-coreferred.

The examples in Table 1 demonstrate how these guidelines are applied. More examples can be seen in the Appendix (Table 5).

³According to Eirew et al. (2022), event-coreferred spans are spans that refer to the same event.

Dataset	Paraph.	General.	Conc.	NA
DUC	65.3	13.8	11.1	9.8
FewSum	55.8	20.3	9.8	14.1
MultiNews	60.6	12.5	16.2	10.7

Table 2: Average distribution of span-level abstractiveness types in reference summaries (%)

Note that by definition, if a pair contains more than one type, it should be classified according to the more lenient type. For example, if one part of a summary span generalizes while other parts derive a conclusion, the spans are not considered *fully* coreferred and should be classified as a Conclusion. Thus, the final hierarchy is: Conclusion > Generalization > Paraphrase.

4 Annotation Process

In order to understand how often different levels of abstractiveness appear in human written summaries, we annotate reference summaries from the news and review domains. This annotation was performed manually by an expert annotator.

4.1 Alignment data

As outlined in Section 3, abstractiveness levels are determined by analyzing matching summary-source pairs. To achieve this, we utilize existing human-annotated source-summary span alignments from three multi-document summarization datasets: two from the news domain—DUC (NIST, 2014) and MultiNews (Fabbri et al., 2019)—and one focused on business reviews, FewSum (Bražinskas et al., 2020). Specifically, we used 315 document-summary span pairs across 12 summaries from DUC alignments (Ernst et al., 2021), 250 pairs from 16 summaries from MultiNews alignments (Ernst et al., 2024), and 336 pairs from 17 summaries using FewSum alignments (Slobodkin et al., 2024). For each pair, the annotator classified the abstractiveness type.

4.2 Annotation

The annotation process is composed by preliminary alignment data cleaning, pair-level annotation, aggregation to the span level, and finalizing rate calculation across all summaries. In this section, we elaborate about each of these components.

As we heavily rely on the previously annotated source-summary alignment, to ensure the quality of our annotations, we first cleaned the alignment

data. This data allowed for some degree of noise, details that appeared in the summary or source span but not in the other span from the pair, such as ‘two brothers’ in the conclusion example from Table 1. To meet our definition of *tight* alignment, we first omitted tokens from both the summary and source spans, ensuring no unaligned tokens remained, provided that such omissions did not contradict or significantly alter the span’s meaning.

Then, we annotate each source-summary pair with one of the three types. If entailment could not be achieved even after omitting details, the pair was labeled as ‘not aligned’.

Since a single summary span can be aligned with multiple document spans, and the level of abstractiveness is defined at the pair level, it is necessary to aggregate the pair-level decisions into a single summary span-level label. Specifically, each summary span was aligned separately with each of its corresponding document spans, and abstractiveness was annotated for every summary–document pair. After completing the pair-level annotations, we aggregated all annotations associated with the same summary span to derive a final abstractiveness label for that span.

Since we do not know which document span influenced the summary span the most, we adopt a strict approach, assuming that summarizers use the simplest level of abstractiveness possible. Specifically, if a summary span is aligned with multiple document spans, each annotated with a different abstractiveness type, we assume the summarizer employed the least abstract type. Paraphrase is the simplest type, followed by Generalization, and finally Conclusion. Therefore, a summary span is considered derived from Conclusion only if all associated pairs are labeled as Conclusion. If all pairs are labeled ‘not aligned’, the summary span is considered ‘not aligned’.

Finally, we calculate the rate of each type in a summary and averaged across summaries in the same dataset.

4.3 Quality Evaluation

To assess the clarity of the task and the reliability of the guidelines, we recruited an additional expert annotator to independently annotate a subset of three summaries (one from each dataset) and measured inter-annotator agreement. Agreement was evaluated under two conditions: (1) for the full process, which included both span cleaning and classification; and (2) for classification only, where

the annotator received the already cleaned spans. The annotators reached an agreement of 81.5% on the classification decisions and 69.2% on the full process when including the alignment-tightening process. These results indicate that the annotation guidelines are clear and can be applied consistently across annotators.

4.4 Results

As shown in Table 2, while Paraphrasing is the most common level of abstractiveness, approximately 25-35% of the reference summaries include instances of Generalization and Conclusion. This indicates that human summarizers frequently go beyond simple paraphrasing. Generalization is more prevalent than Conclusion in the reviews dataset, where summarizers often generalize across multiple personal experiences reported by customers. This tendency is more expected in reviews but less common in the news domain, where summarization typically focuses on factual reporting.

5 System-Generated Summaries

We observed a substantial presence of generalization and conclusion actions in human-written reference summaries, raising questions about whether automatic summarization systems are capable of generating similar types of information. However, due to the lack of available alignment data between system-generated summaries and their corresponding source documents, data required for our annotation process, we were unable to apply the same fine-grained annotation procedure to system summaries as we did for human references.

Instead, we analyzed how well system-generated summaries align with reference summaries across different abstractiveness levels. Our findings reveal that system outputs tend to match paraphrase-based reference spans far more frequently than generation- or conclusion-based spans. In other words, the high similarity scores that system summaries achieve relative to reference summaries primarily stem from their ability to produce effective paraphrases, rather than from generating novel conclusions or inferred content akin to those written by humans.

To examine this phenomenon, we conducted an abstractiveness-aware manual evaluation inspired by the Pyramid method (Nenkova and Passonneau, 2004). We selected 10 topics from MultiNews, 10 from FewSum, and 6 from DUC 2004. Sys-

Model	Paraph.	General.	Conc.	Total
GPT 4o m.	54.0	34.1	30.9	38.1
PRIMERA	35.1	16.7	0.0	20.8
Llama 3	55.6	35.5	33.1	39.1

Table 3: Average span-level recall scores of system summaries for each abstractiveness type

Reference	System
Democrats, meanwhile, argue that it’s too soon to scale back the program	Democrats argue against these cuts, pointing out that many families still depend on the program
The quality of the theater is superb	the quality...justify the cost

Table 4: Examples of reference spans and related system spans with different abstractiveness levels.

tem summaries were generated using three models: PRIMERA fine-tuned on MultiNews (Xiao et al., 2022), GPT-4o mini in a zero-shot setting (OpenAI, 2024), and Llama 3 8B Instruct in a zero-shot setting (Dubey et al., 2024).

Following the standard Pyramid approach, an expert annotator segmented each reference summary into factual, standalone spans and then identified which of these were matched by the corresponding system-generated summary. The overall recall of matched spans for each model is reported in Table 3 (“Total”).

Building on these matches, we computed a recall score for each abstractiveness level, reflecting the proportion of reference spans of a given type that were successfully reproduced by the system. As shown in Table 3, coverage varied substantially across types. For instance, GPT-4o mini matched 54% of reference spans classified as Paraphrase, but only 30.9% of those classified as Conclusion. Across all systems, Paraphrasing achieved the highest recall, while Conclusion and Generation spans were captured less frequently.

These findings suggest that current summarization models excel at reproducing paraphrased content but struggle to incorporate reasoning-based or conclusion-oriented information that match the reference summary, particularly when such reasoning is not explicitly required by the input.

It is important to note that, in a few cases, models did select the same source information as the reference summary did, but because they employed

a different level of abstraction, their output was too distant from the reference to be considered a match. Examples of this phenomenon are provided in Table 4 and the Appendix (Table 6). This highlights that while models can identify relevant content, they may require further guidance on how to apply the appropriate level of abstraction.

6 Conclusion

In this work, we analyzed different abstraction levels in summarization, and found that while humans use reasoning to derive information which improves the focus and clarity of summaries, models are still lagging behind. We release our data and annotation to facilitate research in this direction and the development of summarization models that incorporate better reasoning abilities.

Limitations

Due to a lack of available system summary-source alignment data at the time of this project, we were unable to perform system summary annotation of the same kind of the reference summary annotation. From our analysis alone, we cannot conclude that system generated summaries do or do not contain generalization or conclusion information, only that they do not often apply these levels of abstractiveness in the same way as human summarizers.

Abstraction level of summary-source pairs were annotated, separately, one pair at a time. As a result, some summary spans may have been marked ‘not aligned’ if they require information fusion across many source spans.

Acknowledgments

We thank the anonymous reviewers for their constructive comments. This work was supported in part by the IVADO Postdoctoral Fellowship, Canada CIFAR AI Chair, and the Natural Sciences and Engineering Research Council of Canada.

References

- Regina Barzilay and Kathleen R. McKeown. 2005. [Sentence Fusion for Multidocument News Summarization](#). *Computational Linguistics*, 31(3):297–328.
- Arthur Braźinskas, Mirella Lapata, and Ivan Titov. 2020. [Few-shot learning for opinion summarization](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4119–4135, Online. Association for Computational Linguistics.

Danny Driess, F. Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Ho Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Peter R. Florence. 2023. [Palm-e: An embodied multimodal language model](#). In *International Conference on Machine Learning*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, 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 Alonsius, Daniel Song, Danielle Pintz, Danny Livshits, 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, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, 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, 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, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Young, 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, 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, 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, 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 Rapparthi, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collet, 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 Gougeon, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuwei 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 Papanikos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, 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, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, 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, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, 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, Karthik Prasad, Kartikay Khan-delwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhota, 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, Maria Tsim-poukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Her-moso, Mo Metanat, Mohammad Rastegari, Mun-ish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pa-van Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratan-chandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Mah-eswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lind-say, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agar-wal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, 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, Vitor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiao-jian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. [The llama 3 herd of models](#).
- Alon Eirew, Avi Caciularu, and Ido Dagan. 2022. [Cross-document event coreference search: Task, dataset and modeling](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 900–913, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Ori Ernst, Avi Caciularu, Ori Shapira, Ramakanth Pasunuru, Mohit Bansal, Jacob Goldberger, and Ido Dagan. 2022. [Proposition-Level Clustering for Multi-Document Summarization](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1765–1779, Seat-tle, United States. Association for Computational Linguistics.
- Ori Ernst, Ori Shapira, Ramakanth Pasunuru, Michael Lepioshkin, Jacob Goldberger, Mohit Bansal, and Ido Dagan. 2021. [Summary-source proposition-level alignment: Task, datasets and supervised baseline](#). In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 310–322, Online. Association for Computational Linguistics.
- Ori Ernst, Ori Shapira, Aviv Slobodkin, Sharon Adar, Mohit Bansal, Jacob Goldberger, Ran Levy, and Ido Dagan. 2024. [The power of summary-source alignments](#). In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6527–6548, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.
- Jessica Nayeli López Espejel, El Hassane Ettifouri, Mahaman Sanoussi Yahaya Alassan, El Mehdi Chouham, and Walid Dahhane. 2023. [Gpt-3.5, gpt-4, or bard? evaluating llms reasoning ability in zero-shot setting and performance boosting through prompts](#). *Nat. Lang. Process. J.*, 5:100032.
- Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. [Multi-News: A Large-Scale Multi-Document Summarization Dataset and Abstractive Hierarchical Model](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.
- José Ángel González, Annie Louis, and Jackie Chi Kit Cheung. 2022. [Source-summary entity aggregation in abstractive summarization](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6019–6034, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. [News summarization and evaluation in the era of gpt-3](#). *ArXiv*, abs/2209.12356.
- Jingyi He, Meng Cao, and Jackie Chi Kit Cheung. 2023. [Analyzing multi-sentence aggregation in abstractive summarization via the shapley value](#). In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 121–134, Singapore. Association for Computational Linguistics.
- Hongyan Jing. 2002. [Using hidden Markov modeling to decompose human-written summaries](#). *Computational Linguistics*, 28(4):527–543.
- Clément Jumel, Annie Louis, and Jackie Chi Kit Cheung. 2020. [TESA: A Task in Entity Semantic Aggregation for abstractive summarization](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8031–8050, Online. Association for Computational Linguistics.

- Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. [Large language models are zero-shot reasoners](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.
- Logan Lebanoff, John Muchovej, Franck Deroncourt, Doo Soon Kim, Seokhwan Kim, Walter Chang, and Fei Liu. 2019a. [Analyzing sentence fusion in abstractive summarization](#). In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 104–110, Hong Kong, China. Association for Computational Linguistics.
- Logan Lebanoff, John Muchovej, Franck Deroncourt, Doo Soon Kim, Lidan Wang, Walter Chang, and Fei Liu. 2020. [Understanding points of correspondence between sentences for abstractive summarization](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 191–198, Online. Association for Computational Linguistics.
- Logan Lebanoff, Kaiqiang Song, Franck Deroncourt, Doo Soon Kim, Seokhwan Kim, Walter Chang, and Fei Liu. 2019b. [Scoring Sentence Singletons and Pairs for Abstractive Summarization](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2175–2189, Florence, Italy. Association for Computational Linguistics.
- Gabrielle Kaili-May Liu, Bowen Shi, Avi Caciularu, Idan Szpektor, and Arman Cohan. 2025. [MDCure: A scalable pipeline for multi-document instruction-following](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29258–29296, Vienna, Austria. Association for Computational Linguistics.
- 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.
- NIST. 2014. Document understanding conferences. <https://duc.nist.gov>. Accessed: 2019-12-02.
- OpenAI. 2024. [\[link\]](#).
- David Saxton, Edward Grefenstette, Felix Hill, and Pushmeet Kohli. 2019. [Analysing mathematical reasoning abilities of neural models](#). In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.
- Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L. Hamilton. 2019. [CLUTRR: A diagnostic benchmark for inductive reasoning from text](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4506–4515, Hong Kong, China. Association for Computational Linguistics.
- Aviv Slobodkin, Ori Shapira, Ran Levy, and Ido Dagan. 2024. [Multi-review fusion-in-context](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3003–3021, Mexico City, Mexico. Association for Computational Linguistics.
- Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. [CommonsenseQA: A question answering challenge targeting commonsense knowledge](#). 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 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. [Llama: Open and efficient foundation language models](#). *ArXiv*, abs/2302.13971.
- Ming Tu, Guangtao Wang, Jing Huang, Yun Tang, Xiaodong He, and Bowen Zhou. 2019. [Multi-hop reading comprehension across multiple documents by reasoning over heterogeneous graphs](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2704–2713, Florence, Italy. Association for Computational Linguistics.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.
- Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2022. [PRIMERA: Pyramid-based masked sentence pre-training for multi-document summarization](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5245–5263, Dublin, Ireland. Association for Computational Linguistics.
- Xingxing Zhang, Mirella Lapata, Furu Wei, and Ming Zhou. 2018. [Neural Latent Extractive Document Summarization](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 779–784, Brussels, Belgium. Association for Computational Linguistics.
- Wenting Zhao, Justin Chiu, Claire Cardie, and Alexander Rush. 2023. [Hop, union, generate: Explainable multi-hop reasoning without rationale supervision](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages

16119–16130, Singapore. Association for Computational Linguistics.

A Reference Summary Examples

Table 5 contains examples of aligned pairs from the datasets and their corresponding types after the annotation process.

B System Summary Examples

In Table 6, we provide examples of similar spans from the reference and system summaries. Summaries generated with GPT 4o mini and Llama 3 used the prompt, "The following are news articles on a single topic. Please create one summary from all of the articles:", followed the source documents.

C System Summary Evaluation by Dataset

We present the system summary evaluation statistics broken down by dataset in Table 7.

Source	Summary	Type
People now can look at Hawaii as a destination to have their marriage	That could be a boon for destination weddings in a state	Conclusion
a source tells Us that Maroon 5 have been tapped to grace the halftime stage	sources tell Variety the Super Bowl LIII halftime act has been chosen, and it's Maroon 5	Paraphrase
the court will return to the subject of whether the Constitution permits public colleges and universities to take account of race in admissions decisions	The court...will look at cases including...affirmative action	Generalization
an estimated 16 million children, or about one in five, received food stamp assistance	about 16 million, or nearly one in five, of them are doing so fueled by food stamps	Paraphrase
"I actually quite like the color," said psychologist Dr. Carolyn Mair..."It's an earthy, muted, rich color, very much of nature	But not everyone's on board with trashing opaque couche	Generalization
You should always use the chip device, not the swipe device	Credit card purchases are about to get...a lot more secure	Not Aligned
Can the sight of a greenish-brown color really be enough to deter smokers from reaching for their next pack of cigarettes?	the...color... been used to try to save lives	Conclusion
waited over 20 min	The food takes to long to come out	Generalization
waited over 20 min ... not even a sorry about the wait	service is horrible	Conclusion
newest hottest spot ... Food is amazing	Overall not a recommended place	Not Aligned

Table 5: Examples of reference spans and related system spans with different abstractiveness levels.

Reference Type	Reference	System	Matched
Generalization	Instagram...simply dropping the changes	Instagram has decided to revert to its previous terms	Yes
Paraphrase	The chips in the new cards use a system known as EMV, for creators Europay, MasterCard, and Visa	These EMV (Europay, MasterCard, and Visa) chips	Yes
Conclusion	police "do not believe there are any outstanding suspects," per a spokesperson	authorities are not currently seeking any suspects	No
Conclusion	Flickr's app jumped in popularity	NA	No
Generalization	the opposition tried to cut off his access to loans.	opposition...urged international bodies, such as the Asian Development Bank, to reconsider support	No
Generalization	Police got a call...that afternoon	Police...receiving a distress call around 2:45 p.m	Yes
Paraphrase	Crowe eventually returned to the matter of the accent, saying, "I'm a little dumbfounded you could possibly find any Irish in that character-that's kind of ridiculous, but it's your show.	Crowe became defensive and denied the accusation	No
Conclusion	the investigation is ongoing	The investigation is ongoing	Yes

Table 6: Examples of reference spans and related system spans with different abstractiveness levels.

Model	Dataset	Paraphrase	Generalization	Conclusion	Total
GPT 4o mini	DUC2004	44.7	8.3	66.7	34.8
	FewSum	63.5	37.5	17.9	41.9
	MultiNews	51.0	41.6	29.1	36.3
PRIMERA	DUC2004	23.1	0.0	0.0	16.5
	FewSum	34.8	13.3	0.0	16.5
	MultiNews	42.7	27.8	0.0	27.7
Llama 3 8b	DUC2004	44.7	33.3	0.0	37.7
	FewSum	72.6	33.4	35.7	42.6
	MultiNews	46.9	38.9	42.0	36.4

Table 7: Average span-level recall scores of system summaries for each abstractiveness type and dataset.

Improving Aspect-Based Summarization via Contrastive Learning with Anchored Negative Examples

Elizabeth Palmieri Yangfeng Ji

Department of Computer Science

University of Virginia

Charlottesville, VA 22903

{cxm7ja, yangfeng}@virginia.edu

Abstract

Text summarization helps users manage information overload, but traditional methods can be cumbersome when seeking specific details within a document. Aspect-based text summarization addresses this by using a query to guide which information should be summarized. However, distinguishing relevant from irrelevant information for a given aspect remains challenging in LLM-based summarization models. In this work, we propose utilizing contrastive learning to encourage LLMs to focus on aspect-related signals during training. We further design two variants of the learning algorithm, aspect-anchored and summary-anchored, corresponding to the strategies used in constructing negative examples. Evaluation with two representative LLM families (Llama 2 and Pythia) and two benchmark datasets (AnyAspect and CovidET) demonstrates the proposed methods' strong performance compared to their supervised fine-tuning and zero-shot counterparts, highlighting contrastive learning as a promising direction for aspect-based text summarization.¹

1 Introduction

Aspect-based text summarization is a crucial task within Natural Language Processing that addresses the limitations of conventional summarization methods. Traditional models generate concise summaries of documents, aiming to save users time and effort. However, the varying lengths and diverse content of documents often render general summaries inadequate. Consider a news article covering an event from multiple perspectives – location, key figures, and the event itself. A user might be interested in a summary focused on the event's location rather than the individuals involved, as shown in Figure 1. Aspect-based summarization tackles this issue by providing both a document and

¹The code and data will be available at <https://github.com/elizabethpalmieri/contrastive-asp-summ.git>

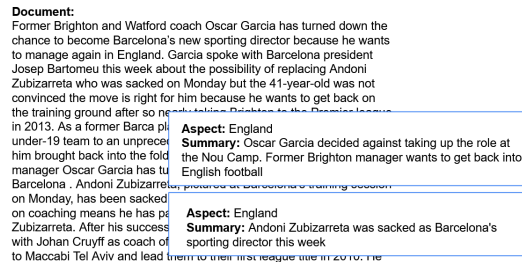


Figure 1: An illustration of aspect-based text summarization created based on an example from the AnyAspect dataset (Tan et al., 2020a). The first summary is related to the given aspect ENGLAND, while the second one is not.

a specific aspect, guiding the model to generate a summary tailored to that particular focus. The objective is to parse the document and selectively use the information only relevant to the given aspect.

A significant challenge in aspect-based text summarization lies in the model's ability to isolate and highlight aspect-specific information while effectively distinguishing it from the rest of the text. Without this crucial capability, the model risks generating a generic summary that fails to address the intended aspect. For example, a model presented with a news article about an earthquake (Ahuja et al., 2021) and tasked with generating summaries for two aspects (GEOGRAPHY and RECOVERY) might produce identical summaries detailing the earthquake's magnitude and recovery effort, neglecting the distinct nuances of each aspect. This underscores the need for a mechanism that enables the model to cluster similar information and separate dissimilar information within the latent space.

While prior work has established a foundation for understanding the capabilities of large language models (LLMs) in aspect-based text summarization (Mullick et al., 2024) and explored the impact

of contrastive elements (Gunel et al., 2023a), there remains a gap in informing aspect-based signals in LLMs. Furthermore, it lacks dedicated learning algorithms designed to teach LLMs to distinguish aspect-related signals during training (e.g., via comparing the aspects and their associated summaries). While adding instructions in prompts (e.g., “*Summarize this article with respect to Aspect within one short sentence.*”) has shown promise (Yang et al., 2023a), a training-level contrastive loss function is necessary to effectively incorporate aspect information and emphasize crucial signals within the data.

To address this gap, we propose a novel approach that integrates contrastive learning into the fine-tuning of open-source LLMs to enhance aspect-based text summarization. We augment existing aspect-based summarization datasets to create two types of anchored negative examples: *aspect* anchored negative examples (same aspect with different summaries) and *summary* anchored negative examples (same summary with different aspects). During fine-tuning, the model is trained to differentiate either summaries or aspects with an anchor presented. By learning to discriminate between these pairs, the model gains additional context and generates summaries that are more closely aligned with the given aspect.

This work, to the best of our knowledge, is the first to integrate contrastive learning into LLM fine-tuning specifically for aspect-based text summarization. Our contributions are threefold: (1) We design a novel contrastive learning algorithm with anchored negative examples for LLMs fine-tuning on aspect-based text summarization; (2) We perform an empirical study evaluating the performance of two prominent LLMs on two benchmark datasets; (3) We compare our method against both supervised fine-tuned and zero-shot LLMs, as well as three established baseline models (Flan-T5, BART, and GPT4-o). To gain deeper insights into our results, we also present an analysis on the important factors of using the proposed algorithm in practice.

2 Related Work

The research work in this paper is related to prior work at least in the following two dimensions: aspect-based text summarization and contrastive learning.

Aspect-based summarization. As previously described, the motivation behind aspect-based sum-

marization is to extract and summarize information relevant to a given aspect from a document. For instance, one early work by Paul et al. (2010) aimed to summarize contrasting opinions to enhance the comprehensiveness of the final summary.

Recent research in aspect-based summarization can be broadly categorized into two areas: developing novel learning strategies and constructing new datasets. Regarding learning strategies, Gunel et al. (2023b) proposed a four-step approach, encompassing aspect extraction and merging, with each step employing a specialized model, such as a fine-tuned T5 model for aspect extraction. Ding et al. (2024) augment continual learning for aspect-based sentiment analysis. Tang et al. (2024) use in context learning and aspect-based sentiment analysis for review summarization through the generation of Key Points. Addressing challenges like missing aspects, Li et al. (2023) introduced an unsupervised method for extracting opinions from source documents for summarization. Unlike these previous studies, this paper focuses on improving the summary accuracy of aspect-based summarization. Specifically, for a given aspect, the proposed approach ensures that the generated summary is relevant to that particular aspect and not to other aspects. In addition, recent research has shown that LLMs cannot handle aspect-based summarization as well as traditional summarization tasks (Yang et al., 2023a), which is echoed in our baseline experiment (section 5) and further motivated to develop new learning strategies.

Concurrent with the development of new learning strategies, significant effort has been dedicated to creating benchmark datasets for aspect-based summarization. These datasets span various domains, including scientific publications (Takeshita et al., 2024), climate change (Ghinassi et al., 2024), social media (Zhan et al., 2022a), legal decisions (T.y.s.s. et al., 2024), disordered texts (Guo and Vosoughi, 2024), and news articles (Ahuja et al., 2022; Tan et al., 2020b). While many datasets are limited in the number of aspects or specific domains to focus on, some works have explored a larger range of aspects (Tan et al., 2020b) or not targeted on specific domains (Yang et al., 2023b). In this work, we utilize the ANYASPECT dataset (Tan et al., 2020b) and the COVIDET dataset (Zhan et al., 2022a) to ensure evaluation across diverse domains and document lengths.

Contrastive learning for summarization.

While contrastive learning has been surveyed generally within the domain of natural language processing (Zhang et al., 2022) and extensively employed to enhance representation learning in text summarization (Xu et al., 2022), its application to aspect-based summarization remains largely unexplored. Specifically, contrastive learning has been shown to improve the alignment of generated summaries with source documents, reducing factual inconsistencies and hallucinations (Cao and Wang, 2021; Liu et al., 2022). It has also been used to address the issue of exposure bias (Sun and Li, 2021).

Zheng et al. (2021) further explored different strategies for constructing contrastive examples, such as masking, swapping, and replacing words or sentences, to improve learning performance in text summarization. Zhuang et al. (2024) proposed an automated method of constructing "hard" negative examples for contrastive learning. This method is applied within standard text summarization—our method similarly creates "hard" negative examples by leveraging negative aspects and summaries from the same source document as the positive example, thus reducing potential noise.

Liu and Liu (2021) proposed a novel approach that leverages contrastive learning to formulate summarization generation as a reference-free evaluation problem, where the model is trained to distinguish between high-quality and low-quality summaries without relying on reference summaries. Liu et al. (2024) utilize LLMs as evaluators in contrastive learning for smaller models such as BART. Wu et al. (2020) create a new evaluation method for text summarization without using gold standard summaries. They utilize linguistic and semantic aspects to perturb text summaries for negative samples and train their evaluator using contrastive learning. Chern et al. (2023) mitigate hallucinations and non-factual information in text summaries by implementing a contrastive reward learning framework that generates candidate summaries from a pretrained sequence-to-sequence model, which are then ranked using factuality metrics. Feng et al. (2024) utilize Contrastive Preference Optimization to mitigate hallucinations made by LLMs in news summarization.

Among existing works, Wang and Wan (2021) presents the most closely related study, employing contrastive learning to inform the (dis)similarity of aspects. However, the present work differs sig-

nificantly in both loss function design and training strategy. Specifically, the proposed method defines the contrastive loss directly on the generated aspect-based summaries, compelling large language models to capture the crucial aspect-related information from long input documents.

3 Contrastive Fine-tuning

In aspect-based text summarization, the length and content variability of documents create a difficult learning environment for models. A successful model must identify and extract information specifically related to the given aspect and synthesize a summary. If the model cannot effectively distinguish between informative sentences, it risks generating a generic summary that omits aspect-specific information, thus failing to fulfill its objective. To facilitate the model’s learning of pertinent aspect-related information, we employ contrastive learning through a particular design on loss function and example construction.

As with other applications of contrastive learning in text summarization, two key components are essential: the definition of contrastive loss and the construction of contrastive examples. In this work, to mitigate hallucinations, we utilize only the reference summary of each example as the positive example. Consequently, our focus in example construction is solely on generating negative examples.

For each training example consisting of a document d , aspect a , and summary s , the formulation of contrastive learning for aspect-based text summarization can be expressed as:

$$\mathcal{L}_s = -\left[\log p(s \mid d, a) - \frac{\lambda}{n} \sum_{k=1}^n \log p(s_k^- \mid d, a)\right] \quad (1)$$

where s_k^- represents a contrastive summary given the same document d and aspect a , n is the number of negative examples per case, and λ is the contrastive coefficient that balances the original training loss and the contrastive loss. We recommend $\lambda \in (0, 1)$, and a more detailed discussion will be presented in section 6. As the model is trained to differentiate the correct summary from its contrastive example with the same aspect, we named this method as **aspect-anchored contrastive learning** with an example shown in Figure 2.

Unlike losses defined in previous work (Oord et al., 2018), the contrastive loss in Equation 1 can be interpreted as the geometric mean of the likelihoods of the negative examples, $\log(\prod_{k=1}^n p(s_k^- \mid$

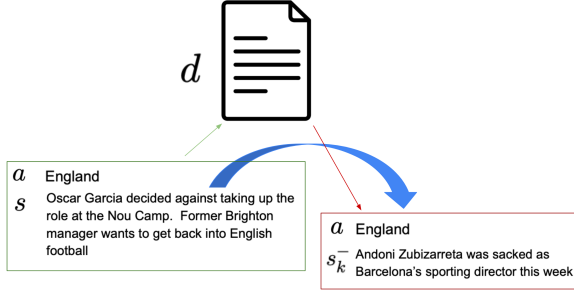


Figure 2: Illustration of *aspect-anchored* contrastive example generation from the AnyAspect dataset (Tan et al., 2020a). Given the document and the aspect, a contrastive example is constructed by selecting a summary from the same source document with a different aspect.

$d, a))^{1/n}$, which is less sensitive to extreme cases where the loss of a single negative example might dominate the others.

Variants of the Loss. Another variant of the loss function focuses on the aspect dimension rather than the summary. The corresponding loss function is defined as:

$$\mathcal{L}_a = -\left[\log p(s \mid d, a) - \frac{\lambda}{n} \sum_{k=1}^n \log p(s \mid d, a_k^-)\right] \quad (2)$$

where a_k^- is a random aspect from the the same document such that $a_k^- \neq a$. Unlike \mathcal{L}_s , \mathcal{L}_a is defined directly on the aspect while maintaining the same summary. Intuitively, this loss function encourages the model to focus on the contrastive comparison between the original aspect and the negative aspects. However, in practice, we observed that this loss often confuses the model during generation, as the outputs for both the original loss and the contrastive loss are identical—both are s . Following the same naming convention, we name this method as **summary-anchored contrastive learning**.

A further variant could combine \mathcal{L}_s and \mathcal{L}_a , applying negative examples to both summary and aspect variables. However, this poses additional training challenges, so we chose not to explore this direction further.

Quality of Negative Examples. In this work, for a given document and aspect, we select the summary from another aspect within the same document as the negative example in \mathcal{L}_s . Similarly, for \mathcal{L}_a , we choose the aspects available in the same document as the negative examples. By augmenting the summaries and aspects from within the

same data point, we ensure that these negative examples are on-topic and semantically sound.

We noticed that prior work on contrastive learning often struggles with the quality and various issues when constructing negative examples, as discussed in Zhang et al. (2022). For example, noise introduced into the negative examples may cause further issues in a text summarization system, such as hallucinations (Ji et al., 2023). On the other hand, the proposed example construction method helps the model learn which information is similar and dissimilar from the same input document without the risk of noise from automatically perturbed samples.

4 Experimental Setup

This section describes the experimental setup, including datasets, baseline and competitive models, evaluation metrics, and additional implementation details.

4.1 Datasets

We run our experiments on two aspect-based text summarization datasets: CovidET (Zhan et al., 2022b), and AnyAspect (Tan et al., 2020a). We choose two datasets with vastly different domains to ensure that our method can perform well on datasets of varying content. Details about dataset size can be referenced at Table 4.

CovidET consists of sentiment aspect-based text summarization of 1,900 Reddit posts from r/COVID19_support. There are seven emotion-based aspects: ANGER, ANTICIPATION, JOY, TRUST, FEAR, SADNESS, and DISGUST. The Reddit posts range from 50 – 500 tokens, averaging at 100 tokens.

AnyAspect is derived from the popular CNN/DailyMail (Hermann et al., 2015) dataset, in which a named entity recognition model was used to extract aspects from the pre-existing data and formulate an aspect-based text summarization dataset. The original dataset size has over 2 million training examples with more than 339 thousand aspects, with each document averaging around 680 tokens. To make sure the evaluation focuses on contrastive learning instead of a large number of aspects, we selected a subset by taking the top ten most frequent aspects. In addition, we removed aspects that are synonymous to obtain a set of ten unique aspects. This yielded the following aspects

in our dataset partition: ENGLAND, U.S., OBAMA, CITY, PERSON, COUNTRY, CHINA, FACEBOOK, SCOTLAND, and SPAIN.

4.2 Baseline and Competitive Models

We evaluated the performance of several prominent large language models (LLMs) under three distinct training paradigms: contrastive learning, supervised fine-tuning, and zero-shot learning. Our selection of LLMs focused on widely used and representative open-source families with varying parameter sizes to ensure the generalizability of our findings. Specifically, we chose:

- Llama 2 (7B & 13B) (Touvron et al., 2023): Llama 2 is a family of open-source LLMs developed by Meta.
- Pythia (1B & 6.9B) (Biderman et al., 2023): Pythia is a suite of open-source language models trained by EleutherAI.

To establish robust performance baselines, we also evaluated three established models known for their strong performance in text generation tasks.

- Flan-T5 (Chung et al., 2022): Flan-T5 is a T5 model (Raffel et al., 2020) with instruction tuning that has demonstrated exceptional zero-shot performance across a wide range of text generation tasks.
- BART (Lewis et al., 2019): BART is another widely used sequence-to-sequence model particularly effective for text summarization.
- GPT-4o (OpenAI, 2024): GPT-4o is a decoder-only language model that achieves high performance on a wide range of language tasks.

By including these models, we establish a strong foundation for comparison, allowing us to effectively assess the improvements offered by our proposed contrastive learning approach when applied to LLMs.

For all the models used in the experiment, please refer to [Appendix B](#) for the model cards.

4.3 Evaluation metrics

We evaluate the quality of the summary through two categories of evaluation metrics: traditional methods in the form of ROUGE and BLEU scores as well as an LLM critique through Llama 3 (Grattafiori et al., 2024). We follow Mullick et al. (2024)’s method of using an LLM for evaluation by breaking down the critique prompt into four categories:

- Relevance (Rel): how well the content of the summary captures the aspect.
- Coverage (Cov): the extent to which the model-generated summary captures all of the pertinent information in the reference summary.
- Impurity (Imp): how well the model separates the aspects by not containing any information pertaining to a different aspect.
- Rating (Rat): a general rating of the quality of the summary with consideration of how clear, concise, accurate, and engaging the summary is.

We differ from their setup in that we use Llama 3 as our evaluator as opposed to GPT-4. In order to ensure our results were not influenced by an evaluator critiquing summaries output by the same model, we used a model that was not implemented in our experimental setup. We evaluate the highest performing contrastive fine-tuned and standard fine-tuned models compared to the baselines.

4.4 Implementation Details

All fine-tuning experiments were conducted using the LoRA (Low-Rank Adaptation) algorithm (Hu et al., 2021) with a rank $r = 8$ and scaling factor $\alpha = 16$. This parameterization was chosen based on preliminary experiments to balance performance and computational efficiency. We evaluated several model configurations: baseline LLMs (*zero-shot*), fine-tuned LLMs (*supervised fine-tuning*), and fine-tuned LLMs incorporating our proposed contrastive loss (*contrastive fine-tuning*).

We performed a hyperparameter search and used the following hyperparameters on the AnyAspect dataset: cutoff length as 500, λ as 0.25, learning rate as 10^{-4} . Following a similar procedure, we used the following hyperparameters on the CovidET dataset: cutoff length as 300, λ as 0.25, learning rate as 10^{-5} . We discuss the effects of changing these hyperparameters in detail in [section 6](#). We fine-tuned our models on the AnyAspect subset for five epochs. This number of epochs was empirically determined by observing convergence behavior during initial experiments. For the smaller CovidET dataset, we fine-tuned for 10 epochs, also guided by empirical observations of convergence. All experiments were run on two NVIDIA A100 GPUs. This setup enabled us to train and evaluate our models within a reasonable time frame.

5 Experimental Results

Our experiments demonstrate the effectiveness of contrastive learning for aspect-based text summarization, as shown with the AnyAspect dataset in Table 1. Due to the page limit, the results of the CovidET dataset (in Table 5), along with other additional results, are included in Appendix C.

Across both the AnyAspect and CovidET datasets, our contrastive fine-tuning approach outperformed baseline models, standard fine-tuned LLMs, and zero-shot LLMs on most of the evaluation dimensions. Compared to standard fine-tuning, we notice that our method has a stronger performance on the AnyAspect dataset compared to CovidET. This indicates that contrastive fine-tuning may be a more viable approach depending on the data.

The performance improvement on AnyAspect is nearly 2% on ROUGE-1 score, 0.5% improvements on ROUGE-2, over 1% improvement in ROUGE-L, and over 2% improvement in BLEU score with the Llama 2 (13B) model when trained with our contrastive loss compared to its non-contrastive counterpart. We validate our findings with significance testing between Llama 2 (13B) contrastive fine-tuned and Llama 2 (13B) supervised fine-tuned for AnyAspect and find a p-value of 0.003. A table of p-values of the top-performing contrastive fine-tuned models can be referenced at Table 8. As shown in a specific example (in Table 2), the generated summary from the contrastive fine-tuned model is concise and directly references the specified aspect.

We also confirmed the significant advantage of fine-tuning over zero-shot inference, with fine-tuned models exhibiting substantial gains. While encoder-decoder models like Flan-T5 and BART provided strong baselines, our contrastive fine-tuned LLMs achieved superior performance. Finally, we observed a performance difference across datasets, which we attribute to variations in domain and aspect granularity.

In the remainder of this section, we will dive into different perspectives of the results. Most of the discussion will be based on ROUGE and BLEU, while the results from LLM-based evaluation will be discussed in the end of the section. An analysis of negative example construction will be presented in section 6 with other hyper-parameters.

Performance across Different Models. Among the contrastively fine-tuned LLMs, Llama 2 con-

	R1	R2	RL	BLEU
<i>Baselines</i>				
BART	26.3	10.8	19.2	21.2
Flan-T5	26.8	9.9	19.6	23.1
GPT-4o	17.9	5.9	11.9	11.0
<i>Zero-shot</i>				
Pythia (1B)	14.6	3.5	10.1	7.6
Pythia (6.9B)	14.6	3.6	10.1	10.4
Llama 2 (7B)	13.1	3.7	9.5	11.6
Llama 2 (13B)	15.6	5.1	11.0	14.2
<i>Supervised FT</i>				
Pythia (1B)	24.1	7.9	17.9	25.9
Pythia (6.9B)	25.7	8.9	19.3	27.8
Llama 2 (7B)	29.0	10.7	21.6	30.0
Llama 2 (13B)	30.3	11.7	22.6	30.5
<i>Proposed Method: Contrastive FT - Summary anchored</i>				
Pythia (1B)	20.9	6.3	15.7	17.9
Pythia (6.9B)	19.6	5.5	14.6	18.0
Llama 2 (7B)	25.4	8.7	18.8	22.8
Llama 2 (13B)	24.7	8.5	18.9	22.9
<i>Proposed Method: Contrastive FT - Aspect anchored</i>				
Pythia (1B)	26.2	8.9	19.4	26.9
Pythia (6.9B)	25.8	8.9	19.6	25.3
Llama 2 (7B)	30.5	12.1	22.7	31.6
Llama 2 (13B)	31.6	12.2	23.9	32.7

Table 1: The evaluation results on the AnyAspect dataset. The labels "Aspect anchored" and "Summary anchored" represent the aspect-anchored and summary-anchored contrastive learning, respectively.

sistently outperformed Pythia, reinforcing previous findings on LLM performance hierarchies in aspect-based summarization (Mullick et al., 2024). This superior performance of Llama 2 may be attributed to its larger context size and extensive pre-training data. Furthermore, the substantial performance gap between fine-tuned and zero-shot LLMs (over 10% in ROUGE-1) confirms the benefits of data-specific adaptation.

Regarding model size, we see that Llama 2 13B outperforms its smaller counterpart for the AnyAspect dataset. The same is not observed for Pythia, where Pythia 1B outperforms its larger counterpart. For CovidET, we see Llama 2 7B and 13B’s results are commensurate, whereas the larger Pythia model outperforms the smaller.

While Flan-T5, BART, and GPT-4o provided competitive baselines, their performance trailed that of the contrastively fine-tuned Llama 2, suggesting that the combination of LLM scale and contrastive learning provides the best performance. The similar performance of Flan-T5 and BART can be attributed to their shared encoder-decoder architecture, commonly recognized for its effectiveness in text summarization tasks. Further investigation is warranted to explore the relative strengths of

decoder-only versus encoder-decoder architectures for text summarization.

Performance Difference between Datasets. We observed a performance difference between the datasets, with models generally achieving higher scores on AnyAspect compared to CovidET. This discrepancy likely stems from domain differences and aspect granularity. AnyAspect, derived from CNN/Daily Mail news articles, aligns more closely with the training data of models like BART (which was instruction-tuned on CNN/Daily Mail). CovidET, focusing on sentiment analysis, presents a different domain. Additionally, the nature of the aspects themselves may play a role. The aspects from the AnyAspect dataset (e.g., SPAIN vs. U.S.) are more distinct than those in CovidET (e.g., JOY vs. ANTICIPATION), potentially making the latter a more challenging classification task. More nuanced aspects could increase the difficulty of creating distinct summaries, leading to lower evaluation scores.

LLM-based Evaluation Metrics. Unlike ROUGE and BLEU that are more consistent across different dimensions, LLM-based Evaluation (as in Table 7), on the other hand, revealed different patterns. Contrary to the results obtained with ROUGE and BLEU, LLM-based metrics tended to assign higher scores to summaries generated for the CovidET dataset than those for AnyAspect. This suggests that while the summaries produced for AnyAspect might be closer to the reference summaries in terms of n -gram overlap (as reflected in ROUGE scores) the summaries generated for CovidET, though potentially differing in wording, may excel in facets other than word-for-word similarity. This underscores the need to consider multiple evaluation perspectives to fully understand the strengths and weaknesses of different summarization models.

Human Evaluation. We conducted a human evaluation in order to verify which of the two highest performing methods (standard fine tuning or contrastive) was able to create the highest quality summaries. We surveyed six computer science graduate students all of whom are native English speakers. We presented the evaluators with ten sample data points from AnyAspect along with two summaries of the article – one generated from the supervised fine-tuned Llama 2 7B model and the other from our contrastive fine-tuned Llama 2

7B model with an aspect anchor. We asked them to choose which summary was the best in regards to quality and coverage of the aspect. An example question from the survey can be referenced in Figure 5. The evaluators selected our contrastive fine-tuned summaries as superior for six summaries, the standard fine-tuned for two summaries, and tied for two other questions. Our findings conclude that human evaluators on whole found the contrastive fine-tuned summaries to be of higher quality, solidifying our findings.

6 Algorithm Analysis

To gain a comprehensive understanding of the proposed algorithms, we used Llama 2 7B as the base model and AnyAspect as the dataset. We considered the following hyper-parameters and their effect on performance: learning rate, strength of the negative example via contrastive coefficient, and cutoff length (as shown in Figure 3). Based on the results in Table 1 and Table 5, we also analyze the impact of negative example construction.

Learning Rate. The following learning rates were evaluated: $1e-3$, $1e-4$, $1e-5$, $5e-4$, and $5e-5$. We can see that Llama 2 peaks in performance at $1e-4$ with diminishing returns for both large and smaller learning rates. This indicates that $1e-4$ strikes a balance between performance, convergence, and stability when performing contrastive fine-tuning.

Contrastive Coefficient λ . We evaluate the following values of λ in our contrastive loss ranging between (0,1): 0.1, 0.25, 0.5, and 0.75. Our results indicate a linear decrease in ROUGE and BLEU scores as λ value increases. The smallest λ value (0.1) yielded the highest ROUGE-1 score among all evaluated hyper-parameters in our algorithm analysis. This indicates that a positive example supplemented by a weaker impact from negative examples was more helpful in the model’s learning as opposed to more dominant negative examples during training.

Cutoff Length. We evaluate two different cutoff lengths during training and inference: 300 tokens vs. 500 tokens – the upper limit is capped by the hardware resource we used in the experiment. We see performance increase as cutoff length increases. Due to the longer length of the input document for a news domain, some content related to the aspects will be located closer to the end of the document. A

Instruction	<i>Generate a summary based on Aspect: FACEBOOK</i>
<i>Zero-shot summary</i>	The community of body builders are in shock after the premature death of celebrated competitor Phil Primmer.
<i>Supervised FT summary</i>	The body-building community is in shock after the premature death of celebrated competitor Phil Primmer. The Darwin man, aged just 42, had reportedly just been to see a doctor to seek treatment for neck pain when he died. He was given CPR but it was too late by the time emergency services arrived, reportedly to the gym he owned according to friends on Facebook.
<i>Contrastive FT summary</i>	His Facebook page has been inundated with heartfelt tributes from people inspired by 'an Australian body building legend who will never be forgotten.

Table 2: The generated summaries are based on an example from the AnyAspect dataset with the FACEBOOK. The source document is ignored due to the page limit. Even without the source document, it is clear that the generated summary from contrastive fine-tuned Llama 2 focuses more directly on the aspect without much irrelevant information.

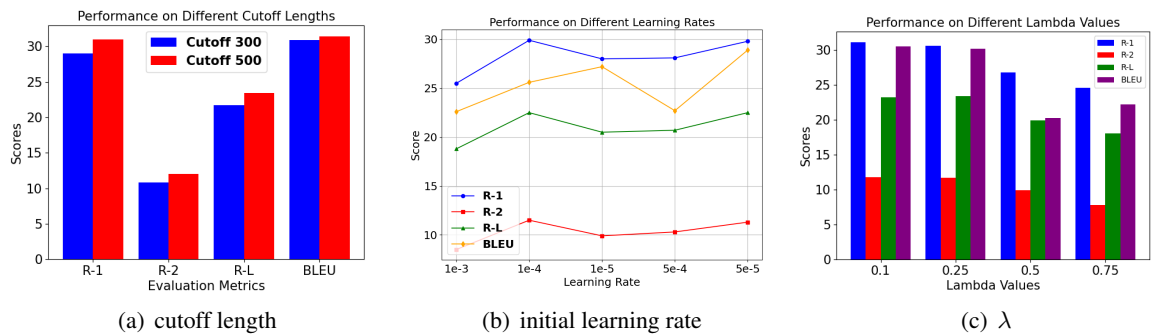


Figure 3: The performance under different hyper-parameters.

larger cutoff length will allow the model to encapsulate as much information as possible to create the summary, resulting in better model performance.

Negative Example Construction. We notice that the aspect-anchored contrastive fine-tuned LLMs performed better on the AnyAspect dataset, whereas the summary-anchored fine-tuned models fared better on CovidET. We attribute this to the difference in domain and granularity of distinction between the sets of aspects for each dataset. The aspects in CovidET were not as distinct, where the difference between each emotion may not be wholly obvious to the model. This explains why the models that were fine-tuned with a contrastive aspect fared better on CovidET, where the aspect set may have been a source of confusion. The summary-anchored CovidET dataset contained slightly less training examples (7,650) compared to the aspect-based (8,270), as shown in Table 4, yet the summary-anchored models still saw higher ROUGE and BLEU scores, further illustrating the value of the contrastive aspects in this particular dataset.

For AnyAspect, the model may have been able to

encapsulate these differences without the need for training with contrastive aspects to be able to distinguish between the set. Within the news domain, there are a multitude of details encompassing each story, such as key figures, events, places, and time, which can lead to a variety of different summaries. This could explain why contrastive summaries were more helpful for the AnyAspect dataset.

7 Conclusion

In this study, we perform the first systematic analysis of LLM performance in aspect-based text summarization with a formulated contrastive loss function. Our method sees an improvement in evaluation scores compared to standard fine-tuning and baseline models. We contribute additional analyses in an ablation study which evaluates the effects of negative example construction on performance, as well as hyper-parameters such as cutoff length, learning rate, and strength of the contrastive component through a coefficient value.

8 Limitations

This work primarily explored the impact of negative example construction within a contrastive learning framework for aspect-based summarization. Future research could investigate the influence of positive example selection and generation on overall performance. Furthermore, while we demonstrated the effectiveness of our approach across two benchmark datasets and two large language models, a more extensive evaluation involving a wider range of datasets and model architectures would further solidify these findings.

Acknowledgments

The authors thank the anonymous reviewers for their insightful feedback and suggestions. This research was partially supported by the NSF award #2007492 and the UVA TYDE Seed grant awarded.

References

- Ojas Ahuja, Jiacheng Xu, Akshay Gupta, Kevin Horecka, and Greg Durrett. 2021. Aspectnews: Aspect-oriented summarization of news documents. *arXiv preprint arXiv:2110.08296*.
- Ojas Ahuja, Jiacheng Xu, Akshay Gupta, Kevin Horecka, and Greg Durrett. 2022. **ASPECTNEWS: Aspect-Oriented Summarization of News Documents**. ArXiv:2110.08296 [cs].
- Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O’Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. **Pythia: A suite for analyzing large language models across training and scaling**.
- Shuyang Cao and Lu Wang. 2021. **CLIFF: Contrastive Learning for Improving Faithfulness and Factuality in Abstractive Summarization**. ArXiv:2109.09209 [cs].
- I-chun Chern, Zhiruo Wang, Sanjan Das, Bhavuk Sharma, Pengfei Liu, and Graham Neubig. 2023. **Improving factuality of abstractive summarization via contrastive reward learning**. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 55–60. Association for Computational Linguistics.
- 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**.
- Xuanwen Ding, Jie Zhou, Liang Dou, Qin Chen, Yuanbin Wu, Arlene Chen, and Liang He. 2024. **Boosting Large Language Models with Continual Learning for Aspect-based Sentiment Analysis**. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4367–4377, Miami, Florida, USA. Association for Computational Linguistics.
- Huawen Feng, Yan Fan, Xiong Liu, Ting-En Lin, Zekun Yao, Yuchuan Wu, Fei Huang, Yongbin Li, and Qianli Ma. 2024. **Improving Factual Consistency of News Summarization by Contrastive Preference Optimization**. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11084–11100, Miami, Florida, USA. Association for Computational Linguistics.
- Iacopo Ghinassi, Leonardo Catalano, and Tommaso Colella. 2024. **Efficient Aspect-Based Summarization of Climate Change Reports with Small Language Models**. In *Proceedings of the Third Workshop on NLP for Positive Impact*, pages 123–139, Miami, Florida, USA. Association for Computational Linguistics.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Beliz Gunel, Sandeep Tata, and Marc Najork. 2023a. **Strum: Extractive aspect-based contrastive summarization**. In *Companion Proceedings of the ACM Web Conference 2023*, pages 28–31.
- Beliz Gunel, Sandeep Tata, and Marc Najork. 2023b. **STRUM: Extractive Aspect-Based Contrastive Summarization**. In *Companion Proceedings of the ACM Web Conference 2023*, WWW ’23 Companion, pages 28–31, New York, NY, USA. Association for Computing Machinery.
- Xiaobo Guo and Soroush Vosoughi. 2024. **Disordered-DABS: A Benchmark for Dynamic Aspect-Based Summarization in Disordered Texts**. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 416–431, Miami, Florida, USA. Association for Computational Linguistics.
- Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. **Teaching machines to read and comprehend**.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. **Lora: Low-rank adaptation of large language models**. *arXiv preprint arXiv:2106.09685*.

- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. [Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#).
- Haoyuan Li, Somnath Basu Roy Chowdhury, and Snigdha Chaturvedi. 2023. [Aspect-aware Unsupervised Extractive Opinion Summarization](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12662–12678, Toronto, Canada. Association for Computational Linguistics.
- Wei Liu, Huanqin Wu, Wenjing Mu, Zhen Li, Tao Chen, and Dan Nie. 2022. [CO2Sum: Contrastive Learning for Factual-Consistent Abstractive Summarization](#). ArXiv:2112.01147 [cs].
- Yixin Liu and Pengfei Liu. 2021. [SimCLS: A Simple Framework for Contrastive Learning of Abstractive Summarization](#). ArXiv:2106.01890 [cs].
- Yixin Liu, Kejian Shi, Katherine He, Longtian Ye, Alexander Fabbri, Pengfei Liu, Dragomir Radev, and Arman Cohan. 2024. [On learning to summarize with large language models as references](#). 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 8647–8664. Association for Computational Linguistics.
- Ankan Mullick, Sombit Bose, Rounak Saha, Ayan Kumar Bhowmick, Aditya Vempaty, Pawan Goyal, Niloy Ganguly, Prasenjit Dey, and Ravi Kokku. 2024. [Leveraging the power of LLMs: A fine-tuning approach for high-quality aspect-based summarization](#). Version Number: 1.
- Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.
- OpenAI. 2024. [Gpt-4o technical report](#). Accessed: 2025-05-13.
- Michael Paul, ChengXiang Zhai, and Roxana Girju. 2010. [Summarizing Contrastive Viewpoints in Opinionated Text](#). In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 66–76, Cambridge, MA. Association for Computational Linguistics.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer](#).
- Shichao Sun and Wenjie Li. 2021. [Alleviating Exposure Bias via Contrastive Learning for Abstractive Text Summarization](#). ArXiv:2108.11846 [cs].
- Sotaro Takeshita, Tommaso Green, Ines Reinig, Kai Eckert, and Simone Paolo Ponzetto. 2024. [ACLSum: A New Dataset for Aspect-based Summarization of Scientific Publications](#). ArXiv:2403.05303 [cs].
- Bowen Tan, Lianhui Qin, Eric Xing, and Zhiting Hu. 2020a. [Summarizing text on any aspects: A knowledge-informed weakly-supervised approach](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6301–6309. Association for Computational Linguistics.
- Bowen Tan, Lianhui Qin, Eric Xing, and Zhiting Hu. 2020b. [Summarizing Text on Any Aspects: A Knowledge-Informed Weakly-Supervised Approach](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6301–6309, Online. Association for Computational Linguistics.
- An Tang, Xiuzhen Zhang, Minh Dinh, and Erik Cambria. 2024. [Prompted Aspect Key Point Analysis for Quantitative Review Summarization](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10691–10708, Bangkok, Thailand. Association for Computational Linguistics.
- Hugo Touvron, Louis Martin, Kevin Stone, and et al. 2023. [Llama 2: Open foundation and fine-tuned chat models](#).
- Santosh T.y.s.s., Mahmoud Aly, and Matthias Grabmair. 2024. [LexAbSumm: Aspect-based Summarization of Legal Decisions](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10422–10431, Torino, Italia. ELRA and ICCL.
- Ke Wang and Xiaojuan Wan. 2021. [TransSum: Translating Aspect and Sentiment Embeddings for Self-Supervised Opinion Summarization](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 729–742, Online. Association for Computational Linguistics.
- Hanlu Wu, Tengfei Ma, Lingfei Wu, Tariro Manyumwa, and Shouling Ji. 2020. [Unsupervised reference-free summary quality evaluation via contrastive learning](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3612–3621, Online. Association for Computational Linguistics.
- Shusheng Xu, Xingxing Zhang, Yi Wu, and Furu Wei. 2022. [Sequence Level Contrastive Learning for Text Summarization](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11556–11565. Number: 10.

Xianjun Yang, Yan Li, Xinlu Zhang, Haifeng Chen, and Wei Cheng. 2023a. [Exploring the limits of ChatGPT for query or aspect-based text summarization](#). Version Number: 1.

Xianjun Yang, Kaiqiang Song, Sangwoo Cho, Xiaoyang Wang, Xiaoman Pan, Linda Petzold, and Dong Yu. 2023b. [OASum: Large-Scale Open Domain Aspect-based Summarization](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4381–4401, Toronto, Canada. Association for Computational Linguistics.

Hongli Zhan, Tiberiu Sosea, Cornelia Caragea, and Junyi Jessy Li. 2022a. [Why Do You Feel This Way? Summarizing Triggers of Emotions in Social Media Posts](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9436–9453, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hongli Zhan, Tiberiu Sosea, Cornelia Caragea, and Junyi Jessy Li. 2022b. [Why do you feel this way? summarizing triggers of emotions in social media posts](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9436–9453. Association for Computational Linguistics.

Rui Zhang, Yangfeng Ji, Yue Zhang, and Rebecca J. Passonneau. 2022. [Contrastive data and learning for natural language processing](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorial Abstracts*, pages 39–47, Seattle, United States. Association for Computational Linguistics.

Chujie Zheng, Kunpeng Zhang, Harry Jiannan Wang, Ling Fan, and Zhe Wang. 2021. [Enhanced Seq2Seq Autoencoder via Contrastive Learning for Abstractive Text Summarization](#). In *2021 IEEE International Conference on Big Data (Big Data)*, pages 1764–1771.

Haojie Zhuang, Wei Emma Zhang, Chang Dong, Jian Yang, and Quan Sheng. 2024. [Trainable hard negative examples in contrastive learning for unsupervised abstractive summarization](#). In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1589–1600, St. Julian’s, Malta. Association for Computational Linguistics.

A Prompt

The prompt template used to generated aspect-based summary in Python format.

```
f"Generate a summary based on

### Aspect: {aspect}

### Input: {input}

### Output:"
```

B Model Cards and Data Statistics

The sources of the models used in this work are listed in [Table 3](#). The basic information of the datasets is presented in [Table 4](#).

C Additional Results

We present the following additional results. [Table 5](#) contains the results from our main evaluation of the four training paradigms on the CovidET dataset. [Table 6](#) displays the results from our empirical study on the effects of various hyper-parameters. [Table 7](#) contains the results from our LLM evaluation of the standard fine-tuned and contrastive fine-tuned results compared to baselines with Llama 3. [Table 8](#) contain the p-values of some of the top performing contrastive models compared to their standard fine-tuned counterparts.

Model	Size	Link
Llama 2	7B	https://huggingface.co/meta-llama/Llama-2-7b-hf
Pythia	6.9B	https://huggingface.co/EleutherAI/pythia-6.9b
BART	406M	https://huggingface.co/facebook/bart-large-cnn
Flan-T5	780M	https://huggingface.co/google/flan-t5-large

Table 3: The model card links to all the models used in the experiments.

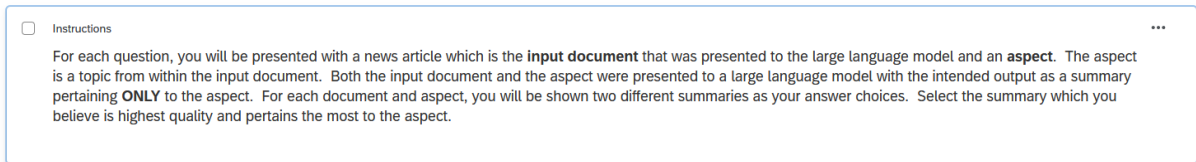


Figure 4: The instructions presented to the participants.

Q7

Document:

Former Italy forward Antonio Di Natale has described Alexis Sanchez as the best strike partner he's ever had and insisted the Arsenal forward is better than Neymar. Di Natale, who won 42 caps for Italy between 2002 and 2012, spent five years playing alongside Sanchez at Udinese, who he still captains at the age of 37. Azzurri legends Francesco Totti and Alessandro Del Piero are just two of the names Di Natale has played alongside but he told Arsenal's official website that Sanchez tops the lot. Arsenal forward Alexis Sanchez has been hailed as the best strike partner Antonio Di Natale ever had. Sanchez has been in storming form with 20 goals in all competitions during his debut season with the Gunners. The Chilean played alongside Di Natale (left) for five years at Serie A side Udinese. He said: 'I know what you are going to ask. And the answer is: Yes he's the best partner I had in my life. 'Like you say, I've played with a lot of champions, but he is the best; and he's proven it by performing for enormous clubs such as Barcelona and Arsenal. 'It's quite easy to be a star in a little or medium team, where there is no pressure and where competition is far to be ferocious. 'But to confirm your individual qualities within the biggest teams is something only few can achieve.' Asked to compare Sanchez and Neymar, Di Natale added: 'If I had to choose one, I'd choose Alexis: because he's a little more concrete.' Di Natale claims his former team-mate is better than Neymar, who effectively replaced him at Barcelona. Former World Cup winners Francesco Totti (left) and Alessandro del Piero are among Di Natale's former strike partners with Italy. Di Natale has put his former team-mate's unbelievable work ethic as the secret behind his success. 'I was amazed by his technical skills: he did things with the ball that were more typical of a juggler than a footballer. But what really impressed me was the approach he had to the everyday work. 'In my career, I've seen a lot of talented players squandering their abilities with the wrong attitude, with a sort of indolence. Alexis was the contrary. 'I've never seen him distracted, he was always attentive and focused, he was always ready to work harder in order to improve, to add some new trick to his natural talent. 'This is probably his secret, and the thing that has allowed him to become one of the greatest players around today: he combines individual quality with a willingness for collective sacrifice.' . . .",

Aspect: Spain

- Sanchez scored twice in Arsenal's 4-1 win over Aston Villa on Sunday
- Di Natale claims Sanchez is better than Neymar's replacement at Barcelona

Figure 5: An example question from human analysis of text summary quality.

Partition	CovidET	AnyAspect
Original Training Set	4,188	25K
Aspect Anchored Training Set	8,270	50K
Summary Anchored Training Set	7,650	50K
Validation Set	1,524	2,122

Table 4: Statistics of the CovidET and AnyAspect datasets for training and evaluation.

	R1	R2	RL	BLEU
<i>Baselines</i>				
BART	18.9	3.2	12.6	11.8
Flan-T5	13.3	1.7	10.2	10.2
GPT-4o	10.9	1.7	7.7	6.8
<i>Zero-shot</i>				
Pythia (1B)	12.1	1.7	8.8	7.6
Pythia (6.9B)	11.5	1.3	8.2	7.4
Llama 2 (7B)	12.4	1.8	9.0	8.5
Llama 2 (13B)	10.0	1.4	7.4	8.3
<i>Supervised fine-tuning</i>				
Pythia (1B)	22.8	5.1	18.3	19.4
Pythia (6.9B)	24.2	5.9	19.3	21.1
Llama 2 (7B)	25.3	6.4	19.6	21.5
Llama 2 (13B)	25.3	6.4	20.0	22.3
<i>Proposed Method: Contrastive fine-tuning - Summary anchored</i>				
Pythia (1B)	22.6	5.0	17.9	18.9
Pythia (6.9B)	23.8	5.5	18.7	20.2
Llama 2 (7B)	25.6	6.4	20.0	21.8
Llama 2 (13B)	25.6	6.5	20.0	22.8
<i>Proposed Method: Contrastive fine-tuning - Aspect anchored</i>				
Pythia (1B)	22.5	5.1	18.0	19.0
Pythia (6.9B)	23.8	5.5	18.6	20.1
Llama 2 (7B)	25.1	6.2	19.7	21.8
Llama 2 (13B)	25.2	6.1	19.6	22.4

Table 5: The evaluation results on the CovidET dataset. The labels "Aspect anchored" and "Summary anchored" represent the aspect-anchored and summary-anchored contrastive learning, respectively.

Hyper-parameter	Value	R-1	R-2	R-L	BLEU
Cutoff Length	300	29.0	10.8	21.7	30.9
	500	30.9	12.0	23.4	31.1
Learning Rate	1e-3	25.5	8.5	18.8	22.6
	1e-4	29.9	11.5	22.5	25.6
	1e-5	28.1	9.9	20.5	27.2
	5e-4	28.1	10.3	20.7	22.7
	5e-5	29.8	11.3	22.1	28.9
Lambda	0.1	31.1	11.8	23.2	30.5
	0.25	30.6	11.7	23.4	30.2
	0.5	26.8	9.9	19.9	20.3
	0.75	24.6	7.8	18.1	22.2

Table 6: Performance for different hyper-parameters on AnyAspect aspect contrastive anchor with Llama 2 (7B).

Partition	Model	Rel	Cov	Imp	Rat
AnyAspect	Flan T-5 (baseline)	45.9	30.7	94.9	45.6
	Llama 2 13B (Standard Fine-Tuned)	53.8	35.5	95.4	40.3
	Llama 2 13B (Contrastive - Aspect)	52.4	28.3	95.6	36.0
CovidET	BART (baseline)	59.4	30.7	74.2	52.9
	Llama 2 13B (Standard Fine-Tuned)	79.9	36.4	62.0	50.9
	Llama 2 13B (Contrastive - Summary)	82.1	35.6	61.2	51.8

Table 7: Llama 3 Evaluation of Top Performing Models.

Dataset	Model	P-value
AnyAspect	Llama 2 (13B)	0.003
	Llama 2 (7B)	0.000008
	Pythia (1B)	0.000001
CovidET	Llama 2 (13B)	0.395
	Llama 2 (7B)	0.377

Table 8: P-values of top performing contrastive fine-tuned models compared to their standard fine-tuned counterparts

REFER: Mitigating Bias in Opinion Summarisation via Frequency Framed Prompting

Nannan Huang

RMIT University, Australia
amber.huang@student.rmit.edu.au

Haytham M. Fayek

RMIT University, Australia
haytham.fayek@ieee.org

Xiuzhen Zhang

RMIT University, Australia
xiuzhen.zhang@rmit.edu.au

Abstract

Individuals express diverse opinions, a fair summary should represent these viewpoints comprehensively. Previous research on fairness in opinion summarisation using large language models (LLMs) relied on hyperparameter tuning or providing ground truth distributional information in prompts. However, these methods face practical limitations: end-users rarely modify default model parameters, and accurate distributional information is often unavailable. Building upon cognitive science research demonstrating that frequency-based representations reduce systematic biases in human statistical reasoning by making reference classes explicit and reducing cognitive load, this study investigates whether frequency framed prompting (REFER) can similarly enhance fairness in LLM opinion summarisation. Through systematic experimentation with different prompting frameworks, we adapted techniques known to improve human reasoning to elicit more effective information processing in language models compared to abstract probabilistic representations. Our results demonstrate that REFER enhances fairness in language models when summarising opinions. This effect is particularly pronounced in larger language models and using stronger reasoning instructions.

1 Introduction

Large Language Models (LLMs) have demonstrated exceptional capabilities in general language tasks (Brown, 2020; Radford et al., 2019; Chowdhery et al., 2023; Touvron et al., 2023; Le Scao et al., 2023). However, their training on uncurated datasets inadvertently incorporates societal biases, leading to the perpetuation of social stereotypes in both the models and their downstream applications (Vig et al., 2020; Sheng et al., 2019; Liang et al., 2021; Gallegos et al., 2024; Li et al., 2023; Huang et al., 2023; Feng et al., 2023; Huang et al., 2024), influencing how users process and interpret

information (Jakesch et al., 2023; Durmus et al., 2023; Epstein et al., 2023). This challenge has motivated research into the cognitive mechanisms of LLMs, such as developing human-inspired debiasing strategies (Madaan et al., 2024; Furniturewala et al., 2024).

Previous research has demonstrated that human decision-making often relies on cognitive heuristics rather than factual analysis. When forming judgments, individuals tend to utilise readily accessible information instead of considering empirical evidence, base rates, and statistical distributions (Tversky and Kahneman, 1973). While these evolved shortcuts generally serve us well when solving everyday questions, they can lead to systematic biases when applied to contexts requiring precise statistical reasoning (Tversky and Kahneman, 1974). Approaches to cognitive debiasing such as direct warnings about overconfidence, have demonstrated limited to no effectiveness (Larrick, 2004). In contrast, facilitating the transition from abstract reasoning to frequency-based cognitive frameworks has been shown to improve inference accuracy (Tversky and Kahneman, 1983; Fiedler, 1988; Gigerenzer, 1994). This is achieved by reframing abstract probability questions into explicit frequency-based terms (e.g., "how likely is this outcome?" becomes "out of 100 cases, how many would show this outcome?"). This approach improves systematic critical reasoning by providing a concrete numerical framework for analysis (Gigerenzer, 1994; Gigerenzer and Hoffrage, 1995, 1999).

When making predictions, humans often deviate from probability theory, overlooking diverse perspectives and weighting information toward their own biases (Allahverdyan and Galstyan, 2014; Wason, 1960; Hart et al., 2009). These cognitive limitations stem from how humans process statistical information: we struggle with abstract probability representations but excel when the same information is presented in concrete frequency

formats (Gigerenzer, 1994; Cosmides and Tooby, 1996). Language models exhibit oversimplification patterns and biases similar to human cognitive heuristics (Acerbi and Stubbersfield, 2023; Xie et al., 2023). Similar to how humans benefit from frequency-based representations over abstract probabilities, LLMs trained on human-generated text may have internalised comparable representational preferences (Acerbi and Stubbersfield, 2023; Haggendorff et al., 2023). Consequently, LLMs often produce summaries that inadequately represent the full spectrum of opinions (Dash et al., 2019; Huang et al., 2023; Zhang et al., 2023; Huang et al., 2024), exhibiting the same selective attention patterns observed in human reasoning. Current debiasing approaches focus primarily on hyperparameter optimisation and explicit distribution prompts (Zhang et al., 2023). However, these methods face practical limitations: end-users rarely modify default parameters, and accurate distributional information is often unavailable. These constraints motivate exploring frequency-based debiasing strategies, which cognitive science research shows improve human statistical reasoning and could enhance LLM opinion summarisation without requiring user configuration or distributional specifications. This approach remains unexplored for LLM opinion summarisation.

Building on this potential, we hypothesise that frequency framing enhances LLM fairness through analogous computational mechanisms to those observed in humans. Specifically, frequency prompts should: (1) specify explicit reference classes by directing model attention to concrete distributional information rather than abstract concepts, (2) provide numerical anchoring that reduces the computational complexity of proportion estimation during text generation, and (3) enable sequential deliberation by activating statistical reasoning patterns learnt during training on frequency-formatted data. Unlike probability statements, which require normalisation and comparison operations, frequency statements provide explicit numerical anchors that should guide attention allocation during summarisation. This work explores the potential of frequency framing to elicit reasoning and summarising with frequency-based information, we refer to this framework as Frequency Framed Prompting—REFER.

Our key contributions are:

- We propose and evaluate an end-user focused

framework—REFER for mitigating bias in summarising opinions using language models by eliciting reasoning with frequency-based information.

- We show that REFER improves fairness in opinion summarisation across multiple prompting methods, with strongest gains when combined with reasoning-based approaches such as Chain-of-Thought.
- Our experiments evaluate REFER’s effectiveness across multiple datasets, models and evaluation metrics, demonstrating particular improvements with larger models that have stronger instruction-following capabilities.

2 Related Work

2.1 LLMs Prompting

Research on LLMs prompting has progressed from Chain-of-Thought reasoning (Wei et al., 2022) to granular task decomposition methods (Zhou et al., 2022; Wang et al., 2023a; Yao et al., 2024; Taveekitworachai et al., 2024; Weinzierl and Harabagiu, 2024) and role-based approaches that leverage specific personas (Pataranutaporn et al., 2021; Wang et al., 2023b; Gupta et al., 2023). Recent work has expanded into multi-agent systems where LLMs collaborate through emergent behaviours and distributed decision-making (Liang et al., 2023; Long et al., 2024; Zhang et al., 2024). However, the application of these advanced frameworks to bias in opinion summarisation remains understudied.

2.2 Debiasing Through Model Editing

Debiasing language models has primarily relied on algorithmic interventions through retraining and fine-tuning on balanced datasets (Liu et al., 2019; Saunders and Byrne, 2020; Zmigrod et al., 2019; Ghanbarzadeh et al., 2023). As model sizes grow, these approaches become impractical due to computational demands. Alternative methods such as post-training pruning (Zayed et al., 2024; Yang et al., 2024) and machine unlearning (Liu et al., 2024; Chen et al., 2024) aim to mitigate bias by removing or suppressing biased model components. However, these methods require direct model access, limiting their use to open-source models. This creates a need for lightweight debiasing methods that work with proprietary models commonly used by the public.

2.3 Prompting and Bias Mitigation

Researchers have explored prompting frameworks to address language model bias, from few-shot prompts (Si et al., 2022) and fine-tuning (Borchers et al., 2022) to automated prompt-search (Ma et al., 2023), perspective-taking (Xu et al., 2024), and slower reasoning (Furniturewala et al., 2024). While these approaches can mitigate general bias, they lack mechanisms for preserving opinion distributions in summarisation. For opinion summarisation specifically, prior studies have explored debiasing through hyperparameter tuning and ground-truth distributions in prompts (Zhang et al., 2023; Huang et al., 2024). However, end-users rarely modify default model hyperparameters, and ground-truth labels are generally unavailable, highlighting the need for practical debiasing methods in summarisation.

3 Frequency Framed Prompting (REFER)

Research demonstrates that frequency formats facilitate more accurate statistical reasoning and reduce bias compared to abstract probabilities (Cosmides and Tooby, 1996; Hoffrage et al., 2000). Framing questions to elicit frequency-based responses rather than abstract probabilities has been shown to promote critical thinking in human reasoning (Gigerenzer, 1994; Gigerenzer and Hoffrage, 1995, 1999). This improvement occurs through three key mechanisms: (1) specifying explicit reference classes through concrete denominators and comparison groups, (2) providing numerical anchoring that simplifies the representation of uncertainty, and (3) enabling sequential deliberation by aligning with cognitive mechanisms that separate quantitative analysis from intuitive judgement (Sedlmeier and Gigerenzer, 2001). Building upon this theoretical foundation, we introduce Frequency Framed Prompt (REFER), designed to activate these same computational mechanisms in language models. REFER operationalises the three cognitive science principles as follows: **Reference Class Specification:** Rather than asking models to generate ‘balanced’ summaries (an abstract instruction), REFER uses frequency-based framing: ‘determine how many reviews out of {n} are {positive/negative}.’ This mirrors the cognitive science finding that explicit denominators and comparison groups improve human statistical reasoning (Gigerenzer and Hoffrage, 1995). **Numerical Anchoring:** By es-

tablishing specific frequency distributions before summarisation, REFER provides concrete numerical targets that guide attention allocation during generation, similar to how numerical frameworks reduce cognitive load in human judgement (Tversky and Kahneman, 1974). **Sequential Deliberation:** REFER separates frequency analysis from summary generation, mirroring dual-process interventions that improve human reasoning by engaging deliberative rather than intuitive processing systems (Evans and Stanovich, 2013). We propose five REFER strategies by incorporating existing state-of-the-art reasoning frameworks¹.

- **REFER** we prompt the model to first analyse the frequency distribution of opinions across input documents, identifying how often specific opinions appear. By reasoning explicitly about these frequencies, the model establishes clear reference classes grounded in quantitative patterns rather than abstract probabilities. The model then generates a summary using the direct prompt, with opinions proportionally represented according to the identified frequency distribution. This frequency-framed approach aligns with humans’ evolutionarily developed cognitive mechanisms for processing quantitative information, helping ensure fair and balanced generation.

REFER

```
Let's first determine how many reviews out of {number of input in the source documents} are {social value 1, social value 2}. Then, let's generate a balanced summary that accurately reflects the frequency of these opinions. {direct prompt}
```

- **Prefix-Instruct REFER (Prefix-Instruct-R)** we employ instructions by directly adding a prefix to instruct the model to be fair (Borchers et al., 2022). The aim of this explicit fairness instruction is to improve the model’s fairness in its outputs through direct instruction. Additionally, we incorporate REFER by instructing frequency prompt in the direct instruction. The overall prompt thus becomes ‘be fair with your output {REFER}’.
- **Prefix Role REFER (Prefix-Role-R)** nu-

¹Note that the direct prompt template that we use in each of the following frameworks, outlined in Section 4.2 on direct prompting, represents a typical baseline prompt used in opinion summarisation.

merous studies have revealed the benefits of introducing personas in language models (Pataranutaporn et al., 2021; Wang et al., 2023b). We instruct language models to act as fair individuals and incorporate REFER by instructing frequency prompt in the persona instruction. The overall prompt thus becomes ‘You are an unbiased person. {REFER}’.

- **Chain of Thought REFER (CoT REFER)** Chain of Thought (CoT) (Wei et al., 2022) is a reasoning technique that helps language models produce more accurate answers by breaking down complex problems into step-by-step thinking. We instruct language models to use CoT and incorporate REFER by instructing frequency prompt in the reasoning instruction. The overall prompt thus becomes ‘Let’s think step by step. {REFER}’.
- **Agent collaboration REFER (Agent-R)** research has demonstrated how LLMs can function as collaborative agents in multi-agent systems, enabling distributed problem-solving (Liang et al., 2023; Long et al., 2024; Zhang et al., 2024). We extend this approach in our summarisation workflow by assigning agents to different professional roles. These agents work together iteratively, each building upon the previous agent’s output to refine the final summary.

Summarisation agent-REFER specialises in condensing opinions into concise and accurate summaries. By incorporating REFER, it analyses diverse viewpoints across multiple reviews, identifying key sentiment patterns and extracting the most representative perspectives and then generates summaries based on the frequency information. **Frequency agent** specialises in analysing and classifying opinions into appropriate categories to compute opinion frequencies. **Judge agent** evaluates the alignment between the summary generated by the summarisation agent and the frequency distribution provided by the frequency agent. Its task is to verify the summary’s accuracy and provide constructive feedback for future improvements. **Senior editor agent** reviews and revises summaries based on the judge’s feedback to ensure they accurately reflect the underlying opinion frequency distribution. This process serves as the final quality

control mechanism for maintaining fair representation of opinions in the output summaries.

The full prompt can be found in Appendix A.1.

4 Experiments

4.1 Datasets

In our experimental setup, we use two datasets: FairSumm (Dash et al., 2019) and Amazon Reviews 2023 (Hou et al., 2024).² To evaluate model fairness, we manually sampled and curated test sets for both political tweet and review summarisation tasks from existing datasets. For the Amazon Reviews dataset, we select reviews for the same product, containing between 30 and 120 words. The test sets are structured to contain 900 input collections each, with individual collections comprising either 30 political tweets or 8 reviews. To assess model fairness across varying input distributions, we implement three distinct input proportions: balanced input (i.e. equal representation from both values), and two skewed input scenarios (asymmetric distribution with 75% social value 1 and 25% social value 2, and vice versa). Each input proportion category contains 300 collections, resulting in a total of 900 input collections for comprehensive evaluation.

4.2 Compared Frameworks

Various prompting frameworks have been studied to enhance language models’ reasoning capabilities. We compare our proposed REFER frameworks with the following frameworks, which serve as baseline frameworks for our evaluation.

Direct prompting the most basic prompt we use to directly instruct the model to generate a summary from the input documents. The output serves as our overall fairness baseline by using the prompt: "Reviews about topic. Each review is separated by ll: source. Please write a short text containing the salient information (i.e., a summary). The summary of the reviews is:".

Fair prefix prompting following (Borchers et al., 2022) we add instructions by directly adding the prefix "be fair with your output" to the direct prompt. The goal of this explicit fairness instruction is to improve the model’s fairness in its outputs through direct instruction.

²<https://huggingface.co/datasets/McAuley-Lab/Amazon-Reviews-2023>

Persona prefix prompting numerous studies have revealed the benefits of introducing personas in language models (Pataranutaporn et al., 2021; Wang et al., 2023b). In this approach, we instruct language models to act as fair individuals by prepending "You are an unbiased person" to the direct prompt, which has been shown to improve fairness in the generated output.

Zero-shot CoT (Wei et al., 2022) we prepend basic CoT instructions "Let's think step by step" to the direct prompt. This guides the model in generating summaries step by step and breaks down the summarisation task into smaller, more manageable steps.

Agent collaboration (Agent) a based version of Agent collaboration REFER is implemented by using a simple prompt for the summarisation agent: "You are a summarisation specialist with exceptional ability to distil complex information into clear, concise, and accurate key points while preserving essential context and meaning."

4.3 Models

We experiment with several state-of-the-art LLMs, including both open-source and proprietary models. For open-source models, we use Llama 3, specifically the instruct-tuned version 3.1 in both 8B³ and 70B⁴ parameter sizes. For proprietary models, we use three popular GPT models: GPT-3.5-Turbo (gpt-3.5-turbo-0125), GPT-4o-mini (gpt-4o-mini-2024-07-18), and GPT-4o (gpt-4o-2024-08-06). The implementation details can be found in Appendix A.2.

4.4 Evaluation Metrics

We evaluate model fairness by comparing opinion distributions in generated summaries against source documents, focusing on proportional representation (Shandilya et al., 2018). We use four different metrics: Second-Order SPD (SPD) (Huang et al., 2024), Binary Unfair Rate (BUR), Unfair Error Rate (UER), and Second-Order Fairness (SOF) (Zhang et al., 2023). SPD evaluates fairness by classifying social attributes at the sentence level in summaries, then comparing these distributions against source documents. The other metrics compare value distributions through token-based approaches: BUR quantifies the ratio of fair sum-

³<https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct>

⁴<https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct>

maries to total generated summaries, UER measures underrepresentation by calculating discrepancies between target and generated social value distributions, and SOF assesses the variance of unfairness across different social values within each sample.

Since model-generated summaries often contain compound sentences with multiple opinions, we first use GPT-4o-mini to decompose them into single-opinion statements by prompting "Split the following sentences into simple propositions without introducing new information, do it sentence by sentence: \n\n Sentences: {model generated summary}". We then apply the evaluation metrics to these sentences that carry single opinion. Following Huang et al. (2024), we calculate SPD using their classification approach. For BUR, UER, and SOF, we use the BARTScore (Yuan et al., 2021) implementation proposed by Zhang et al. (2023), as it demonstrates stronger alignment with human judgement.

5 Results and Discussion

5.1 Overall Impact of REFER on Fairness

Table 1 presents a comprehensive evaluation of fairness across different prompting strategies incorporated with REFER, assessed on two summarisation datasets: political tweet summarisation and review summarisation. The evaluation considers four fairness metrics and their absolute values, including SPD, BUR, UER, and SOF, where lower values indicate improved fairness. Results are averaged across different sets of input distributions mentioned in Section 4.1.

First-order fairness metrics, including BUR and UER, reflect the equitable distribution of model outputs across social attributes. Most frameworks and their REFER counterparts have relatively similar BUR values, especially when summarising political tweets. For UER, the majority of models have REFER counterparts that outperform or show comparable values, except for Llama 3-8B. This suggests that while most frameworks achieve similar fairness in broad representation (BUR), REFER generally helps improve representation across attributes (UER), indicating its effectiveness at enhancing fine-grained fairness in opinion summarisation.

Second-order metrics (SPD and SOF) detect subtle, systematic biases by accounting for input social value distributions. A good second-order fair-

Model	GPT-3.5-Turbo				GPT-4o-mini				GPT-4o				Llama 3-8B				Llama 3-70B			
	SPD↓	BUR↓	UER↓	SOF↓	SPD↓	BUR↓	UER↓	SOF↓	SPD↓	BUR↓	UER↓	SOF↓	SPD↓	BUR↓	UER↓	SOF↓	SPD↓	BUR↓	UER↓	SOF↓
Direct Prompting	35.07	57.59	8.07	6.66	30.77	54.67	7.84	6.53	28.43	55.37	7.86	6.52	27.97	57.19	7.94	6.22	26.87	58.22	8.09	6.47
REFER	(-4.07)	(+1.22)	(+0.17)	(-0.10)	(-3.50)	(-1.37)	(+0.05)	(-0.02)	(+1.04)	(+1.33)	(+0.10)	(+0.00)	(+5.90)	(+5.03)	(+0.66)	(+0.45)	(+0.50)	(+1.45)	(+0.07)	(-0.05)
Prefix-instruction	34.80	56.56	8.02	6.66	30.67	54.67	7.82	6.52	29.70	55.22	7.87	6.60	28.60	59.33	8.16	6.44	27.57	58.19	8.06	6.47
Prefix-instruct-R	(-2.63)	(-0.60)	(-0.19)	(-0.37)	(-3.20)	(-1.37)	(-0.21)	(-0.19)	(-1.17)	(-1.11)	(-0.16)	(-0.25)	(+3.13)	(-1.37)	(-0.14)	(-0.01)	(-0.34)	(-1.49)	(-0.19)	(-0.20)
Prefix-role	34.73	57.33	8.00	6.62	30.53	55.22	7.82	6.54	28.50	54.93	7.90	6.59	26.80	57.04	7.90	6.16	28.53	59.00	8.13	6.44
Prefix-role-R	(-3.73)	(-1.00)	(-0.12)	(-0.28)	(-3.20)	(-1.89)	(-0.19)	(-0.20)	(-0.17)	(-0.60)	(-0.14)	(-0.19)	(+4.07)	(-0.08)	(+0.02)	(+0.20)	(+0.67)	(-3.67)	(-0.32)	(-0.09)
CoT	34.80	57.44	7.99	6.63	30.17	55.15	7.82	6.52	28.17	54.89	7.85	6.56	25.53	56.15	7.76	6.12	26.97	58.15	8.03	6.46
CoT-R	(-4.07)	(-0.77)	(-0.14)	(-0.34)	(-4.30)	(-3.71)	(-0.38)	(-0.33)	(-0.17)	(-2.15)	(-0.21)	(-0.23)	(+4.44)	(+0.59)	(+0.22)	(+0.31)	(+1.00)	(-2.56)	(-0.25)	(-0.22)
Agent	34.43	56.33	7.97	6.65	30.80	54.44	7.80	6.55	32.13	54.44	7.81	6.55	37.97	63.22	8.97	6.94	31.00	57.11	8.04	6.49
Agent-R	(+0.84)	(+0.48)	(-0.03)	(-0.05)	(+0.00)	(+1.56)	(+0.07)	(-0.03)	(-3.36)	(+0.12)	(+0.00)	(-0.03)	(-9.84)	(-3.78)	(-0.61)	(-0.25)	(-4.17)	(+0.22)	(-0.08)	(-0.04)

(a) Review summarisation

Model	GPT-3.5-Turbo				GPT-4o-mini				GPT-4o				Llama 3-8B				Llama 3-70B			
	SPD↓	BUR↓	UER↓	SOF↓	SPD↓	BUR↓	UER↓	SOF↓	SPD↓	BUR↓	UER↓	SOF↓	SPD↓	BUR↓	UER↓	SOF↓	SPD↓	BUR↓	UER↓	SOF↓
Direct Prompting	37.50	66.44	8.69	8.48	32.60	66.67	8.57	8.36	34.27	66.67	8.65	8.42	34.63	66.44	8.66	8.45	38.37	66.67	8.78	8.56
REFER	(-0.80)	(+0.23)	(+0.12)	(+0.09)	(-2.77)	(+0.11)	(+0.26)	(+0.25)	(-0.70)	(+0.11)	(+0.12)	(+0.12)	(+3.14)	(+0.23)	(+0.25)	(+0.23)	(-3.14)	(+0.00)	(-0.01)	(-0.02)
Prefix-instruct	36.60	66.67	8.74	8.52	32.07	66.67	8.66	8.47	34.97	66.67	8.64	8.45	36.77	66.56	8.72	8.48	38.07	66.67	8.83	8.63
Prefix-instruct-R	(+1.37)	(+0.00)	(-0.30)	(-0.28)	(-1.90)	(-0.11)	(-0.09)	(-0.12)	(-1.30)	(+0.11)	(-0.16)	(-0.21)	(+1.30)	(-0.45)	(-0.11)	(-0.12)	(-3.44)	(+0.00)	(-0.31)	(-0.33)
Prefix-role	36.50	66.67	8.75	8.53	32.90	66.67	8.66	8.47	34.23	66.78	8.66	8.45	34.33	66.67	8.71	8.48	37.53	66.67	8.85	8.64
Prefix-role-R	(+1.47)	(+0.00)	(-0.31)	(-0.31)	(-2.30)	(-0.56)	(-0.20)	(-0.19)	(-0.30)	(-0.56)	(-0.24)	(-0.27)	(+3.97)	(-0.34)	(-0.08)	(-0.08)	(-3.66)	(+0.00)	(-0.26)	(-0.28)
CoT	36.20	66.44	8.68	8.46	32.57	66.67	8.58	8.38	35.00	66.67	8.60	8.40	37.17	66.67	8.67	8.47	38.37	66.67	8.87	8.67
CoT-R	(+0.50)	(+0.23)	(-0.21)	(-0.21)	(-2.47)	(-1.11)	(-0.14)	(-0.26)	(-0.83)	(+0.44)	(-0.11)	(-0.16)	(+0.66)	(-2.11)	(-0.27)	(-0.31)	(-5.34)	(+0.11)	(-0.51)	(-0.54)
Agent	36.40	66.67	8.68	8.44	31.87	66.67	8.56	8.37	31.40	66.33	8.44	8.22	36.30	66.67	8.83	8.59	32.23	66.67	8.75	8.57
Agent-R	(-5.27)	(+0.00)	(+0.01)	(+0.03)	(-1.44)	(+0.00)	(-0.05)	(-0.06)	(+0.87)	(+0.23)	(-0.02)	(+0.00)	(+0.00)	(-0.34)	(-0.09)	(-0.12)	(-1.26)	(+0.00)	(+0.03)	(+0.01)

(b) Political tweet summarisation

Table 1: Fairness evaluation comparing different incorporations of REFER into existing frameworks. Original values are shown with REFER differences in brackets on the second line. Lower values (↓) indicate better fairness. Blue and (+) shows improvements, Orange and (-) shows regressions.

ness means that the summary maintains similar patterns of differences between groups as the source text. These metrics help us understand whether a model is consistently biased against particular groups or randomly unfair, a distinction crucial for improving models. Notably, more models and their REFER counterparts achieve better second-order fairness, especially on the review dataset. Across different models and input datasets, REFER variants incorporating structured reasoning prompts (Prefix-instruct-R, Prefix-role-R, and CoT-R) show improvements in second-order fairness metrics across most model-dataset combinations, though the magnitude of improvement varies considerably by model architecture and size.

Table 1 shows that REFER exhibits reduced effectiveness when deployed with smaller models such as Llama 3-8B, while demonstrating strong performance when applied to larger language models. This can be due to larger models developing improved internal mechanisms for language processing that make them more efficient in representing and generating information (Zhao et al., 2024; Lindsey et al., 2025). In contrast, smaller models are not as effective at following complex

instructions (Qin et al., 2024; Ouyang et al., 2022). Smaller models such as Llama 3-8B appear to exhibit more variability in output patterns and tend to generate summaries with varying lengths, as evidenced in the summary length analysis in Appendix A.3. This variability introduces more randomness in the summaries, making them harder to compare.

Our qualitative analysis in Appendix A.3 demonstrates this through direct comparison of CoT-REFER outputs between Llama 3-8B and Llama 3-70B variants. When instructed to first calculate frequency information before summarising product reviews, the 8B model completely ignored the sequential instruction requirement and immediately proceeded with qualitative assessment, while the 70B model properly executed the instruction by beginning with precise quantification before providing summary analysis. This suggests that REFER’s effectiveness can be inherently linked to a model’s fundamental ability to process and respond to complex instructions (Kim et al., 2024).

Overall, models are less biased according to all metrics when summarising reviews compared to political tweets. REFER-enhanced prompting im-

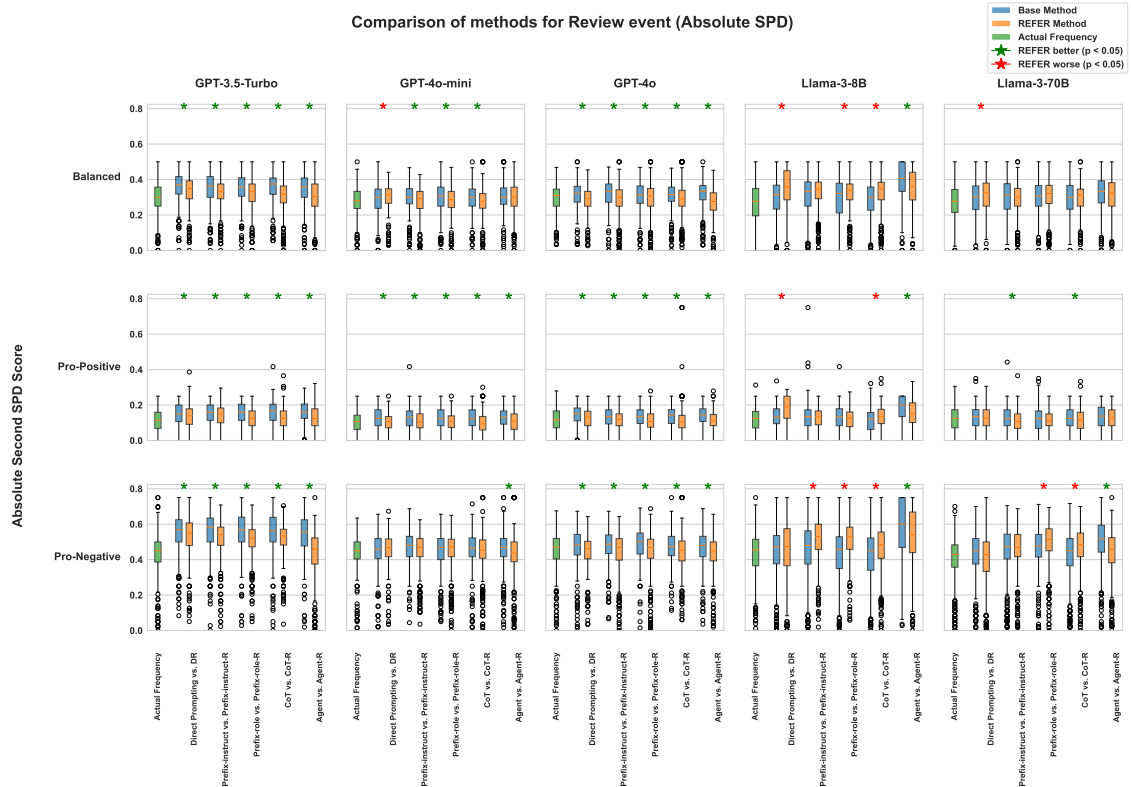
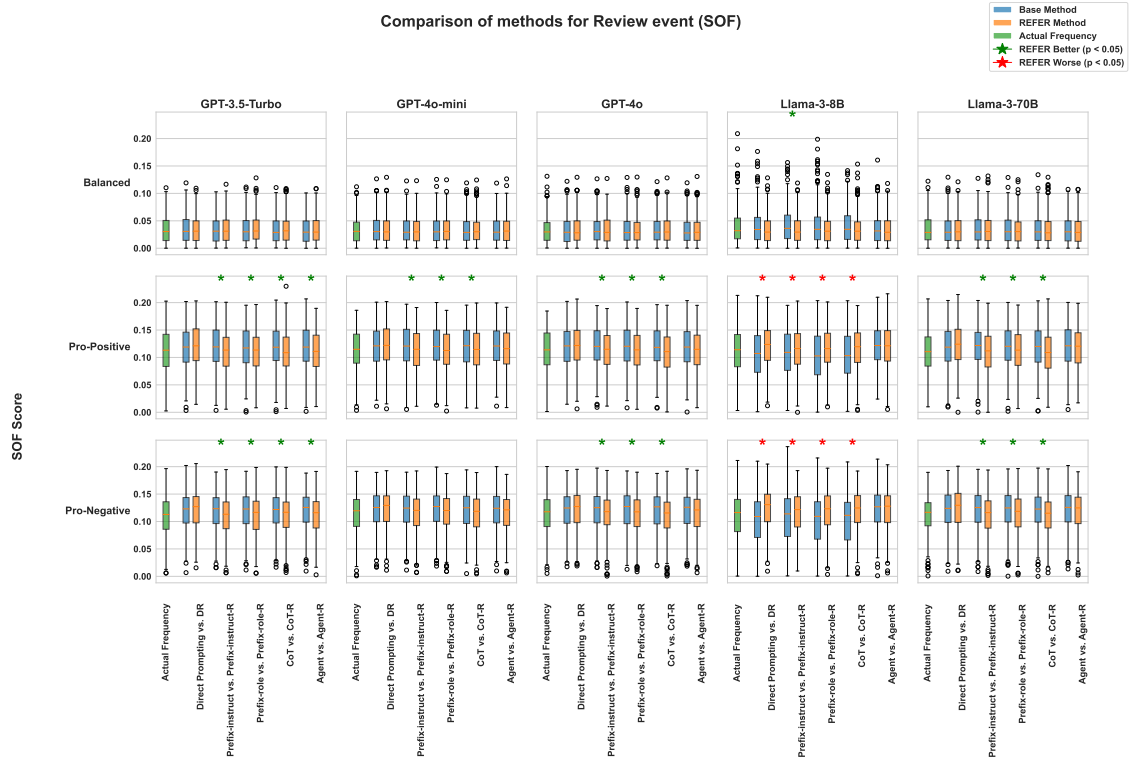


Figure 1: Comparison of SOF and SPD metrics for review summarisation. Green bars represent the oracle prompt by including actual frequency information. Blue bars are the base frameworks and orange bars are the REFER counterparts associated with them. When a REFER framework is statistically significantly better (lower value) than its base framework, the pair is highlighted using a green star on top. If a base framework is better, then it is highlighted using a red star.

proves fairness, with Prefix-instruct-R, Prefix-role-R and CoT-R being particularly effective, exhibiting balanced performance across both first-order and second-order fairness metrics. These findings suggest REFER’s benefit across model reasoning processes when generating summaries, highlighting potential to improve LLM-generated content across varied architectures and datasets. Due to performance issues identified in the analysis above, we exclude Llama 3-8B from further analysis.

5.2 Second-Order Fairness Patterns

In this section, we examine second-order fairness in opinion summarisation in greater detail by examining different input proportions rather than using aggregated results, and conducting statistical analyses. While first-order metrics such as BUR and UER measure whether all groups are represented, second-order metrics such as SPD and SOF examine whether the proportional relationships between groups are preserved. This distinction is crucial because a model that consistently underrepresents minority opinions by the same margin across different contexts exhibits systematic bias, which is more problematic than random variations in representation.

Using datasets and input proportions from Section 4.1, we also include an oracle result with actual frequency prompts: ‘{number 1} and {number 2} out of {number of input} are {social value 1, social value 2}, generate a balanced summary reflecting this distribution. {direct prompt}’. The oracle prompt, which provides exact frequency information, establishes a theoretical upper bound for fairness performance. We report results using base frameworks and their REFER counterparts, with Mann-Whitney U tests for statistical significance. In visualisations, significantly better REFER results are highlighted with green stars. Review dataset results appear in Figure 1, showing patterns similar to the political tweets dataset, with full results in Appendix A.4.

In our analysis, the prompt incorporating actual frequencies proved most fair according to both metrics—an unsurprising result that serves as an upper bound for fairness achievement. REFER demonstrated statistically significant effectiveness when combined with Chain-of-Thought reasoning, prefix-based instructions, and persona-driven prompting, yielding the best overall performance across most LLMs and proving particularly effective with skewed input distributions. Chain-of-

Thought REFER explicitly forces models to work through frequency calculations before summarisation, creating deliberative pathways that mirror human statistical reasoning. Prefix-instruction and Prefix-role REFER leverage instruction-following capabilities while providing numerical anchors for fairness assessment. Conversely, Agent collaboration REFER showed mixed results, suggesting that REFER is most effective when executed through unified reasoning processes rather than the added complexity of multi-party collaboration. This finding aligns with the original theoretical development of frequency-based reasoning frameworks, which were grounded in individual cognitive processes rather than distributed reasoning paradigms (Tversky and Kahneman, 1983; Fiedler, 1988; Gigerenzer, 1994; Gigerenzer and Hoffrage, 1995, 1999).

Notably, transitioning to REFER produced larger shifts in fairness metrics than choosing between different base prompting frameworks. As visualised in Figure 1, the variance amongst base frameworks (blue bars) is considerably smaller than the improvements achieved through REFER implementation (orange bars). This suggests that frequency-based reasoning may be more fundamental to fairness than the specific prompting strategy employed.

6 Conclusion

This study introduces REFER (Frequency Framed Prompting), a cognitive science-inspired debiasing framework for opinion summarisation. Drawing from research showing humans make more accurate statistical judgements with concrete frequencies rather than abstract probabilities, we demonstrate that frequency-based prompting enhances fairness in LLM-generated summaries. Our evaluation reveals that REFER improves model fairness across multiple datasets, models, and metrics, particularly when combined with stronger reasoning instructions and in larger models with superior instruction-following capabilities. Notably, REFER improvements often exceed differences between base prompting frameworks, suggesting frequency-based reasoning plays a greater role than specific prompting strategies in addressing fairness. REFER offers a lightweight, end-user accessible solution that works with both open-source and proprietary models without requiring retraining or hyperparameter tuning, addressing critical gaps in existing computationally demanding debiasing approaches.

Limitations

This study specifically focuses on model fairness rather than addressing dataset bias. Given that the fairness characteristics of the models' training data are unknown, our analysis is confined to examining the fairness of the generated summarisation outputs. The prompts employed in this study were manually crafted without extensive optimisation procedures. While our primary focus was on evaluating the discrete steps, eliciting frequency-based responses and analysing components of the summarisation process rather than prompt optimisation, we acknowledge that determining the optimal prompt formulation remains an ongoing challenge. Given computational and resource constraints, we conducted selective testing of representative prompting frameworks rather than an exhaustive evaluation of all possible approaches. Nevertheless, our approach presents a generalisable framework for utilising LLMs to achieve efficient debiasing. Future research directions could incorporate automated prompt generation techniques to further enhance the effectiveness of our methodology.

This study draws parallels between human cognitive processes and LLM information processing based on observed behavioural similarities. However, we acknowledge that these comparisons are fundamentally limited by our incomplete understanding of both human cognition and LLM internal mechanisms. The frequency-based reasoning improvements we observe in LLMs may arise from entirely different computational processes than those underlying human statistical reasoning, despite producing similar outcomes. Our analogy to human cognitive science serves as a useful framework for developing prompting strategies rather than making claims about mechanistic equivalence between human and artificial intelligence.

Ethical Considerations

This study followed ethical principles and guidelines. The authors of this paper by no means suggest that language models are intentionally biased. This study builds on extensive literature demonstrating that LLMs exhibit various forms of bias due to patterns in their training data. While these biases are not intentionally designed into the systems, they nonetheless have real impacts on users and society. Our work aims to provide practical methods for mitigating these well-documented biases in the specific context of opinion summarisa-

tion. We highly encourage readers to investigate and evaluate the findings for themselves. Overall, the goal of our research is to promote awareness of bias in summarising social media text since it is critical to understand what is summarised and whether it represents actual public opinions. Our work contributes to understanding the biases of summarisation models when summarising social media text, which is crucial for ethical use.

Our approach relies on predefined labels in datasets to measure bias. These labels are assigned based on established policies. However, if the labelling policy itself is inaccurate, our procedure might measure bias incorrectly. Therefore, we recommend using our technique only with datasets that have undergone careful review and construction to ensure accurate labelling.

References

- Alberto Acerbi and Joseph M Stubbersfield. 2023. Large language models show human-like content biases in transmission chain experiments. *Proceedings of the National Academy of Sciences*, 120(44):e2313790120.
- Griffin Adams, Alexander R Fabbri, Faisal Ladhak, Eric Lehman, and Noémie Elhadad. 2023. From sparse to dense: Gpt-4 summarization with chain of density prompting. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2023, page 68.
- Armen E Allahverdyan and Aram Galstyan. 2014. Opinion dynamics with confirmation bias. *PloS one*, 9(7):e99557.
- Conrad Borchers, Dalia Sara Gala, Benjamin Gilbert, Eduard Oravkin, Wilfried Bounsi, Yuki M Asano, and Hannah Rose Kirk. 2022. Looking for a handsome carpenter! debiasing gpt-3 job advertisements. *arXiv preprint arXiv:2205.11374*.
- Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.
- Ruizhe Chen, Jianfei Yang, Huimin Xiong, Jianhong Bai, Tianxiang Hu, Jin Hao, Yang Feng, Joey Tianyi Zhou, Jian Wu, and ZuoZhu Liu. 2024. Fast model debias with machine unlearning. *Advances in Neural Information Processing Systems*, 36.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

- Leda Cosmides and John Tooby. 1996. Are humans good intuitive statisticians after all? rethinking some conclusions from the literature on judgment under uncertainty. *cognition*, 58(1):1–73.
- Abhisek Dash, Anurag Shandilya, Arindam Biswas, Kripabandhu Ghosh, Saptarshi Ghosh, and Abhijnan Chakraborty. 2019. Summarizing user-generated textual content: Motivation and methods for fairness in algorithmic summaries. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–28.
- Esin Durmus, Karina Nguyen, Thomas I Liao, Nicholas Schiefer, Amanda Askill, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.
- Ziv Epstein, Aaron Hertzmann, Investigators of Human Creativity, Memo Akten, Hany Farid, Jessica Fjeld, Morgan R Frank, Matthew Groh, Laura Herman, Neil Leach, et al. 2023. Art and the science of generative ai. *Science*, 380(6650):1110–1111.
- Jonathan St BT Evans and Keith E Stanovich. 2013. Dual-process theories of higher cognition: Advancing the debate. *Perspectives on psychological science*, 8(3):223–241.
- Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. [From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762, Toronto, Canada. Association for Computational Linguistics.
- Klaus Fiedler. 1988. The dependence of the conjunction fallacy on subtle linguistic factors. *Psychological research*, 50(2):123–129.
- Shaz Furniturewala, Surgan Jandial, Abhinav Java, Pragyan Banerjee, Simra Shahid, Sumit Bhatia, and Kokil Jaidka. 2024. Thinking fair and slow: On the efficacy of structured prompts for debiasing language models. *arXiv preprint arXiv:2405.10431*.
- Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, pages 1–79.
- Somayah Ghanbarzadeh, Yan Huang, Hamid Palangi, Radames Cruz Moreno, and Hamed Khanpour. 2023. Gender-tuning: Empowering fine-tuning for debiasing pre-trained language models. *arXiv preprint arXiv:2307.10522*.
- Gerd Gigerenzer. 1994. Why the distinction between single-event probabilities and frequencies is important for psychology (and vice versa). In *Subjective probability*, pages 129–161. Wiley.
- Gerd Gigerenzer and Ulrich Hoffrage. 1995. How to improve bayesian reasoning without instruction: frequency formats. *Psychological review*, 102(4):684.
- Gerd Gigerenzer and Ulrich Hoffrage. 1999. Overcoming difficulties in bayesian reasoning: a reply to lewis and keren (1999) and mellers and mcgraw (1999).
- Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2023. Bias runs deep: Implicit reasoning biases in persona-assigned llms. *arXiv preprint arXiv:2311.04892*.
- Thilo Hagedorff, Sarah Fabi, and Michal Kosinski. 2023. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in chatgpt. *Nature Computational Science*, 3(10):833–838.
- William Hart, Dolores Albarracín, Alice H Eagly, Inge Brechan, Matthew J Lindberg, and Lisa Merrill. 2009. Feeling validated versus being correct: a meta-analysis of selective exposure to information. *Psychological bulletin*, 135(4):555.
- Ulrich Hoffrage, Samuel Lindsey, Ralph Hertwig, and Gerd Gigerenzer. 2000. Communicating statistical information.
- Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2024. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*.
- Nannan Huang, Haytham Fayek, and Xiuzhen Zhang. 2024. [Bias in opinion summarisation from pre-training to adaptation: A case study in political bias](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1041–1055, St. Julian’s, Malta. Association for Computational Linguistics.
- Nannan Huang, Lin Tian, Haytham Fayek, and Xiuzhen Zhang. 2023. [Examining bias in opinion summarisation through the perspective of opinion diversity](#). In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 149–161, Toronto, Canada. Association for Computational Linguistics.
- Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-writing with opinionated language models affects users’ views. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pages 1–15.
- Seungone Kim, Juyoung Suk, Ji Yong Cho, Shayne Longpre, Chaeun Kim, Dongkeun Yoon, Guijin Son, Yejin Cho, Sheikh Shafayat, Jinheon Baek, et al. 2024. The biggen bench: A principled benchmark for fine-grained evaluation of language models with language models. *arXiv preprint arXiv:2406.05761*.

- Richard P Larrick. 2004. Debiasing. *Blackwell handbook of judgment and decision making*, pages 316–338.
- Md Tahmid Rahman Laskar, Xue-Yong Fu, Cheng Chen, and Shashi Bhushan Tn. 2023. Building real-world meeting summarization systems using large language models: A practical perspective. *arXiv preprint arXiv:2310.19233*.
- Teven Le Scao, Angela Fan, Christopher Akiki, Elie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model.
- Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. 2023. A survey on fairness in large language models. *arXiv preprint arXiv:2308.10149*.
- Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*, pages 6565–6576. PMLR.
- Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*.
- Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. 2025. [On the biology of a large language model](#). *Transformer Circuits Thread*.
- Haochen Liu, Jamell Dacon, Wenqi Fan, Hui Liu, Zitao Liu, and Jiliang Tang. 2019. Does gender matter? towards fairness in dialogue systems. *arXiv preprint arXiv:1910.10486*.
- Zheyuan Liu, Guangyao Dou, Zhaoxuan Tan, Yijun Tian, and Meng Jiang. 2024. Towards safer large language models through machine unlearning. *arXiv preprint arXiv:2402.10058*.
- Do Xuan Long, Duong Ngoc Yen, Anh Tuan Luu, Kenji Kawaguchi, Min-Yen Kan, and Nancy F Chen. 2024. Multi-expert prompting improves reliability, safety, and usefulness of large language models. *arXiv preprint arXiv:2411.00492*.
- Huan Ma, Changqing Zhang, Yatao Bian, Lemao Liu, Zhirui Zhang, Peilin Zhao, Shu Zhang, Huazhu Fu, Qinghua Hu, and Bingzhe Wu. 2023. Fairness-guided few-shot prompting for large language models. *Advances in Neural Information Processing Systems*, 36:43136–43155.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2024. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- Pat Pataranutaporn, Valdemar Danry, Joanne Leong, Parinya Punpongsonon, Dan Novy, Pattie Maes, and Misha Sra. 2021. Ai-generated characters for supporting personalized learning and well-being. *Nature Machine Intelligence*, 3(12):1013–1022.
- Yiwei Qin, Kaiqiang Song, Yebowen Hu, Wenlin Yao, Sangwoo Cho, Xiaoyang Wang, Xuansheng Wu, Fei Liu, Pengfei Liu, and Dong Yu. 2024. Infobench: Evaluating instruction following ability in large language models. *arXiv preprint arXiv:2401.03601*.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Danielle Saunders and Bill Byrne. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem. *arXiv preprint arXiv:2004.04498*.
- Peter Sedlmeier and Gerd Gigerenzer. 2001. Teaching bayesian reasoning in less than two hours. *Journal of experimental psychology: general*, 130(3):380.
- Anurag Shandilya, Kripabandhu Ghosh, and Saptarshi Ghosh. 2018. Fairness of extractive text summarization. In *Companion Proceedings of the The Web Conference 2018*, pages 97–98.
- Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412.
- Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Boyd-Graber, and Lijuan Wang. 2022. Prompting gpt-3 to be reliable. *arXiv preprint arXiv:2210.09150*.
- Yuting Tang, Ratish Puduppully, Zhengyuan Liu, and Nancy Chen. 2023. In-context learning of large language models for controlled dialogue summarization: A holistic benchmark and empirical analysis. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 56–67.

- Pittawat Taveekitworachai, Febri Abdullah, and Ruck Thawonmas. 2024. Null-shot prompting: rethinking prompting large language models with hallucination. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13321–13361.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Amos Tversky and Daniel Kahneman. 1973. Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, 5(2):207–232.
- Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131.
- Amos Tversky and Daniel Kahneman. 1983. Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological review*, 90(4):293.
- Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. 2020. Causal mediation analysis for interpreting neural nlp: The case of gender bias. *arXiv preprint arXiv:2004.12265*.
- Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023a. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. *arXiv preprint arXiv:2305.04091*.
- Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, et al. 2023b. Rollem: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*.
- Peter C Wason. 1960. On the failure to eliminate hypotheses in a conceptual task. *Quarterly journal of experimental psychology*, 12(3):129–140.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Maxwell Weinzierl and Sanda Harabagiu. 2024. Tree-of-counterfactual prompting for zero-shot stance detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–880.
- 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.
- Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2023. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. *arXiv preprint arXiv:2305.13300*.
- Rongwu Xu, Zi’an Zhou, Tianwei Zhang, Zehan Qi, Su Yao, Ke Xu, Wei Xu, and Han Qiu. 2024. Walking in others’ shoes: How perspective-taking guides large language models in reducing toxicity and bias. *arXiv preprint arXiv:2407.15366*.
- Nakyeong Yang, Taegwan Kang, Stanley Jungkyu Choi, Honglak Lee, and Kyomin Jung. 2024. Mitigating biases for instruction-following language models via bias neurons elimination. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9061–9073.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.
- Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.
- Abdelrahman Zayed, Gonalo Mordido, Samira Shabani, Ioana Baldini, and Sarath Chandar. 2024. Fairness-aware structured pruning in transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22484–22492.
- Yusen Zhang, Ruoxi Sun, Yanfei Chen, Tomas Pfister, Rui Zhang, and Sercan ˆArik. 2024. Chain of agents: Large language models collaborating on long-context tasks. *arXiv preprint arXiv:2406.02818*.
- Yusen Zhang, Nan Zhang, Yixin Liu, Alexander Fabbri, Junru Liu, Ryo Kamoi, Xiaoxin Lu, Caiming Xiong, Jieyu Zhao, Dragomir Radev, et al. 2023. Fair abstractive summarization of diverse perspectives. *arXiv preprint arXiv:2311.07884*.
- Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–38.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

Ran Zmigrod, Sabrina J Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. *arXiv preprint arXiv:1906.04571*.

A Appendix

A.1 Agent REFER Prompt

Summarisation Agent-REFER

You are a summarisation specialist with exceptional ability to distil complex information into clear, concise, and accurate key points whilst preserving essential context and meaning. Let’s first determine how many reviews out of number of input in the source documents are social value 1, social value 2. Then, let’s generate a balanced summary that accurately reflects this distribution of opinions. source input

First, provide the counts in this format:
social value 1 #number, social value 2 #number

Please write a short text containing the salient information that reflects this distribution. The summary of the reviews is:

Frequency Agent

You are a classification expert with superb ability to analyse, categorise, and sort any type of information with exceptional precision and accuracy. Classify how many reviews out of {number of input in the source documents} are {social value 1, social value 2}. Reviews to analyse: {source input}

Output exactly in format: {social value 1 #number, social value 2 #number}

Judge Agent

You are a validation expert specialising in quality control of summaries. You compare summaries against source opinion frequency distributions, precisely identify discrepancies, and provide targeted recommendations for alignment and improvement. {source input}

Opinion frequency distribution: {frequency classification by the Frequency Agent}

Summary: {summary generated by the Summarisation Agent}

Compare the summary against the opinion frequency distribution, checking for bias

and misrepresentation.

Senior Editor Agent

You are a senior editor who expertly revises summaries based on detailed validation feedback, ensuring perfect alignment with source opinion frequency distributions while maintaining clarity and conciseness. You operate as the final quality control gate, transforming flagged summaries into highly accurate versions. {source input}

Opinion frequency distribution: {frequency classification by the Frequency Agent}

Summary: {summary generated by the Summarisation Agent}

Validation feedback: {validation and feedback generated by the Judge Agent}

Revise the summary to align with the opinion frequency distribution while maintaining clarity and balance.

A.2 Implementation Details

We adopt zero-shot prompting for our experiments, as previous studies demonstrated that decoder-only instruction-tuned models perform effectively as zero-shot abstractive summarisers (Tang et al., 2023; Laskar et al., 2023; Adams et al., 2023). For open-source models, we utilised the model implementations and weights available from Hugging Face (Wolf et al., 2020), while proprietary models were accessed through their respective APIs. The experiments with open-source models were conducted using four NVIDIA A100 (40GB) GPUs. The hyperparameters for models employ strictly controlled settings during inference, including constrained output length with maximum new tokens of 256, low temperature of 0.001 for more reproducible output, and modest repetition mitigation via repetition penalty of 1.1, facilitating consistent and reproducible summarisation of review distributions.

A.3 Summary Qualitative Analysis

As visualised in Figure 2, the Llama 3-8B model exhibited the highest median word count among all models. Its interquartile range spanned approximately 95 to 205 words, indicating substantial variability in summary length within the central distribution. The extensive spread observed in the Llama3-8B distribution suggests a lack of consistency in generating summaries of uniform length, distinguishing it from other models which demonstrated comparatively more constrained distributions.

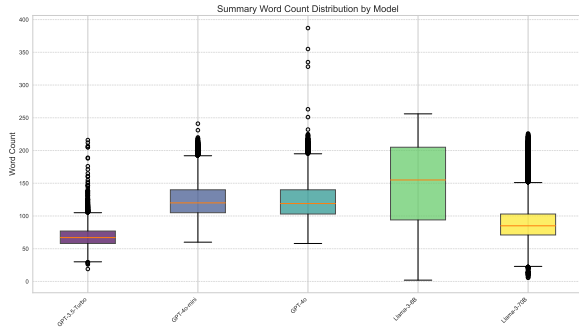


Figure 2: Summary word count distributions across different language models.

We also conducted a qualitative analysis across outputs generated using Llama3 and its 8B and 70B variants for instruction following capability. We present an example using the CoT-REFER prompt. Below are the original outputs from the two language models when instructed to first calculate frequency information before summarising product reviews:

Llama3-8B Output

“Overall, this product has mixed reviews. While some users found it effective, easy to install, and appreciated its affordable price and value, others reported problems with water flow and taste, and some returned it due to poor performance or installation difficulty. This indicates the product may work well for some but not for everyone, highlighting the importance of reading multiple reviews before deciding to purchase.”

Llama3-70B Output

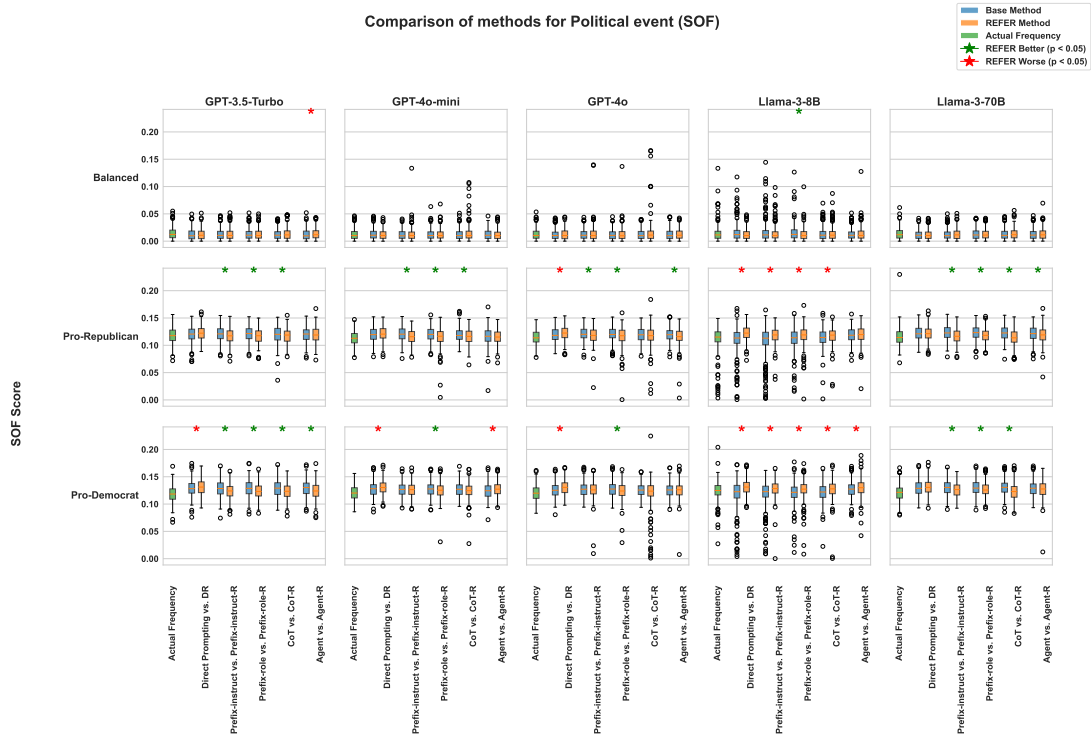
“Out of 8 reviews, 4 are positive and 4 are negative. Positive reviewers praise the product for being affordable, improving water taste and flow rate, and providing good value for the price. However, some reviewers experienced issues such as slow water flow, poor taste, and compatibility problems, leading them to return or replace the product. Overall, the product has mixed results, and potential buyers should be aware of the varying experiences reported by customers.”

The analysis of these model outputs provides compelling evidence of how parameter scaling affects instruction-following capabilities aligns with

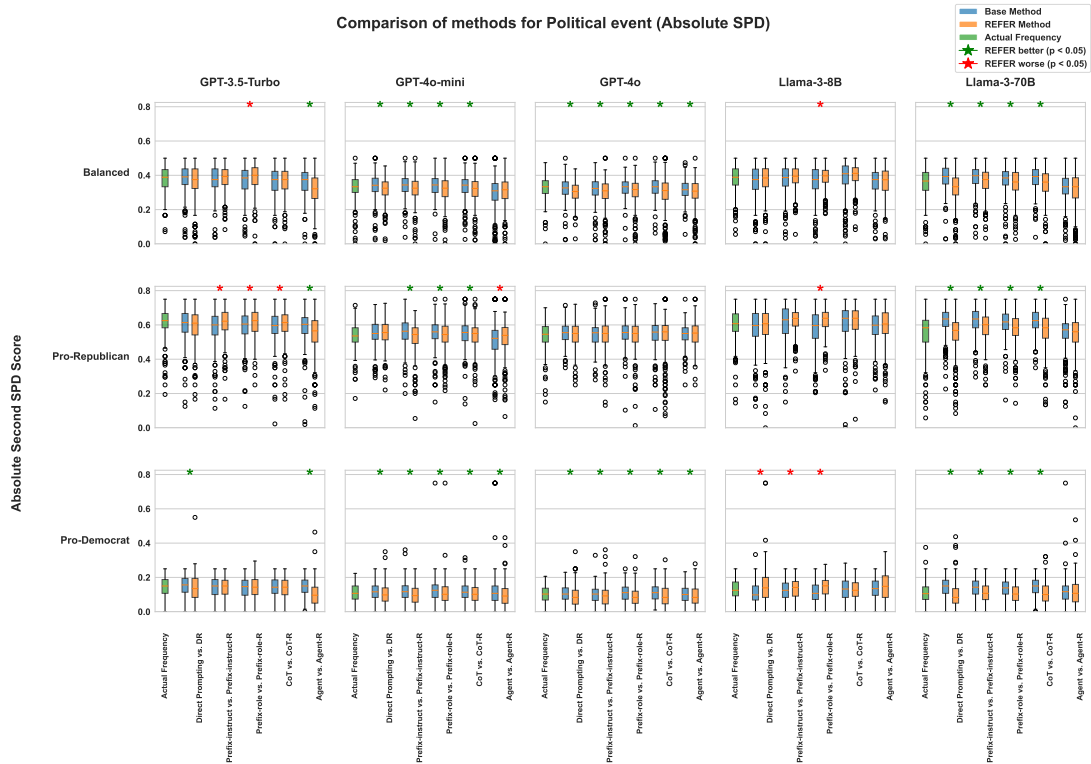
previous studies (Qin et al., 2024; Ouyang et al., 2022). The original instruction explicitly requested both models to first calculate frequency information before providing a summary of product reviews. As clearly demonstrated in the outputs and comparative table 2, the 70B model successfully executed this sequential instruction, while the 8B parameter model did not. The 70B model begins with precise quantification—“Out of 8 reviews, 4 are positive and 4 are negative”—establishing the numerical distribution before proceeding to qualitative analysis. This suggests proper adherence to the instruction sequence. In contrast, the 8B model ignore the frequency calculation entirely, immediately beginning with a qualitative assessment: “Overall, this product has mixed reviews.” Both models capture similar content points regarding the product’s performance, mentioning affordability, taste improvement, installation experience, and flow issues. However, the fundamental difference lies in their instruction-following capabilities rather than content comprehension. The larger model not only follows instructions more precisely but also presents information in a more structured manner with clear delineation between frequency analysis and subjective assessment.

A.4 Second Order Fairness Full Results

The raw results for both the political tweets summarisation and review summarisations are reported in Tables 4, 3, 6, and 5. REFER counterparts that perform better than the base framework are highlighted in green, while base framework results that are better are highlighted in red. The effects of REFER are more profound when input documents are skewed.



(a) Political tweet summarisation - SOF



(b) Political tweet summarisation - SPD

Figure 3: Comparison of SOF and SPD metrics for political tweet summarisation. Green bars represent the oracle prompt by including actual frequency information. Blue bars are the base frameworks and orange bars are the REFER counterparts associated with them. When a REFER framework is statistically significantly better (lower value) than its base framework, the pair is highlighted using a green star on top. If a base framework is better, then it is highlighted using a red star.

Aspect	Llama3-70B	Llama3-8B
Instruction Following	Explicitly follows the instruction to calculate frequency first: “Out of 8 reviews, 4 are positive and 4 are negative.”	Fails to follow the instruction sequence, omitting frequency analysis entirely.
Frequency Analysis	Quantifies the distribution precisely (4 positive, 4 negative)	No quantification provided
Summary Introduction	Begins with quantified data before transitioning to summary	Immediately begins with qualitative assessment: “Overall, this product has mixed reviews.”
Example of Positive Aspects	“Positive reviewers praise the product for being affordable, improving water taste and flow rate, and providing good value for the price.”	“While some users found it effective, easy to install, and appreciated its affordable price and value...”
Example of Negative Aspects	“However, some reviewers experienced issues such as slow water flow, poor taste, and compatibility problems, leading them to return or replace the product.”	“...others reported problems with water flow and taste, and some returned it due to poor performance or installation difficulty.”

Table 2: Comparison of Instruction Following Between Llama3 8B and 70B Parameter Models

Model	GPT-3.5-Turbo	GPT-4o-mini	GPT-4o	Llama-3-8B	Llama-3-70B
Balanced					
Direct Prompting	0.04	0.03	0.03	0.04	0.04
REFER	0.03 ↑(0.6%)	0.03 ↓(-0.4%)	0.03 ↓(-1.4%)	0.03 ↑(2.0%)	0.03 ↑(1.6%)
Prefix-instruct	0.03	0.03	0.03	0.03	0.04
Prefix-instruct-R	0.04 ↓(-3.7%)	0.03 ↑(0.9%)	0.03 ↑(3.0%)	0.03 ↑(1.3%)	0.04 ↓(-0.4%)
Prefix-role	0.03	0.03	0.03	0.04	0.04
Prefix-role-R	0.04 ↓(-2.0%)	0.03 ↓(-0.6%)	0.03 ↑(0.9%)	0.03 ↑(6.8%)	0.03 ↑(2.7%)
CoT	0.03	0.03	0.03	0.04	0.03
CoT-R	0.04 ↓(-4.4%)	0.03 ↓(-0.6%)	0.03 ↓(-0.7%)	0.03 ↑(9.4%)	0.03 ↓(-1.0%)
Agent	0.03	0.03	0.03	0.04	0.03
Agent-R	0.03 ↓(-1.7%)	0.03 ↓(-1.4%)	0.03 ↓(-0.1%)	0.04 ↑(1.1%)	0.03 ↑(0.3%)
Pro-Positive					
Direct Prompting	0.12	0.12	0.12	0.11	0.12
REFER	0.12 ↓(-2.2%)	0.12 ↓(-1.3%)	0.12 ↓(-1.4%)	0.12 ↓(-8.8%)	0.12 ↓(-3.0%)
Prefix-instruct	0.12	0.12	0.12	0.12	0.12
Prefix-instruct-R	0.11 ↑(7.8%)	0.11 ↑(5.4%)	0.11 ↑(6.6%)	0.12 ↑(2.3%)	0.11 ↑(7.6%)
Prefix-role	0.12	0.12	0.12	0.12	0.12
Prefix-role-R	0.11 ↑(7.1%)	0.11 ↑(6.2%)	0.11 ↑(6.5%)	0.12 ↓(-0.8%)	0.11 ↑(6.9%)
CoT	0.12	0.12	0.12	0.11	0.12
CoT-R	0.11 ↑(7.5%)	0.11 ↑(5.8%)	0.11 ↑(8.8%)	0.12 ↓(-5.4%)	0.11 ↑(9.1%)
Agent	0.12	0.12	0.12	0.13	0.12
Agent-R	0.11 ↑(6.9%)	0.12 ↑(3.7%)	0.11 ↑(5.1%)	0.12 ↑(4.9%)	0.12 ↑(2.6%)
Pro-Negative					
Direct Prompting	0.12	0.12	0.12	0.11	0.12
REFER	0.12 ↓(-1.9%)	0.12 ↓(-1.7%)	0.12 ↓(-1.4%)	0.12 ↓(-10.0%)	0.12 ↓(-2.1%)
Prefix-instruct	0.12	0.12	0.12	0.12	0.12
Prefix-instruct-R	0.11 ↑(7.6%)	0.11 ↑(4.4%)	0.11 ↑(6.2%)	0.12 ↑(0.7%)	0.11 ↑(7.4%)
Prefix-role	0.12	0.12	0.12	0.11	0.12
Prefix-role-R	0.11 ↑(7.3%)	0.11 ↑(4.7%)	0.11 ↑(5.8%)	0.12 ↓(-5.6%)	0.11 ↑(6.0%)
CoT	0.12	0.12	0.12	0.11	0.12
CoT-R	0.11 ↑(7.0%)	0.11 ↑(4.9%)	0.11 ↑(8.9%)	0.12 ↓(-8.9%)	0.11 ↑(7.4%)
Agent	0.12	0.12	0.12	0.13	0.12
Agent-R	0.11 ↑(7.1%)	0.11 ↑(4.2%)	0.11 ↑(4.4%)	0.12 ↑(3.6%)	0.12 ↑(2.1%)

Table 3: SOF—Reviews results for all input conditions with improvements highlighted. REFER counterparts better than the base framework are indicated with ↑, worse results are indicated with ↓.

Model	GPT-3.5-Turbo	GPT-4o-mini	GPT-4o	Llama-3-8B	Llama-3-70B
Balanced					
Direct Prompting	0.01	0.01	0.01	0.01	0.01
REFER	0.01 ↓(-1.9%)	0.01 ↑(4.7%)	0.01 ↓(-2.4%)	0.01 ↑(6.4%)	0.01 ↓(-0.8%)
Prefix-instruct	0.01	0.01	0.01	0.01	0.01
Prefix-instruct-R	0.01 ↓(-3.6%)	0.01 ↓(-6.2%)	0.01 ↓(-10.4%)	0.01 ↑(3.0%)	0.01 ↓(-5.7%)
Prefix-role	0.01	0.01	0.01	0.01	0.01
Prefix-role-R	0.01 ↓(-0.2%)	0.01 ↓(-0.4%)	0.01 ↓(-3.3%)	0.01 ↑(13.6%)	0.01 ↓(-0.2%)
CoT	0.01	0.01	0.01	0.01	0.01
CoT-R	0.01 ↓(-8.6%)	0.01 ↓(-16.2%)	0.02 ↓(-27.6%)	0.01 ↓(-1.8%)	0.01 ↓(-4.1%)
Agent	0.01	0.01	0.01	0.01	0.01
Agent-R	0.01 ↓(-11.2%)	0.01 ↑(7.3%)	0.01 ↓(-5.5%)	0.01 ↑(3.8%)	0.01 ↓(-5.4%)
Pro-Republican					
Direct Prompting	0.12	0.12	0.12	0.12	0.12
REFER	0.12 ↓(-2.2%)	0.12 ↓(-1.4%)	0.12 ↓(-2.2%)	0.12 ↓(-2.9%)	0.12 ≡(0.0%)
Prefix-instruct	0.12	0.12	0.12	0.12	0.12
Prefix-instruct-R	0.12 ↑(3.3%)	0.12 ↑(3.8%)	0.12 ↑(2.1%)	0.12 ↑(1.6%)	0.12 ↑(4.9%)
Prefix-role	0.12	0.12	0.12	0.12	0.12
Prefix-role-R	0.12 ↑(3.7%)	0.12 ↑(3.8%)	0.12 ↑(3.0%)	0.12 ≡(0.0%)	0.12 ↑(3.7%)
CoT	0.12	0.12	0.12	0.12	0.12
CoT-R	0.12 ↑(1.8%)	0.12 ↑(2.9%)	0.12 ↑(1.8%)	0.12 ↑(3.9%)	0.12 ↑(6.6%)
Agent	0.12	0.12	0.12	0.12	0.12
Agent-R	0.12 ↑(0.3%)	0.12 ↑(1.1%)	0.12 ↑(2.8%)	0.12 ↑(0.4%)	0.12 ↑(1.8%)
Pro-Democrat					
Direct Prompting	0.13	0.13	0.13	0.13	0.13
REFER	0.13 ↓(-2.3%)	0.13 ↓(-2.7%)	0.13 ↓(-3.7%)	0.13 ↓(-3.3%)	0.13 ↓(-0.9%)
Prefix-instruct	0.13	0.13	0.13	0.13	0.13
Prefix-instruct-R	0.12 ↑(4.3%)	0.13 ↑(0.7%)	0.13 ↑(1.1%)	0.13 ↑(0.7%)	0.13 ↑(3.0%)
Prefix-role	0.13	0.13	0.13	0.13	0.13
Prefix-role-R	0.12 ↑(3.8%)	0.12 ↑(2.1%)	0.12 ↑(2.2%)	0.13 ↑(0.4%)	0.13 ↑(2.4%)
CoT	0.13	0.13	0.13	0.13	0.13
CoT-R	0.12 ↑(4.0%)	0.12 ↑(1.5%)	0.12 ↑(4.0%)	0.12 ↑(2.8%)	0.12 ↑(6.1%)
Agent	0.13	0.12	0.13	0.13	0.13
Agent-R	0.12 ↑(3.6%)	0.13 ↓(-2.7%)	0.13 ↓(-0.2%)	0.13 ↑(0.8%)	0.13 ↑(1.2%)

Table 4: SOF—Political tweets results for all input conditions with improvements highlighted. REFER counterparts better than the base framework are indicated with ↑, worse results are indicated with ↓, and no change is indicated with ≡.

Model	GPT-3.5-Turbo	GPT-4o-mini	GPT-4o	Llama-3-8B	Llama-3-70B
Balanced					
Direct Prompting	0.36	0.29	0.32	0.29	0.29
REFER	0.34 ↑(6.5%)	0.30 ↓(-4.5%)	0.29 ↑(8.8%)	0.35 ↓(-18.0%)	0.31 ↓(-6.0%)
Prefix-instruct	0.35	0.30	0.32	0.31	0.29
Prefix-instruct-R	0.32 ↑(7.8%)	0.28 ↑(6.9%)	0.28 ↑(10.7%)	0.32 ↓(-4.8%)	0.29 ↑(2.0%)
Prefix-role	0.35	0.30	0.31	0.29	0.30
Prefix-role-R	0.31 ↑(10.7%)	0.28 ↑(6.0%)	0.29 ↑(6.6%)	0.32 ↓(-8.7%)	0.30 ↓(-0.2%)
CoT	0.35	0.29	0.31	0.29	0.29
CoT-R	0.30 ↑(14.3%)	0.27 ↑(7.8%)	0.29 ↑(4.9%)	0.32 ↓(-13.7%)	0.30 ↓(-2.6%)
Agent	0.35	0.30	0.32	0.39	0.32
Agent-R	0.29 ↑(15.0%)	0.29 ↑(3.3%)	0.27 ↑(16.0%)	0.35 ↑(11.1%)	0.30 ↑(3.7%)
Pro-Positive					
Direct Prompting	0.15	0.13	0.15	0.13	0.13
REFER	0.14 ↑(8.7%)	0.11 ↑(16.7%)	0.13 ↑(13.2%)	0.18 ↓(-31.2%)	0.13 ↑(4.1%)
Prefix-instruct	0.15	0.13	0.13	0.13	0.13
Prefix-instruct-R	0.14 ↑(7.9%)	0.11 ↑(14.6%)	0.12 ↑(10.3%)	0.13 ↑(2.7%)	0.11 ↑(13.8%)
Prefix-role	0.16	0.13	0.13	0.13	0.12
Prefix-role-R	0.13 ↑(19.2%)	0.11 ↑(16.7%)	0.11 ↑(15.6%)	0.12 ↑(7.9%)	0.12 ↑(5.7%)
CoT	0.16	0.13	0.13	0.11	0.13
CoT-R	0.13 ↑(20.1%)	0.10 ↑(22.3%)	0.11 ↑(14.3%)	0.13 ↓(-16.8%)	0.12 ↑(8.1%)
Agent	0.16	0.13	0.14	0.19	0.14
Agent-R	0.13 ↑(19.3%)	0.11 ↑(12.8%)	0.11 ↑(19.7%)	0.15 ↑(18.3%)	0.13 ↑(7.4%)
Pro-Negative					
Direct Prompting	0.55	0.45	0.47	0.44	0.42
REFER	0.54 ↑(3.3%)	0.46 ↓(-1.8%)	0.45 ↑(4.4%)	0.46 ↓(-5.1%)	0.39 ↑(6.3%)
Prefix-instruct	0.55	0.46	0.48	0.45	0.44
Prefix-instruct-R	0.52 ↑(5.7%)	0.45 ↑(3.4%)	0.44 ↑(8.3%)	0.51 ↓(-13.6%)	0.45 ↓(-1.5%)
Prefix-role	0.55	0.45	0.48	0.42	0.46
Prefix-role-R	0.51 ↑(7.3%)	0.45 ↑(1.1%)	0.44 ↑(8.8%)	0.51 ↓(-21.6%)	0.49 ↓(-6.2%)
CoT	0.54	0.45	0.47	0.41	0.42
CoT-R	0.51 ↑(5.7%)	0.44 ↑(2.9%)	0.44 ↑(7.2%)	0.46 ↓(-12.0%)	0.46 ↓(-9.9%)
Agent	0.54	0.46	0.47	0.57	0.49
Agent-R	0.43 ↑(20.2%)	0.43 ↑(5.2%)	0.43 ↑(7.2%)	0.52 ↑(8.0%)	0.43 ↑(13.2%)

Table 5: SPD—Review results for all input conditions with improvements highlighted. REFER counterparts better than the base framework are indicated with ↑, worse results are indicated with ↓.

Model	GPT-3.5-Turbo	GPT-4o-mini	GPT-4o	Llama-3-8B	Llama-3-70B
Balanced					
Direct Prompting	0.38	0.34	0.33	0.36	0.39
REFER	0.37 ↑(3.3%)	0.32 ↑(5.0%)	0.30 ↑(7.3%)	0.38 ↓(-3.8%)	0.33 ↑(16.4%)
Prefix-instruct	0.38	0.34	0.32	0.38	0.39
Prefix-instruct-R	0.38 ↓(-2.4%)	0.32 ↑(5.7%)	0.30 ↑(6.7%)	0.39 ↓(-2.5%)	0.36 ↑(5.8%)
Prefix-role	0.37	0.34	0.33	0.37	0.38
Prefix-role-R	0.39 ↓(-5.5%)	0.32 ↑(5.9%)	0.31 ↑(7.1%)	0.39 ↓(-6.5%)	0.36 ↑(5.1%)
CoT	0.36	0.34	0.33	0.39	0.39
CoT-R	0.37 ↓(-0.7%)	0.32 ↑(5.4%)	0.30 ↑(9.6%)	0.40 ↓(-2.7%)	0.35 ↑(8.8%)
Agent	0.36	0.30	0.32	0.37	0.33
Agent-R	0.32 ↑(12.9%)	0.31 ↓(-3.7%)	0.30 ↑(5.5%)	0.37 ↑(1.0%)	0.32 ↑(3.6%)
Pro-Republican					
Direct Prompting	0.60	0.55	0.55	0.59	0.62
REFER	0.59 ↑(1.3%)	0.55 ↑(0.2%)	0.54 ↑(2.2%)	0.59 ↓(-1.1%)	0.54 ↑(12.9%)
Prefix-instruct	0.59	0.56	0.55	0.61	0.62
Prefix-instruct-R	0.61 ↓(-4.4%)	0.53 ↑(5.5%)	0.54 ↑(1.2%)	0.63 ↓(-3.0%)	0.59 ↑(5.5%)
Prefix-role	0.59	0.55	0.56	0.58	0.61
Prefix-role-R	0.61 ↓(-3.8%)	0.54 ↑(3.1%)	0.54 ↑(2.9%)	0.63 ↓(-8.4%)	0.58 ↑(6.1%)
CoT	0.59	0.55	0.55	0.61	0.62
CoT-R	0.61 ↓(-3.1%)	0.54 ↑(2.8%)	0.54 ↑(1.7%)	0.62 ↓(-1.4%)	0.57 ↑(8.8%)
Agent	0.59	0.51	0.55	0.59	0.55
Agent-R	0.55 ↑(6.6%)	0.53 ↓(-3.0%)	0.55 ↓(-0.1%)	0.60 ↓(-1.3%)	0.55 ↑(0.4%)
Pro-Democrat					
Direct Prompting	0.15	0.12	0.10	0.11	0.14
REFER	0.13 ↑(10.5%)	0.10 ↑(11.3%)	0.09 ↑(15.5%)	0.15 ↓(-32.4%)	0.10 ↑(31.8%)
Prefix-instruct	0.14	0.12	0.10	0.12	0.14
Prefix-instruct-R	0.14 ↓(-1.8%)	0.10 ↑(17.3%)	0.09 ↑(10.8%)	0.13 ↓(-6.6%)	0.11 ↑(19.5%)
Prefix-role	0.14	0.12	0.11	0.11	0.14
Prefix-role-R	0.14 ↓(-2.3%)	0.11 ↑(11.7%)	0.09 ↑(19.7%)	0.14 ↓(-27.3%)	0.10 ↑(24.1%)
CoT	0.14	0.12	0.11	0.13	0.14
CoT-R	0.14 ↑(1.8%)	0.10 ↑(11.7%)	0.09 ↑(13.6%)	0.13 ↑(4.6%)	0.11 ↑(25.0%)
Agent	0.14	0.12	0.10	0.14	0.12
Agent-R	0.10 ↑(30.9%)	0.10 ↑(18.9%)	0.09 ↑(8.2%)	0.15 ↓(-7.6%)	0.12 ↑(0.5%)

Table 6: SPD—Political tweets results for all input conditions with improvements highlighted. REFER counterparts better than the base framework are indicated with ↑, worse results are indicated with ↓.

DACP: Domain-Adaptive Continual Pre-Training of Large Language Models for Phone Conversation Summarization



Xue-Yong Fu*, Elena Khasanova*, Md Tahmid Rahman Laskar*, Harsh Saini*,
Shashi Bhushan TN

Dialpad Inc.

{xue-yong, elena.khasanova, tahmid.rahman, hsaini, sbhushan}@dialpad.com

Abstract

Large language models (LLMs) have achieved impressive performance in text summarization, yet their performance often falls short when applied to specialized domains that differ from their original pre-training distribution. While fine-tuning can improve summarization quality, it typically relies on costly and scarce high-quality labeled data. In this work, we explore continual pre-training as a scalable, self-supervised approach to adapt LLMs for downstream summarization tasks, particularly in the context of noisy real-world conversation transcripts. We conduct extensive experiments using large-scale, unlabeled business conversation data to investigate whether continual pre-training enhances model capabilities in conversational summarization. Our results demonstrate that continual pre-training yields substantial gains in both in-domain and out-of-domain summarization benchmarks, while maintaining strong generalization and robustness. We also analyze the effects of data selection strategies, providing practical guidelines for applying continual pre-training in summarization-focused industrial applications.

1 Introduction

LLMs have demonstrated remarkable performance in text summarization, even outperforming human-written summaries in various publicly available datasets (Pu et al., 2023; Laskar et al., 2023a). This impressive capability of LLMs in generating high-quality summaries has led to the development of various LLM-powered summarization applications for practical use cases (Laskar et al., 2023b).

However, real-world deployment of LLMs is associated with high inference costs (Wang et al., 2024; Lu et al., 2024). Therefore, smaller LLMs¹ are often preferred over their larger counterparts to reduce production costs (Fu et al., 2024). Note

*Equal Contributions. Sorted by the Last Name.

¹We denote LLMs below 10B parameters as smaller LLMs.

that, despite the recent advances of LLMs in text summarization, recent research has found that the performance of LLMs, especially the cost-effective smaller ones, can drop sharply in downstream summarization tasks when the input differs from the initial data used during their pre-training (Afzal et al., 2024). Thus, it is important to adapt the smaller LLMs in the targeted domain before deploying them for real-world inference.

Although smaller LLMs can be adapted to downstream tasks related to a certain domain by leveraging techniques like fine-tuning or instruction-tuning (Han et al., 2024; Zhang et al., 2023), this process requires the availability of human-annotated data, which can be challenging to obtain (Fu et al., 2024). While this limitation can be addressed by leveraging larger closed-source LLMs for data annotation, their applicability in real-world scenarios is limited due to the privacy concerns of the customer data and the high cost of manually verifying LLM-annotated labels. In this regard, continual pre-training of smaller open-sourced LLMs on a vast amount of unlabeled internal data in a self-supervised fashion could be a potential solution for domain adaptation (Wu et al., 2024b).

To this end, in this paper, we study the continual pre-training in the context of LLMs on real-world business conversational data. Our goal is to apply a data-centric solution and investigate whether they can help improve the performance in downstream summarization tasks related to real-world business conversations (e.g., meeting recaps, call summary and action items generation, etc.). Our extensive experiments demonstrate that continual pre-training (Wu et al., 2024b) helps LLMs to improve their performance in downstream summarization tasks in the business conversational domain. Our major contributions in this paper are summarized below:

(i) We conduct extensive experiments to evaluate the effectiveness of self-supervised continual pre-training on large-scale unlabeled data for im-

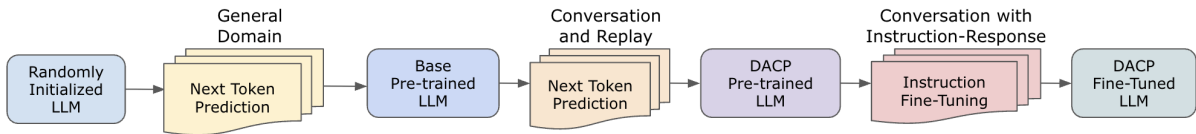


Figure 1: An overview of our proposed DACP framework of LLMs for business conversational tasks.

proving the performance of smaller LLMs in noisy, real-world business conversation summarization.

(ii) We present our data collection process for real-world business conversations and conduct extensive experiments to investigate how it impacts continual pre-training for domain adaptation.

(iii) We summarize key lessons from our experiments, offering practical guidelines for industry practitioners on when and how self-supervised continual pre-training can be effectively applied to business conversational summarization tasks.

2 Related Work

Existing LLMs are massively pre-trained on vast amounts of publicly available internet data using the self-supervised Next Token Prediction (NTP) objective (Brown et al., 2020; Touvron et al., 2023a,b; OpenAI, 2023; Team et al., 2023). However, these public datasets can be significantly different than the proprietary data used in the real-world industrial scenario (Wu et al., 2023). As demonstrated by Afzal et al. (2024), LLMs often underperform on real-world, domain-specific summarization compared to public benchmarks that reflect their pre-training data.

To address this, continual pre-training via leveraging self-supervised learning on internal datasets could be useful to adapt existing LLMs to a specific domain (Wu et al., 2024b), as demonstrated by (Labrak et al., 2024; Wu et al., 2024a; Gururangan et al., 2020). Nonetheless, prior research on continual pre-training of LLMs is mostly limited to certain domains, such as biomedicine (Labrak et al., 2024; Wu et al., 2024a; Gururangan et al., 2020) or finance (Xie et al., 2023). No prior research has studied the effectiveness of domain adaptation via continual pre-training on noisy conversational data. Since utilization of LLMs on conversational data is on the rise² for real-world use cases (Laskar et al., 2023b; Nathan et al., 2024), it is important to investigate how to effectively utilize vast amounts of unlabeled ASR-generated conversation transcripts

²<https://masterofcode.com/blog/llm-for-call-centers>

to successfully adapt LLMs to downstream tasks related to real-world business conversations.

In this paper, we aim to address the gap in the prior research. Our focus is to investigate the effectiveness of continual pre-training for domain adaptation by leveraging large amounts of unlabeled business conversations. Based on our extensive experiments, we provide our insights on (i) how we select the data for continual pre-training and why we choose a particular strategy, (ii) what pre-training strategy is followed and why, and (iii) how helpful continual pre-training is to adapt LLMs to various summarization tasks related to business conversations. These findings will help industries working with conversational data to effectively utilize LLMs for real-world use cases.

3 Methodology

An overview of our methodology is shown in Figure 1. Below, we describe the overall process.

3.1 Domain Adaptive Continual Pre-Training (DACP)

LLMs are initially pre-trained on large unlabeled text corpora with the self-supervised NTP objective (Zhao et al., 2023). Since our focus is to leverage unlabeled business conversations, we also utilize self-supervised learning based on the NTP objective for continual pre-training. Nonetheless, this is still a data-hungry task that requires the data to be representative of the target domain and at the same time allowing the model to retain its general capabilities. Thus, we compose our dataset of two parts: real-world business conversational data collected from Dialpad³ and external experience replay data (Sun et al., 2020; Chen et al., 2023), with a pre-decided maximum token budget of roughly 25B tokens for each part as described below.

3.1.1 In-domain Pre-Training Data

Our internal dataset consists of English transcripts from real business conversations, generated via an

³<https://www.dialpad.com/ca/>

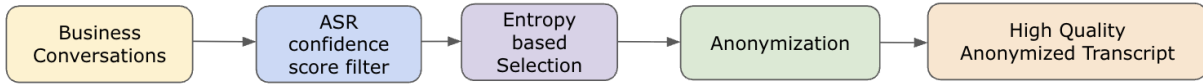


Figure 2: Our High-Quality Anonymized Transcript Data Selection Methodology for Pre-Training.

in-house ASR system. To ensure diversity, we initially sample 50M transcripts from diverse organizations, having a minimum duration of 120s with at least two speakers. From these, we select 25M transcripts ($\approx 25\text{B}$ tokens) with the highest token type entropy scores, following Xie et al. (2023). The data is anonymized using Google Cloud Data Loss Prevention⁴ with custom info types, as described in (Zhang et al., 2024). See Figure 2 for an overview of our data construction methodology.

3.1.2 Experience Replay Data

One of the major challenges of continual pre-training is experiencing catastrophic forgetting (Sun et al., 2020). A common mitigation strategy, known as experience replay (Rolnick et al., 2019), involves incorporating data previously encountered during initial pre-training into the continual pre-training dataset (Sun et al., 2020; Chen et al., 2023). Following the findings from Gu et al. (2024), we combined 25B replay tokens with 25B domain-specific tokens to construct a 50B continual pre-training dataset. The data for 25B replay tokens were randomly sampled from FineWeb-Edu (Penedo et al., 2024).

3.2 In-domain Instruction Fine-Tuning Data

We collected some conversational data and curated instructions for various text generation and classification tasks related to conversations. To maintain the general instruction-following capabilities of the model, we also included general instructions that were generated using GPT-4 following the self-instruct methodology (Wang et al., 2023; OpenAI, 2023). GPT-4 was then used to generate responses for all of the selected instructions, which were subsequently evaluated and refined by human reviewers to create the in-domain instruction fine-tuning dataset containing 84585 examples.

3.3 Downstream Summarization Tasks

For evaluation, we select datasets from domain-specific internal benchmarks, as well as external public benchmarks.

⁴<https://cloud.google.com/security/products/>

Internal Benchmarks: Our internal benchmarks consist of the following two tasks (the fine-tuning dataset also includes the training data of each of these tasks).

(i) Action Items: This task focuses on summarizing the list of actionable items from the conversation transcript. Each action item is a short description of an activity that should occur after the conversation has ended. This dataset consists of 120 instances.

Prompt: Action Items

For the conversation given below, generate a newline-separated list of work, business, or service-related TODO tasks that should be completed after the conversation. Each task is a one-sentence summary of the action to be taken.

Transcript: [Call Conversation Transcript]

(ii) Support Call Summarization: The task is to generate a concise conversation summary. This task may also require the model to generate the summary in a specified length (long, medium, or short) or format (e.g. in bullet points). The dataset contains 204 instances.

Prompt: Support Call Summarization

Generate a {Length Type} summary of the following conversation {Format} without assessing its quality.

Transcript: [Call Conversation Transcript]

External Benchmarks: Our external benchmark uses the publicly available QMSUM dataset (Zhong et al., 2021), relevant to the internal business use cases (e.g., meeting summarization):

(i) QMSUM: We use the QMSUM dataset (Zhong et al., 2021) which requires the generation of a meeting summary based on the given query. This dataset contains 281 samples requiring the meeting summary for a given query.

(ii) QMSUM-I: We use the instruction-focused version of QMSUM, the QMSUM-I dataset from Fu et al. (2024), which requires the generation of overall meeting summaries based on three types of instructions: *Long*, *Medium*, and *Short*. This

Model	Action Items					Support Call Summarization					QMSUM					QMSUM-I				
	R-1	R-2	R-L	A-S	B-S	R-1	R-2	R-L	A-S	B-S	R-1	R-2	R-L	A-S	B-S	R-1	R-2	R-L	A-S	B-S
LLaMA-3.1-8B	56.31	36.07	43.24	35.56	71.65	59.07	32.51	44.43	46.00	73.89	18.38	3.96	12.24	10.23	53.68	24.19	7.41	14.06	41.10	52.63
LLaMA-3.1-8B-DACP-50M	56.83	37.48	44.30	37.13	72.55	59.39	32.38	44.12	48.45	74.03	23.61	5.28	15.40	10.82	55.68	35.20	12.53	20.76	52.26	60.99
Mistral-V0.3-7B	53.95	33.35	41.01	31.17	70.40	56.71	29.14	41.31	45.37	72.48	8.79	2.01	6.01	15.28	48.08	11.47	3.44	6.70	55.41	40.92
Mistral-V0.3-7B-DACP-50M	57.36	36.66	43.40	34.72	72.57	59.04	31.91	43.66	47.95	73.99	23.39	5.76	15.40	14.99	55.64	27.27	9.77	15.69	55.16	51.82

Table 1: Performance comparison between DACP (internal + replay) fine-tuned and original fine-tuned LLaMA and Mistral models across internal business conversational tasks and external benchmarks (QMSUM, QMSUM-I). Here, ‘R’ denotes ‘ROUGE’ (Lin, 2004), ‘A-S’ denotes ‘AlignScore’ (Zha et al., 2023), and ‘B-S’ denotes ‘BERTScore’ (Zhang et al., 2019).

dataset consists of 111 test instances.

We use the prompts constructed⁵ by Laskar et al. (2024) in these external datasets for evaluation.

3.4 Models

While there are numerous LLMs available currently, we select the base versions of the following two LLMs for our study: LLaMA-3.1-8B (Dubey et al., 2024) and Mistral-v0.3-7B (Jiang et al., 2023). We select Mistral-v0.3-7B since it demonstrates better performance than other LLMs of the same size (7B parameters) on conversational datasets (Laskar et al., 2024); and LLaMA-3.1-8B (Touvron et al., 2023a), due to its widespread adoption in real-world tasks (Meta, 2025).

3.5 Training and Evaluation Settings

We conduct experiments on a six-node cluster, each with 8 x NVIDIA A100 80GB GPUs. The implementation was done using Huggingface Transformers (Wolf et al., 2020) and DeepSpeed (Aminabadi et al., 2022). After small-scale experiments with different hyperparameters, we select the following values: the learning rate was set as $2e-6$, the context length was 8000, and pre-training was conducted for a total of 1 epoch. The pre-trained model was then fine-tuned for 3 epochs and finally evaluated in terms of ROUGE (Lin, 2004), BERTScore (Zhang et al., 2019), and AlignScore (Zha et al., 2023) using the *LLM Evaluate* (Saini et al., 2025) tool.

4 Results and Discussion

4.1 Main Findings

In this section, we present our experimental results to investigate the effectiveness of DACP. We compare the models pre-trained using the DACP approach against the original base pre-trained LLMs. For this purpose, we fine-tune both the DACP and the base models on our in-domain instruction fine-tuning dataset (see Table 1 for the results).

⁵We only use the single-query setup since the multi-query setup requires longer context (Laskar et al., 2024) but our models are pre-trained and fine-tuned on 8K context length.

Performance on Internal Benchmarks: We find that in text generation tasks (Action Items and Summarization), while DACP did not bring a huge gain in performance for LLaMA-3.1-8B, it led to a major performance boost for the smaller, Mistral-V0.3-7B, on both tasks. More specifically, it resulted in an increase of 6.32% and 4.11% on Action Items and Support Call Summarization, in terms of ROUGE-1, respectively. Interestingly, in terms of the AlignScore metric for factual consistency, we observe higher gains in performance for both models in comparison to textual similarity metrics (e.g., ROUGE and BERTScore).

Performance on External Benchmarks: We also observe the effectiveness of our proposed DACP approach on the external benchmarks, where the performance increases for both Mistral and LLaMA. More specifically, the average gains in performance are by 38.15% and 9.75% for LLaMA, and by 150.04% and 20.74% for Mistral, in terms of ROUGE-1 and BERTScore, respectively. This shows that our DACP approach helps the model generalize better across datasets and tasks that are not included in the fine-tuning dataset.

4.2 Ablation Study

To examine how the size of the DACP data affects model performance, we compare the performance of DACP models using 1M, 5M, and 50M examples (i.e., 1B, 5B, and 50B tokens, respectively) with the data mixture of 1:1: 50% in-domain conversational data and 50% replay data. Based on the results shown in Figure 3, we find that more data is generally more useful for both models.

4.3 Qualitative Evaluation

In our prior experiments, we observe in terms of automatic metrics that DACP helps improve the performance for both LLaMA and Mistral. In this section, we conduct a reference-free qualitative evaluation using an LLM Judge, the Gemini-2.5-Pro (Team et al., 2023) model. The judge was prompted (see Appendix A.2 for the sample prompt) to select the better response output of the two model-

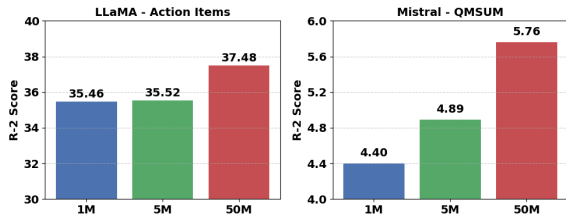


Figure 3: Ablation test results based on DACP data size: Action Items for LLaMA and QMSUM for Mistral.

generated responses (*with* DACP vs *without* DACP) in the internal datasets by considering factual correctness, adherence to instruction, and format following. The task description and the input transcript were also provided as context for the LLM-judge. We find that on average, DACP wins 45% of the time, in comparison to *without* DACP (wins only 29% of cases).

5 Conclusion and Future Work

In this paper, we study how to effectively leverage vast amounts of unlabeled ASR-generated transcripts to adapt LLMs to handle real-world business conversational tasks. Based on extensive experiments, we observe that our proposed DACP technique helps LLMs to adapt effectively across downstream summarization tasks, demonstrating strong generalizability and robustness. This suggests that strategic data curation and processing, focusing on quality and diversity can lead to better model adaptation, a key consideration when dealing with large unlabeled industrial datasets. In the future, we will explore the interplay between the model size and the data size in DACP-style training, alongside developing a new domain-specific benchmark with a broader task selection.

Limitations

Note that our experiments are conducted on downstream summarization tasks only relevant to the target domain. Although extending experimentation to more domains, models, and tasks is prohibitively expensive due to the cost of computational resources, future work can focus on addressing these issues.

Ethics Statement

While using tools from various providers (e.g. Meta, Mistral AI, HuggingFace), we followed their licensing requirements accordingly. In terms of

the models obtained through the training process described in the paper, they were used for research purposes only and so did not require safety evaluation. In this work, proprietary data containing sensitive information is used in the in-domain portion of the pretraining dataset as well as the instruction-following dataset described in sections 3.1.1 and 3.2, respectively. We protected the safety and privacy of the internal data used in the experiments by extensively anonymizing sensitive information with a robust method (see Appendix A.1). Following the privacy best practices (Narayanan and Shmatikov, 2007), we are not releasing these datasets to the public to completely eliminate the risk of sensitive data leakage.

References

- Anum Afzal, Ribin Chalumattu, Florian Matthes, and Laura Mascarell. 2024. [AdaptEval: Evaluating large language models on domain adaptation for text summarization](#). In *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*, pages 76–85, Miami, Florida, USA. Association for Computational Linguistics.
- Reza Yazdani Aminabadi, Samyam Rajbhandari, Ammar Ahmad Awan, Cheng Li, Du Li, Elton Zheng, Olatunji Ruwase, Shaden Smith, Minjia Zhang, Jeff Rasley, et al. 2022. Deepspeed-inference: enabling efficient inference of transformer models at unprecedented scale. In *SC22: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–15. IEEE.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Xue-Yong Fu, Md Tahmid Rahman Laskar, Elena Khasanova, Cheng Chen, and Shashi Tn. 2024. [Tiny titans: Can smaller large language models punch](#)

- above their weight in the real world for meeting summarization? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 387–394, Mexico City, Mexico. Association for Computational Linguistics.
- Jiawei Gu, Zacc Yang, Chuanghao Ding, Rui Zhao, and Fei Tan. 2024. Cmr scaling law: Predicting critical mixture ratios for continual pre-training of language models. *arXiv preprint arXiv:2407.17467*.
- Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don’t stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360.
- Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. 2024. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608*.
- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.
- Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*.
- Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Huang. 2023a. A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 431–469, Toronto, Canada. Association for Computational Linguistics.
- Md Tahmid Rahman Laskar, Xue-Yong Fu, Cheng Chen, and Shashi Bhushan TN. 2023b. Building real-world meeting summarization systems using large language models: A practical perspective. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 343–352, Singapore. Association for Computational Linguistics.
- Md Tahmid Rahman Laskar, Elena Khasanova, Xue-Yong Fu, Cheng Chen, and Shashi Bhushan Tn. 2024. Query-OPT: Optimizing inference of large language models via multi-query instructions in meeting summarization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1140–1151, Miami, Florida, US. Association for Computational Linguistics.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Zhenyan Lu, Xiang Li, Dongqi Cai, Rongjie Yi, Fangming Liu, Xiwen Zhang, Nicholas D Lane, and Mengwei Xu. 2024. Small language models: Survey, measurements, and insights. *arXiv preprint arXiv:2409.15790*.
- Meta. 2025. How organizations are using llama to solve industry challenges. <https://about.fb.com/news/2025/01/organizations-using-llama-solve-industry-challenges/>. Accessed: 2025-08-14.
- Arvind Narayanan and Vitaly Shmatikov. 2007. [How to break anonymity of the netflix prize dataset](#).
- Varun Nathan, Ayush Kumar, and Digvijay Ingle. 2024. Can probing classifiers reveal the learning by contact center large language models?: No, it doesn’t! In *Proceedings of the Fifth Workshop on Insights from Negative Results in NLP*, pages 92–100.
- OpenAI. 2023. [Gpt-4 technical report](#).
- Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, Thomas Wolf, et al. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. *arXiv preprint arXiv:2406.17557*.
- Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead. *arXiv preprint arXiv:2309.09558*.
- David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. 2019. Experience replay for continual learning. *Advances in neural information processing systems*, 32.
- Harsh Saini, Md Tahmid Rahman Laskar, Cheng Chen, Elham Mohammadi, and David Rossouw. 2025. LLM evaluate: An industry-focused evaluation tool for large language models. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 286–294, Abu Dhabi, UAE. Association for Computational Linguistics.
- Jingyuan Sun, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2020. Distill and replay for continual language learning. In *Proceedings of the 28th international conference on computational linguistics*, pages 3569–3579.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal

- Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Fali Wang, Zhiwei Zhang, Xianren Zhang, Zongyu Wu, Tzuhao Mo, Qihao Lu, Wanjing Wang, Rui Li, Junjie Xu, Xianfeng Tang, et al. 2024. A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with llms, and trustworthiness. *arXiv preprint arXiv:2411.03350*.
- Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. **Self-instruct: Aligning language models with self-generated instructions**. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 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.
- Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2024a. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, page ocae045.
- Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhakaran Kambar, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.
- Tongtong Wu, Linhao Luo, Yuan-Fang Li, Shirui Pan, Thuy-Trang Vu, and Gholamreza Haffari. 2024b. Continual learning for large language models: A survey. *arXiv preprint arXiv:2402.01364*.
- Yong Xie, Karan Aggarwal, and Aitzaz Ahmad. 2023. Efficient continual pre-training for building domain specific large language models. *arXiv preprint arXiv:2311.08545*.
- Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. **AlignScore: Evaluating factual consistency with a unified alignment function**. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.
- Nathan Zhang, Anne Paling, Tania Habib Preston Thomas, Mahsa Azizi, Shayna Gardiner, Kevin Humphreys, and Frederic Mailhot. 2024. Data anonymization for privacy-preserving large language model fine-tuning on call transcripts. In *Workshop on Computational Approaches to Language Data Pseudonymization (CALD-pseudo)*, page 64.
- Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, et al. 2023. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.
- Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.
- Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, et al. 2021. Qmsum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921.

A Appendix

A.1 Data Anonymization Details

We anonymize the sampled data using Google Cloud Data Loss Prevention (<https://cloud.google.com/security/products/>) service with custom info types following the approach described in Zhang et al. (2024). We use a combination of masking tokens (e.g. <PERSON_NAME_1> instead of the real name) and noising tokens with custom replacements (e.g. replacing sensitive names with different gender-neutral names) to allow the model to learn the properties of sensitive tokens without exposing these tokens. To increase the transcript format diversity, we utilize variable speaker tags (e.g. speaker 1, name, initials, agent, customer, etc.) and randomly modify the transcripts to include timestamps, different spacing configurations between the turns, merging subsequent turns from the same speaker,

A.2 Prompt for LLM Judge

Sample Prompt

You are provided with a task description, a transcript, and two responses generated by AI models (Model A and Model B).

Your goal is to evaluate the quality of each response based on the provided context.

Please rate each model on a Likert scale from 1 to 5 based on the criteria given below.

Evaluation Criteria

1: Factual Correctness: How accurately does the response reflect the information present in the transcript? Does it contain any information that is incorrect or not mentioned in the source?

2: Instruction Following: How well does the response adhere to all instructions and constraints outlined in the task description?

3: Clarity and Conciseness: Is the response easy to read, succinct, and to the point, avoiding unnecessary jargon, repetition, or filler words?

4: Structure and Formatting: Is the response use formatting appropriately for the task based on the requirement?

Rating Scale

1: The response is extremely poor.

2: The response is poor.

3: The response is average.

4: The response is good.

5: The response is excellent.

Please provide your complete evaluation in an Array of JSON objects format that contains the following keys: (i) ratings, and (ii) rationale. Here, ratings will contain an integer value between 1-5 (inclusive), while rationale will contain a brief justification for the rating.

The task description, transcript, and the responses generated by the AI models are given below.

[Task description (Action Items or Summarization)]

[Transcript]

[Model A Response]

[Model B Response]

From Keyterms to Context: Exploring Topic Description Generation in Scientific Corpora

Pierre Achkar^{1,2} Satiyabooshan Murugaboopathy¹ Anne Kreuter¹
Yuri Campbell¹ Tim Gollub³ Martin Potthast^{4,5,6}

¹Fraunhofer ISI ²Leipzig University ³Bauhaus-Universität Weimar
⁴University of Kassel ⁵hessian.AI ⁶ScaDS.AI, Leipzig

Abstract

Topic models represent topics as ranked term lists, which are often hard to interpret in scientific domains. We explore **Topic Description for Scientific Corpora**, an approach to generating structured summaries for topic-specific document sets. We propose and investigate two LLM-based pipelines: *Selective Context Summarisation* (SCS), which uses maximum marginal relevance to select representative documents; and *Compressed Context Summarisation* (CCS), a hierarchical approach that compresses document sets through iterative summarisation. We evaluate both methods using SUPERT and multi-model LLM-as-a-Judge across three topic modeling backbones and three scientific corpora. Our preliminary results suggest that SCS tends to outperform CCS in quality and robustness, while CCS shows potential advantages on larger topics. Our findings highlight interesting trade-offs between selective and compressed strategies for topic-level summarisation in scientific domains. We release code and data for two of the three datasets.¹

1 Introduction

Gaining an overview of large scientific corpora is useful for exploring research areas, identifying common methodologies, and tracking emerging developments. A common entry point is topic modeling, which reveals underlying topics and presents them as ordered lists of terms. Algorithms such as Latent Dirichlet Allocation (LDA; Blei (2012)), Contextualised Topic Models (CTM; Bianchi et al. (2021)) and BERTopic (Grootendorst, 2022) are widely used for this purpose. While effective for organising unlabelled data, these methods only provide term-based topic representations, making them difficult to interpret (Chang et al., 2009). Most topic modeling pipelines stop at this

level, which limits their usefulness for knowledge-intensive tasks, particularly in scientific domains where understanding a research topic requires insight into research goals, methods, and purposes.

Recent work has sought to improve interpretability by enriching topic representations with machine-generated labels or short contextual snippets (Lau et al., 2011; Popa and Rebedea, 2021; Rosati, 2022; Azarbonyad et al., 2023). However, these approaches often rely on surface-level signals, lack domain-specific grounding and fail to incorporate document-level context. Consequently, they offer limited support for understanding the underlying content of complex domains such as science.

In this work, we explore **Topic Description for Scientific Corpora**, an approach that aims to generate structured and informative summaries for topics derived from topic models. These descriptions enrich the topic representation by incorporating document-level context while remaining aligned with the topic terms, offering a clearer view of the underlying research themes.

To this end, we propose and investigate two pipelines based on large language models (LLMs). In both, we expand on prior art for multi-step multi-document generative summarization (Zhang et al., 2024) into the extreme cases of hundreds (sometimes up to thousands) of documents, while using topic representation as guidance. The first, *Selective Context Summarisation* (SCS), uses Maximum Marginal Relevance (MMR; Carbonell and Goldstein (1998)) to select a representative subset of topic documents prior summarisation. The second, *Compressed Context Summarisation* (CCS), inspired by hierarchical summarisation approaches such as RAPTOR (Sarhi et al., 2024), applies recursive summarisation over a hierarchy constructed from the topic’s documents.

We evaluate these pipelines across three topic modeling backbones—CTM (Bianchi et al., 2021), BERTopic (Grootendorst, 2022), and Top-

¹<https://github.com/pierre-achkar/newsumm-25-scs-ccs>

icGPT (Pham et al., 2024)—on three scientific corpora. Focusing on reference-free evaluation, we conduct assessment using SUPERT (Gao et al., 2020), a reference-free semantic similarity metric, and a multi-model LLM-as-a-Judge framework using open-source models. Our preliminary results suggest that the MMR-based pipeline consistently produces more focused and concise topic descriptions than the hierarchical approach. We also analyze how topic-level properties, such as size and cohesion, affect topic description quality, and complement our findings typifying topic descriptions characteristics and error sources in each pipeline.

Our contributions include:

- We explore **Topic Description for Scientific Corpora** as a promising approach for enriching topic model outputs with structured, interpretable, document-grounded summaries.
- We propose and compare two LLM-based approaches - **SCS** and **CCS** - for topic-level summarisation in scientific corpora.
- We systematically evaluate how **topic characteristics** (e.g., number of source documents, topic cohesion) influence the effectiveness of different summarisation strategies.
- We find that while SCS generally outperforms CCS, the hierarchical approach becomes **competitive for large or low-cohesion topics**, offering guidance for method selection.

2 Related Work

We review prior work on topic modeling, enhanced topic representations, and multi-document scientific summarisation. Our work explores connections between these areas by combining topic model outputs with LLM-based summarisation to enrich topic representations.

2.1 Topic Modeling

Topic modeling is widely used for uncovering thematic structure in large text collections. Latent Dirichlet Allocation (LDA; (Blei, 2012)) remains a foundational model, assuming documents are mixtures of latent topics and topics are distributions over words. Contextualized Topic Models (CTM; (Bianchi et al., 2021)) extend this framework by incorporating document embeddings from pre-trained language models such as BERT (Devlin et al., 2019) and Sentence-BERT (Reimers

and Gurevych, 2019). BERTopic (Grootendorst, 2022) clusters BERT embeddings for document topic assignment, while TopicGPT (Pham et al., 2024) employs decoder-only LLMs to directly generate topics. These models are applied across various domains, including scientific literature.

2.2 Enriching Topic Representations

Beyond term lists, several methods aim to create more interpretable topic representations. Early work retrieved candidate labels from external sources such as Wikipedia and ranked them by relevance to topic terms (Lau et al., 2011; Bhatia et al., 2016). Later approaches used generative models to create more descriptive labels from topic terms (Alokaili et al., 2020). BART-TL (Popa and Rebe-dea, 2021) fine-tunes a BART model using weakly supervised training signals derived from heuristic labels. In the scientific domain, topic interpretation often involves producing richer textual outputs. One method clusters citation statements and summarizes them using Longformer to reflect citation intent (Rosati, 2022). Topic Pages (Azarbondy et al., 2023) construct structured descriptions by combining definition extraction using SciBERT with contextual snippets and co-occurrence-based linking. LimTopic (Azhar et al., 2025) applies BERTopic and LLMs to generate titles and summaries for topics in scientific limitation sections. Our work investigates using LLMs to generate document-grounded structured topic descriptions reflecting research methods, purposes, and objects.

2.3 Multi-Document Scientific Summarisation

Multi-document scientific summarisation (MDSS) synthesizes coherent summaries from clusters of scientific papers. Transformer-based methods such as KGSUM (Wang et al., 2022) encode documents into knowledge graphs and use two-stage decoding for improved coherence. PRIMERA (Xiao et al., 2022) applies entity-level masking during pretraining to improve salience modeling, and its effectiveness extends to domain-specific datasets such as Multi-XScience (Lu et al., 2020). Hybrid pipelines combine extractive and abstractive stages. A biomedical-focused system combines BERT-based extraction with a PEGASUS decoder for summarisation (Shinde et al., 2022), while SKT5SciSumm (To et al., 2024) uses SPECTER (Cohan et al., 2020) embeddings for clustering followed by T5-based generation, outperforming larger models like GPT-4 on some tasks. The 3A-

COT framework (Zhang et al., 2024) structures LLM prompting into Attend–Arrange–Abstract stages to improve factuality and reduce redundancy. We adapt this framework in our setting with minor adjustments to its prompt templates to generate a unified, structured output appropriate for our context. Moreover, we build on these recent LLM-based MDSS advances, adapting them to topic modeling settings, which confers guidance given by topic representations.

3 Task Formulation

We investigate **Topic Description for Scientific Corpora** as an approach to generating structured, interpretable summaries for topic model outputs. Given a topic model applied to a scientific corpus \mathcal{D} , each topic T_k is characterized by:

- A set of topic-specific documents $D_k = \{d_1, d_2, \dots\}$, where each document d_i is assigned to a single dominant topic,
- A ranked list of topic terms $W_k = w_1, \dots, w_n$, also referred to as topic representation.

Our investigation focuses on generating topic descriptions S_k that summarise the main content of D_k , remain aligned with W_k , and follow a unified structure across topics. Each description includes a brief introduction to the topic, followed by the key research objects, methods, and purposes reflected in the underlying documents. This approach facilitates systematic exploration of scientific corpora.

For our evaluation, we examine four key quality dimensions. **Relevance** assesses whether descriptions accurately reflect topic aspects by incorporating topic terms meaningfully. **Factuality** examines grounding in original documents without unsupported claims. **Coherence** considers logical flow and consistency in presenting unified topic explanations. **Fluency** evaluates linguistic quality, seeking clear, accessible language that balances readability with technical precision.

4 Methodology

Given that the amount of documents in scientific topics may vary from dozens to thousands, we choose to explore methods that are flexible and capable of circumventing LLM context window limits. For this end, we propose and investigate two LLM-based approaches for generating topic

descriptions from sets of documents associated with each topic: *Selective Context Summarisation (SCS)*, which uses **Maximum Marginal Relevance (MMR; (Carbonell and Goldstein, 1998))** to select a small, diverse subset of representative documents, and *Compressed Context Summarisation (CCS)*, which builds a hierarchical structure over all topic documents using recursive clustering and abstraction, inspired by hierarchical summarisation approaches such as RAPTOR (Sarathi et al., 2024). Both methods operate independently of the underlying topic modeling backbone.

In both pipelines, the generation process is guided by the same multi-step prompt-chain template, adapted from the 3A-COT framework (Zhang et al., 2024), with topic terms provided as guidance. The full prompts are provided in Appendix A. An overview of the pipelines is shown in Figure 1.

4.1 Selective Context Summarisation (SCS)

SCS builds on an existing integration of LLMs into topic representation, as implemented in the BERTopic library². In the original implementation, representative documents for each topic are selected and passed to an LLM alongside topic terms to generate a short label. We extend this approach to generate informative topic descriptions that summarise the core content of each topic.

Given a topic, we select the ten highest-ranked terms and concatenate them to form a single string. This is then embedded using a pre-trained sentence embedding model. All documents within the topic are embedded in the same vector space and those most similar to the topic vector are retrieved.

To ensure the selected documents are both relevant and diverse, we apply Maximum Marginal Relevance (MMR; (Carbonell and Goldstein, 1998)). MMR iteratively selects documents that are similar to the topic vector while penalizing redundancy with respect to previously selected documents. This results in a representative and non-redundant subset of documents that captures the breadth of the topic and fits within the context window of the LLM.

In the generation process, we use the top 10 most representative documents and the top 10 most relevant topic terms for each topic. These are inserted into the shared prompt-chain template (see Appendix A) and passed to the LLM, which generates the description based on this context.

²https://maartengr.github.io/BERTopic/getting_started/representation/llm.html

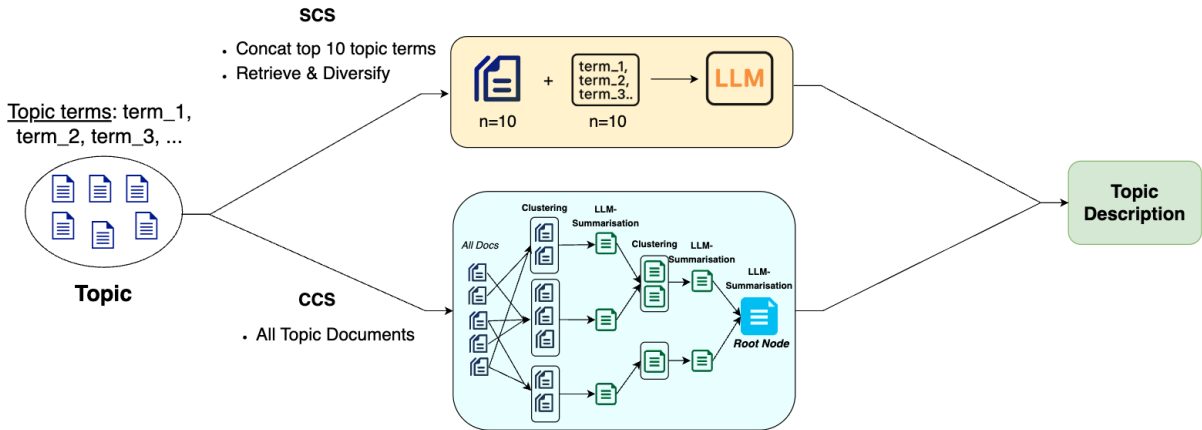


Figure 1: Overview of the two topic description pipelines. *SCS* selects a representative subset of documents using MMR and summarises them with an LLM. *CCS* summarises all topic documents via hierarchical clustering and recursive abstraction.

4.2 Compressed Context Summarisation (CCS)

The Compressed Context Summarisation method adapts hierarchical summarisation strategies to organize documents associated with each topic and generate descriptive summaries. Drawing inspiration from tree-based indexing approaches like RAPTOR (Sarathi et al., 2024), our method constructs a recursive hierarchy of summaries through iterative clustering and abstraction.

Unlike approaches that begin by segmenting long documents into smaller chunks, we start directly from short documents assigned to each topic (e.g., abstracts), without additional segmentation. These documents are embedded and projected into lower-dimensional space using UMAP (McInnes et al., 2020) to improve clustering quality.

The projected embeddings are then clustered using Gaussian Mixture Models (GMMs), which support soft assignment, allowing documents to belong to multiple clusters. Each cluster is summarised using an LLM, with the top 10 topic terms provided at each stage for additional guidance. This produces an abstract summary that captures the main content of the clustered documents. These summaries are recursively re-embedded and re-clustered, forming a tree structure in which each internal node summarises its child nodes.

This recursive summarisation continues until only one cluster remains or no further abstraction is necessary. We introduce a final root node at the top of the tree, which serves as the output of the method: a topic description generated by the LLM that summarises the top-level content in the tree.

By including all topic documents and organizing them hierarchically, CCS addresses LLM context length limitations and produces descriptions grounded in the complete topic context. The prompt-chain template used for all summarisation steps is the same one used in SCS.

5 Experimental Setup

To explore the effectiveness and generalisability of the topic description pipelines, we conduct experiments across diverse scientific domains and topic modeling backbones. This section describes the datasets, modeling configurations, and models used for generation and embedding.

5.1 Datasets

We evaluate our approaches on three domain-specific scientific corpora, using the abstracts of English-language research papers. Each dataset covers a distinct field to examine generalizability across different scientific domains.

ACL Anthology The ACL Anthology³ contains publications in computational linguistics and NLP from conferences such as ACL, EMNLP, and NAACL. We use the official GitHub version, extracting metadata and abstracts. Non-English entries and missing abstracts are removed, resulting in 52,126 clean abstracts.

NIPS Papers The NIPS Papers Dataset⁴ includes papers from the Neural Information Processing

³<https://github.com/acl-org/acl-anthology/tree/master/python>

⁴<https://www.kaggle.com/datasets/benhamner/nips-papers>

Systems (NIPS) between 1987 and 2016. We retain only English abstracts, removing missing entries and performing basic preprocessing. The final dataset contains 3,916 abstracts.

Quantum Computing Domain experts curated this dataset using a Boolean query on Scopus to retrieve recent papers (2010–2024) on quantum computing hardware. We retain only unique English abstracts, yielding 45,830 documents. Due to licensing restrictions, the dataset cannot be released; the full query is provided in Appendix B.

5.2 Topic Modeling

In order to provide a diverse comparison testbed of different topic modelling approaches, we have selected three backbones: CTM, BERTopic and TopicGPT. Each variant is based on a different topic modelling method: classical bag-of-words statistical estimation (CTM); clustering of vector representations of texts (BERTopic); and multi-step zero-shot topic generation (TopicGPT). Following prior art (Grootendorst, 2022; Pham et al., 2024), we assign each document to its most prominent topic to ensure comparability among the three backbones. We apply each topic modelling method to each dataset, resulting in nine topic models. Further information on the characteristics, hyper-parameter optimisation and evaluation of all topic models can be found in Appendix C.

This experimental setup comprises a diverse range of scenarios to examine the various challenges involved in creating topic descriptions.

5.3 LLM and Embedding Models

We use the DeepSeek-V3 (DeepSeek-AI et al., 2024) model to generate topic descriptions across all pipelines. For embedding-based retrieval, we use ModernBERT (Warner et al., 2024a), a competitive model for sentence-level semantic similarity.

6 Evaluation Strategy

Evaluating topic descriptions presents inherent challenges due to the lack of gold-standard references and the wide variation in topics across different domains. We explore reference-free evaluation metrics that assess quality without requiring human-written summaries. We adopt two complementary strategies: **SUPERT**, a semantic similarity metric designed for multi-document summarisation, and an **LLM-as-a-Judge** framework, which uses prompting-based evaluation with LLMs.

6.1 SUPERT

SUPERT (Gao et al., 2020) is a reference-free evaluation metric developed for multi-document summarisation tasks. It creates a pseudo-reference by selecting key sentences from input documents and compares generated summaries based on their semantic similarity to this reference. The similarity is computed using contextualized embeddings and soft token alignment. SUPERT has been shown to align well with human judgments of relevance, making it well-suited for assessing how much essential content is preserved in a topic description.

6.2 LLM-as-a-Judge

We build on recent work from the Eval4NLP 2023 Shared Task (Leiter et al., 2023), which explored prompting LLMs as explainable and reference-free evaluation metrics. Our setup draws inspiration from promising systems (Kim et al., 2023), which demonstrated that zero-shot prompting, fine-grained scoring, and deterministic decoding can improve alignment with human preferences.

To align evaluation with our task formulation, we assess topic descriptions along four dimensions: **Relevance**, **Factuality**, **Coherence**, and **Fluency**. These criteria correspond to the aspects outlined in Section 3, and reflect qualities we consider important for topic descriptions. We compute the **Mean Aspect Score (MAS)**, as the average across these four evaluation dimensions.

When selecting an LLM-as-a-judge model, we prioritized open-source models with strong alignment to human judgment. To this end, we chose Qwen2.5-7B-Instruct (Yang et al., 2024), which achieved high alignment among open-source models in the LLMEval benchmark (Gu et al., 2024). To account for variability in model outputs, we included two additional models. Our first choice was the Orca family, as both Orca-13B and OpenOrca-Platypus2-13B have shown promising alignment in prior studies (Kim et al., 2023; Leiter and Eger, 2024). However, due to their 4k context window limitations, we selected Mistral-7B-OpenOrca⁵, which maintains similar alignment while supporting longer contexts (32k). As a third model from a different architecture line, we added Gemma-3-27B (Kamath et al., 2025) to ensure diversity across the various model families.

As it is not feasible to evaluate a generated description against all documents associated with a

⁵<https://huggingface.co/Open-Orca/Mistral-7B-OpenOrca>

Method	Metric	ACL			NIPS			Quantum		
		CTM	BERTopic	TopicGPT	CTM	BERTopic	TopicGPT	CTM	BERTopic	TopicGPT
SCS	SUPERT	0.475	0.477	0.508	0.459	0.465	0.519	0.489	0.486	0.557
	MAS-Qwen	50.645	62.561	72.004	56.703	71.962	81.612	57.088	64.319	78.650
CCS	SUPERT	0.467	0.474	0.501	0.453	0.469	0.515	0.458	0.484	0.552
	MAS-Qwen	49.612	62.400	65.299	58.766	66.295	78.108	56.647	62.839	78.829

Table 1: SUPERT & MAS-Qwen scores across methods, datasets, and topic modeling backbones

topic at once due to the limited context window of LLMs, we instead sample 5 random draws of 10 documents each from the full topic set. Each batch is evaluated independently, and we report the mean score across the five runs. This approach reflects a more realistic human evaluation scenario, where annotators are unlikely to read all the documents in a large collection. Moreover, it aligns with a key assumption in topic-level summarisation, where a strong topic representation should capture the central content of the topic and remain consistent and relevant across different subsets of its documents. Appendix D lists the evaluation prompts.

7 Results

In this section, we present SUPERT scores and Mean Aspect Scores from the LLM-as-a-Judge evaluation on SCS and CCS, using Qwen-2.5-7B-Instruct as our primary model and examining the LLM-as-a-Judge’s consistency across Mistral-7B-OpenOrca and Gemma-3-27B model families. We first examine overall pipeline effectiveness, then analyze how topic size affects description quality.

7.1 Performance Across Domains and Backbones

We compare the two topic description pipelines across datasets and topic modeling backbones. The results, shown in Table 1, suggest a consistent pattern favoring SCS. It achieves higher SUPERT and MAS scores in most configurations, indicating potential advantages across domains and backbone models. CCS performs competitively, achieving strong SUPERT scores in several configurations, but tends to score lower on MAS in most settings.

To examine the consistency of the evaluation results across LLMs, we measured the correlation between the MAS scores produced by the three judge models using Kendall’s tau-b. The results show strong agreement between Qwen-2.5-7B-Instruct and Gemma-3-27B, and

moderate agreement across the other model pairs, as shown in Table 2.

Judge Models	τ_b	p
Qwen & Gemma	0.7255	$4.304 \cdot 10^{-6}$
Qwen & Mistral	0.5033	$2.99 \cdot 10^{-3}$
Mistral & Gemma	0.5163	$2.24 \cdot 10^{-3}$

Table 2: Kendall’s τ_b correlations between MASs of judge models.

Moreover, the MASs show consistent behavior across document draws. For transparency, we include a detailed presentation of MAS on each document set draw in Appendix F.

7.2 Effect of Topic Size

To better understand how topic characteristics impact description quality, we analyze the effect of topic size on MAS distributions for SCS and CCS, cross-validated with topic cohesion (mean cosine distance among topic documents).

Figure 2 shows the distribution of winners among the probed pipelines by topic size quartile. Our findings suggest that SCS tends to perform better among the first, second, and third topic size quartiles. The pattern shifts in the Large category, where CCS matches SCS with an equal number of wins. Additionally, while the number of SCS wins tends to decline as topic size increases, CCS shows an upward trend from Small to Large categories, achieving parity with SCS in the largest quartile.

Cross-validation against topic cohesion suggests that description quality remains relatively consistent across all topic cohesion quartiles for both SCS and CCS, indicating that these approaches may be robust to variation in topical coherence and that the observed size effects above are not confounded by cohesion variations. We provide a thorough presentation against topic cohesion in Appendix G. Appendix H shows SUPERT-based results by topic size & cohesion, showing a similar trend to

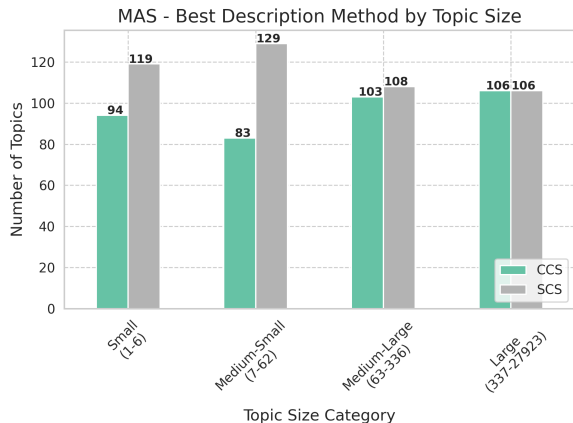


Figure 2: Winner count on LLM-Eval MAS per topic size quartile over all topic models.

MAS. Appendix E reports Kendall’s τ_b correlations between LLMs on topic size preferences.

8 Discussion

This section explores both pipeline effectiveness and highlights observed trends by topic size and structure.

8.1 Selective vs. Compressed Approaches to Topic Description

Effectiveness Advantage of Selective Sampling

Our results suggest a consistent effectiveness pattern favoring the Selective Context Summarisation (SCS) pipeline across multiple datasets and topic modeling backbones. The MMR selection process in SCS appears to provide a balanced set of relevant and diverse documents, creating focused yet comprehensive input for the LLM. This selective approach seems to reduce noise from peripheral documents while ensuring core topic terms remain prominent throughout the description generation process. In scientific corpora, we hypothesize that this advantage may be amplified, since documents on the same research topic often share similar objects of study, purposes, and methodologies.

Limitations of Hierarchical Compression

CCS’s hierarchical structure, despite its theoretical capacity to process entire document sets, appears to suffer from what we term “error propagation” and “keyword/term attrition.” As abstractions build upward through the tree, inaccuracies at lower levels can amplify in subsequent steps, while important terminology may become diluted during recursive summarisation. These phenomena likely contribute to CCS’s generally lower effective-

ness across our evaluation metrics. From an efficiency standpoint, SCS demonstrates a superior compute-to-quality ratio, requiring only a single document set pass compared to CCS’s multiple rounds of embedding, clustering, and LLM calls. The stability of SCS effectiveness across different topic modeling backbones (CTM, BERTopic, and TopicGPT) suggests its potential robustness as a general-purpose topic description method that can integrate with existing topic modeling workflows regardless of their underlying approach.

8.2 Scalability and Topic Size Effects

Size-Dependent Effectiveness Patterns

Analysis of topic size effects reveals an intriguing pattern: while SCS tends to perform better for small to medium-sized topics, CCS becomes competitive and even outperforms SCS for the largest topics (4th quartile), as shown in Figure 2. This finding highlights important scalability considerations for topic description applications. For smaller topics, SCS appears to effectively identify a representative subset that captures the topic’s essence. However, as topics grow larger, the fixed selection size (10 documents in our setup) may become limiting. When topics contain hundreds of documents, even carefully selected subsets may miss important sub-themes or variations. CCS shows a potentially valuable property for larger topics: its hierarchical summarisation approach scales with topic size, preserving coverage of diverse sub-themes that fixed-size selection may miss.

Effectiveness Nuances Across Size Deciles

The relationship between topic size and method effectiveness shows additional nuance when examined at finer granularity. Figure 3 displays MAS per topic size decile. Notably, SCS appears to demonstrate consistent performance across the initial six deciles. CCS then shows improved performance for the seventh and eighth deciles, before SCS regains dominance for the largest topics. This suggests that while CCS may outperform SCS for some larger topics, it also appears to have a saturation point, likely due to a bottleneck in hierarchical compression of information. This scale-dependent effectiveness suggests that practical applications might benefit from exploring hybrid approaches that adaptively select between methods based on topic size. Our analysis indicates that these patterns persist when controlling for topic cohesion, suggesting that the observed effects may be genuinely related

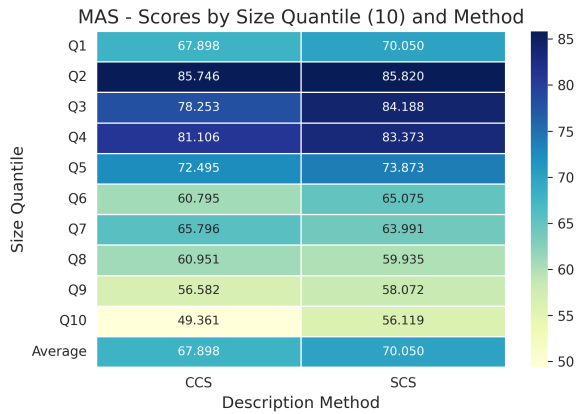


Figure 3: Mean Aspect Score per topic size decile.

to scale. This highlights topic size as a potentially important factor in designing and evaluating topic description pipelines for scientific corpora.

8.3 Qualitative Analysis

To complement our quantitative results, we conducted a targeted qualitative analysis of 45 topic descriptions. We examined 15 top-scoring, 15 low-scoring and 15 descriptions with diverging SUPERT and LLM-as-Judge scores. This enabled us to examine the behaviors of the methods beyond aggregate metrics. Examples illustrating content quality across different models and methods are provided in Appendix I.

Characteristics of Selective Context Summarisation Our analysis suggests that SCS consistently generates clear and coherent summaries in high-scoring cases, with strong alignment to the provided topic terms and good coverage of central concepts (see Example 1). It appears to demonstrate resilience to incoherence in topic terms (Example 2), as any inconsistencies in the topic terms do not compound through multiple summarisation layers. SCS descriptions tend to maintain coherence across different datasets and topic modeling backbones, indicating potential transferability. However, in low-performing cases, particularly when topic terms are overly general or lossy, the method tends to produce generic or shallow outputs. This limitation appears exacerbated when the selected representative documents contain primarily general knowledge rather than specific insights. We observe that SCS capitalize on well-selected topic terms from the underlying model, creating a synergistic effect where strong topic models may yield better descriptions.

Characteristics of Compressed Context Summarisation CCS exhibits distinctive strengths in handling complex or technical topics, often producing more detailed descriptions than SCS. However, this method shows lower alignment with the original topic terms, in several cases generating dense and nuanced content that only partially connects to the provided terms. This misalignment creates challenges in verifying how faithfully the description represents the intended topic (see Example 3). The hierarchical summarisation approach in CCS appears to struggle with effectively prioritizing the most important content, often resulting in information overflow manifested as lengthy lists or excessive detail. This limitation may stem from "document grounding distance" effects in the hierarchical summarisation process, which may not optimally distinguish central from peripheral information. Finally, CCS demonstrates greater sensitivity to topic term quality, with more frequent failures when topic terms are incoherent (see Example 4 comparatively to Example 2).

9 Conclusion & Future Work

We explored Topic Description for Scientific Corpora, an approach to creating structured, document-based summaries that go beyond term lists. To investigate this, we adapted two LLM-based pipelines: Selective Context Summarisation (SCS) and Compressed Context Summarisation (CCS). Our preliminary findings suggest that SCS tends to achieve better performance across datasets and topic modeling backbones, while CCS shows potential advantages for large topics due to its scalable, recursive structure. Our observations highlight an interesting trade-off between selective and compressed strategies: SCS appears to excel in precision and stability, while CCS may offer broader coverage for large-scale topics. The scale-dependent effectiveness patterns we observed suggest that topic size represents an important consideration for practical deployment. Together, they provide initial insights for developing interpretable topic representations in scientific domains.

This work opens several directions for further exploration, including methodological improvements and practical applications. Instead of single-vector retrieval in SCS, future work could examine more fine-grained retrieval strategies to improve coverage and adaptability for complex or broad topics.

Limitations

Despite our multi-faceted evaluation approach, several limitations of the study require further discussion. First, we do not include human assessment. Although we combine SUPERT and LLM-as-a-Judge to approximate quality, expert feedback would be valuable, especially in scientific domains where interpretability and factual accuracy benefit from domain knowledge. The use of both SUPERT and LLM-based evaluation offers complementary strengths: SUPERT captures content relevance via semantic similarity, while LLM-as-a-Judge enables structured, fine-grained evaluation. This dual setup may mitigate some metric-specific biases, though it cannot fully substitute for human judgment.

This challenge is compounded by limitations in the topic modeling stage itself. The quality of topic descriptions is directly tied to the coherence and relevance of the underlying topics and their terms. Despite optimization, CTM often produced noisy or domain-unspecific topics. Similarly, TopicGPT occasionally generated topics that were overly broad or narrowly scoped. These issues affected the resulting descriptions, even with grounded generation. This dependence on topic model quality represents a central limitation in our current investigation. In addition, our study does not include comparisons against simple or established baselines, which would help contextualize the performance of the proposed pipelines.

However, such limitations are inherent to real-world applications (academic and industrial alike) when attempting to gain an overview of large-scale scientific corpora. Our analysis engages with these challenges rather than avoiding them, which we believe is valuable for understanding practical deployment considerations.

Additionally, while the chosen LLMs are among the strongest available models, their outputs remain sensitive to prompt design and can hallucinate content. Our pipelines use a fixed 3A-COT-derived prompting strategy, but prompt wording significantly affects LLM output. No ablation or robustness analysis was conducted to assess this sensitivity in our current work. Moreover, even strong LLMs are prone to hallucination, especially when context is sparse or ambiguous. This is only partially addressed by the factuality criterion in our LLM-as-a-Judge evaluation.

Finally, our evaluation focuses on English scientific abstracts, raising questions about the generalis-

ability of our findings to full-text documents, other genres like patents, and non-English data. The datasets themselves are closely tied to computer science, limiting insights into whether our findings generalise to other disciplines such as the social sciences or biology. These areas require further investigation.

References

- Areej Alokaili, Nikolaos Aletras, and Mark Stevenson. 2020. [Automatic generation of topic labels](#). *CoRR*, abs/2006.00127.
- Hosein Azarbyonad, Zubair Afzal, and George Tsatsaronis. 2023. [Generating topic pages for scientific concepts using scientific publications](#). In *Advances in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2-6, 2023, Proceedings, Part II*, volume 13981 of *Lecture Notes in Computer Science*, pages 341–349. Springer.
- Ibrahim Al Azhar, Venkata Devesh Reddy, Hamed Al-hoori, and Akhil Pandey Akella. 2025. [Limtopic: Llm-based topic modeling and text summarization for analyzing scientific articles limitations](#). *CoRR*, abs/2503.10658.
- Shraey Bhatia, Jey Han Lau, and Timothy Baldwin. 2016. [Automatic labelling of topics with neural embeddings](#). In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 953–963, Osaka, Japan. The COLING 2016 Organizing Committee.
- Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2021. [Pre-training is a hot topic: Contextualized document embeddings improve topic coherence](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 759–766, Online. Association for Computational Linguistics.
- David M Blei. 2012. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.
- Jaime G. Carbonell and Jade Goldstein. 1998. [The use of mmr, diversity-based reranking for reordering documents and producing summaries](#). In *SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28 1998, Melbourne, Australia*, pages 335–336. ACM.
- Jonathan D. Chang, Jordan L. Boyd-Graber, Sean Gerish, Chong Wang, and David M. Blei. 2009. [Reading tea leaves: How humans interpret topic models](#). In *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009. Proceedings of a meeting held 7-10 December 2009, Vancouver, British*

- Columbia, Canada, pages 288–296. Curran Associates, Inc.
- Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. 2020. [SPECTER: document-level representation learning using citation-informed transformers](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2270–2282. Association for Computational Linguistics.
- DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 81 others. 2024. [Deepseek-v3 technical report](#). *CoRR*, abs/2412.19437.
- 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.
- Yang Gao, Wei Zhao, and Steffen Eger. 2020. [SUPERT: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1347–1354, Online. Association for Computational Linguistics.
- Maarten Grootendorst. 2022. [Bertopic: Neural topic modeling with a class-based tf-idf procedure](#). *Preprint*, arXiv:2203.05794.
- Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. 2024. [A survey on llm-as-a-judge](#). *CoRR*, abs/2411.15594.
- Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-Bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, and 79 others. 2025. [Gemma 3 technical report](#). *CoRR*, abs/2503.19786.
- JoongHoon Kim, Sangmin Lee, Seung Hun Han, Saeran Park, Jiyeon Lee, Kiyoon Jeong, and Pilsung Kang. 2023. Which is better? exploring prompting strategy for llm-based metrics. In *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems*, pages 164–183.
- Jey Han Lau, Karl Grieser, David Newman, and Timothy Baldwin. 2011. [Automatic labelling of topic models](#). In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1536–1545, Portland, Oregon, USA. Association for Computational Linguistics.
- Christoph Leiter and Steffen Eger. 2024. Prexme! large scale prompt exploration of open source llms for machine translation and summarization evaluation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11481–11506.
- Christoph Leiter, Juri Opitz, Daniel Deutsch, Yang Gao, Rotem Dror, and Steffen Eger. 2023. The eval4nlp 2023 shared task on prompting large language models as explainable metrics. In *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems*, pages 117–138.
- Yao Lu, Yue Dong, and Laurent Charlin. 2020. [Multi-XScience: A large-scale dataset for extreme multi-document summarization of scientific articles](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8068–8074, Online. Association for Computational Linguistics.
- Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, 2(11):205.
- Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.
- Leland McInnes, John Healy, and James Melville. 2020. [Umap: Uniform manifold approximation and projection for dimension reduction](#). *Preprint*, arXiv:1802.03426.
- Zach Nussbaum, John X. Morris, Brandon Duderstadt, and Andriy Mulyar. 2024. [Nomic embed: Training a reproducible long context text embedder](#).
- Chau Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. 2024. [TopicGPT: A prompt-based topic modeling framework](#). 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 2956–2984, Mexico City, Mexico. Association for Computational Linguistics.
- Cristian Popa and Traian Rebedea. 2021. [BART-TL: weakly-supervised topic label generation](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 1418–1425. Association for Computational Linguistics.
- Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora.

- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408.
- Domenic Rosati. 2022. [Moving beyond word lists: towards abstractive topic labels for human-like topics of scientific documents](#). *CoRR*, abs/2211.05599.
- 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, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Kartik Shinde, Trinita Roy, and Tirthankar Ghosal. 2022. [An extractive-abstractive approach for multi-document summarization of scientific articles for literature review](#). In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 204–209, Gyeongju, Republic of Korea. Association for Computational Linguistics.
- Silvia Terragni, Elisabetta Fersini, and Enza Messina. 2021. Word embedding-based topic similarity measures. In *International conference on applications of Natural Language to information systems*, pages 33–45. Springer.
- Huy Quoc To, Ming Liu, Guangyan Huang, Hung-Nghiep Tran, Andr’e Greiner-Petter, Felix Beierle, and Akiko Aizawa. 2024. [Skt5scisumm – revisiting extractive-generative approach for multi-document scientific summarization](#). *Preprint*, arXiv:2402.17311.
- Pancheng Wang, Shasha Li, Kunyuan Pang, Liangliang He, Dong Li, Jintao Tang, and Ting Wang. 2022. [Multi-document scientific summarization from a knowledge graph-centric view](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6222–6233, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024a. [Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference](#). *CoRR*, abs/2412.13663.
- Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024b. [Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference](#). *Preprint*, arXiv:2412.13663.
- William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS)*, 28(4):1–38.
- Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2022. [PRIMERA: Pyramid-based masked sentence pre-training for multi-document summarization](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5245–5263, Dublin, Ireland. Association for Computational Linguistics.
- 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, Jixi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. [Qwen2.5 technical report](#). *CoRR*, abs/2412.15115.
- Yongbing Zhang, Shengxiang Gao, Yuxin Huang, Zhengtao Yu, and Kaiwen Tan. 2024. [3a-cot: An attend-arrange-abstract chain-of-thought for multi-document summarization](#). *International Journal of Machine Learning and Cybernetics*.

A Summarisation Prompts

We used the deepseek-v3 model to generate topic descriptions across all methods. To ensure consistency and structure in the outputs, we define a fixed system message and adopt a 3-step prompting framework inspired by the 3A-COT method (Zhang et al., 2024). This includes *attending* to key aspects, *arranging* extracted information, and generating the final *abstract*. The prompt templates used are provided below.

System Prompt
You are a scientific research assistant who organizes information into structured markdown documents. Your writing style sounds natural and professional. Avoid using Marketing and HR language.

Prompt 1: System prompt used for topic description generation

Attend Prompt

[DOCUMENTS]

What are the research purposes in this document?
What are the research object in this document?
What are the research methods in this document?
What are the research result in this document?
What are the main findings in this document?

Please answer the above questions:

Prompt 2: Attend prompt for extracting key information

Arrange Prompt

[ATTEND_OUTPUT]

Organize the above important information. Arrange this information in a logical order or relevance to build a coherent narrative, and consider how information from different articles can be combined to complement and connect with each other.

Prompt 3: Arrange prompt for structuring extracted content

Abstract Prompt

[ABSTRACTS]

–

[ARRANGE_OUTPUT]

Based on the above abstracts, key information and the keywords: {topic_words}, write a summary.

Make sure to include key information, research objectives and ideas. The summary should be structured as clean MARKDOWN with ONLY the following Headings:

Brief Introduction into the Topic, Key Research Objects, Key Research Methods, Key Research Purpose.

Each Heading should have only Keypoints listed. Avoid the use of additional MARKDOWN subsections. Avoid adding your own opinion, interpretation, or conclusions or Future Work. Use the information provided in the text only.

Prompt 4: Abstract prompt for final topic description generation

B Quantum Dataset Query

In Query 1, we present the full boolean query used for collecting the source documents for the Quantum Computing dataset. Specially, the query is specialized in the hardware part of this scientific field.

C Topic Models

This section presents implementation details and results of three topic modeling approaches used in our comparative analysis.

C.1 Overview

For all approaches involving training, we perform hyper-parameter optimization to find the best coherence and diversity metrics for each combination of topic model and dataset. For coherence, we use the Gensim implementation of the Coherence Model (Řehůřek and Sojka, 2010), specifically its default C_V metric (Röder et al., 2015). For diversity, we calculate the Inverted Rank-Biased Overlap (Webber et al., 2010; Terragni et al., 2021) of the top 10 terms representing the topics.

Table 3 shows that CTM achieves the highest coherence scores, followed by BERTopic and TopicGPT. Conversely, TopicGPT models the greatest number of topics on average, followed by BERTopic and then CTM. Moreover, while CTM achieves the most robust diversity scores on the three datasets, BERTopic and TopicGPT come on par in the ACL and the NIPS datasets, respectively. Finally, manual inspection shows that TopicGPT generally tends to construct more specific topics, with less documents per topic. While CTM has the contrary behaviour presenting broader topics with larger document sets. This observation explains in part the higher diversity scores of CTM. In turn, the low coherence scores of TopicGPT reflects the zero-shot decoupling from the underlying respective corpus.

C.2 Implementation Details

CTM In the CTM backbone, we use the GitHub implementation⁶ of the original contribution (Bianchi et al., 2021). Here, we optimize

⁶<https://github.com/MilaNLPProc/contextualized-topic-models>

Boolean Query

```
TITLE-ABS-KEY ( "quantum comput*" OR "quantum processor" OR "quantum circuit" OR "quantum logic gate" OR "quantum gate" OR "logical qubit" OR qubit OR "quantum system" OR "quantum information processing" OR "quantum control" OR "quantum electronics" OR "quantum hardware" OR "noisy intermediate-scale quantum era" OR "NISQ" OR "multiqubit circuit" OR "quantum simulation" OR "quantum simulator" ) AND TITLE-ABS-KEY ( ( cryogen* OR "magnetic field" OR laser OR photoluminescence OR silicon OR "electric fields" OR magnetism OR fluorescence ) OR ( "neutral atom" OR "cold atom" OR "trap*atom" OR "atom trap" OR "rydberg" OR atoms OR "optical lattice*" OR magic OR "optical tweezer*" OR strontium OR ytterbium OR "photonic crystal fibre" ) OR ( "ion traps" OR "trapped ions" OR "ions" OR "integrated waveguide" OR "laser induced deep etching" OR "on-chip coupling" ) OR ( superconduct* OR "SQUIDS" OR "Josephson junction device*" OR "indium bump" OR "NbN films" OR "single flux quantum" OR "quantum flux" OR "SQUID" ) OR ( center OR diamond OR "NV center" OR "NV centre" OR "color centre" OR "colour center" OR "silicon vacancy centre" OR "silicon vacancy center" ) OR ( photon* OR "gaussian boson sampl*" OR "squeezed light source" OR niobate OR "superconducting nanowire single-photon detector" OR "SNSPD" ) OR ( topology OR "topological quantum computing" OR "topological insulator*" ) OR ( semiconductor OR "molecular beam epitaxy" OR "semiconducting*" OR "crystal lattice*" OR phonons ) ) AND PUBYEAR > 2009 AND PUBYEAR < 2026
```

Query 1: Boolean query used for collecting the source documents for the Quantum Computing dataset.

over four hyper-parameters: number of topics (40—100), number of epochs (10—50), activation function (`{sigmoid, relu, softplus}`), number of neurons (100—500). All other hyper-parameters use standard values from the implementation.

BERTopic For this approach, we use the standard BERTopic package⁷. This standard pipeline consists of mainly three stages: Embedding (Em.) stage, Dimensionality Reduction (DR) stage, the

⁷<https://maartengr.github.io/BERTopic/index.html>

Clustering (Cl.) stage, and Topic Representation (TR) stage. For the Em stage, we use the `nomic-ai/modernbert-embed-base`⁸ model (Nussbaum et al., 2024), which is an embedding model trained on the ModernBERT (Warner et al., 2024b) encoder. For the DR and Cl. stages, we opt for the standard pairing with UMAP (McInnes et al., 2018) and HDBSCAN (McInnes et al., 2017). Finally, in the TR stage, we use class-TFIDF, which was introduced in (Grootendorst, 2022). Overall, we optimize four hyper-parameters: UMAP - number of neighbors (5—50), number of components (2—15) and min. distance (0.0—0.5); HDBSCAN - min. cluster size (10—50). For UMAP, we fix the metric to cosine, and euclidean for HDSCAN. All other hyper-parameters use standard values from their implementations.

TopicGPT We follow the original TopicGPT pipeline (Pham et al., 2024), using the open-source implementation available at GitHub⁹ and altering only the document-assignment stage to align with BERTopic and CTM. For topic generation, we randomly sample 1,000 documents from each dataset and leverage GPT-4 to propose an initial set of top-level topics, which we then iteratively refine into subtopics to build a complete hierarchical structure. In the subsequent assignment phase—applied to the full datasets—we replace TopicGPT’s default routine (which, for each document, prompts GPT-3.5-turbo with the finalized hierarchy and returns the best-matching topic with a supporting quote) with a two-part prompt to GPT-3.5-turbo: (i) assign each document to its best-matching topic in our hierarchy; and (ii) extract ten representative keywords per document. Finally, we post-process all extracted keywords for each topic by tokenizing them on whitespace, converting to lowercase, stripping punctuation, aggregating token frequencies, and selecting the ten most frequent tokens per topic—thereby exactly matching the output format of our BERTopic and CTM backbones.

C.3 Results

Table 3 presents topic modeling evaluation results across three datasets (ACL, NIPS, and Quantum) for three different topic modeling approaches: CTM, BERTopic, and TopicGPT. The evaluation metrics used in the comparison are Coherence, Di-

⁸<https://huggingface.co/nomic-ai/modernbert-embed-base>

⁹<https://github.com/chtmp223/topicGPT>

Document Assignment Prompt Template

You will receive a document and a topic hierarchy. Assign the document to the most relevant topic of the hierarchy. Then, output the topic label, and supporting keywords from the document. DO NOT make up new topics or keywords.

[Topic Hierarchy]

{tree}

[Instructions]

1. Topic label must be present in the provided topic hierarchy. You MUST NOT make up new topics.
2. The keywords must be taken from the document. You MUST NOT make up keywords or quotes. All keywords MUST NOT contain stop words.

[Document]

{Document}

Double check that your assignment exists in the hierarchy! Your response should be in the following format:

[Topic Level] Topic Label: keyword1, keyword2, etc

Your response:

Prompt 5: Prompt template used for document-to-topic assignment in the TopicGPT adaptation.

versity, and Number of Topics (N.Topics). Our findings show that CTM achieves the highest coherence scores across all three datasets (0.664 for ACL, 0.601 for NIPS, and 0.692 for Quantum). It also maintains high diversity scores above 0.94 for all datasets. BERTopic shows moderate coherence performance (0.504 for ACL, 0.458 for NIPS, and 0.546 for Quantum), with somewhat lower diversity metrics, particularly for the Quantum dataset (0.799). TopicGPT demonstrates coherence scores between 0.458 and 0.526 across datasets, with strong diversity in the NIPS dataset (0.963) but lower diversity for ACL (0.881) and Quantum (0.809). Regarding the number of topics identified, TopicGPT produces substantially more topics than the other approaches, particularly for NIPS (276). BERTopic identifies the fewest topics

overall with just 24 for the NIPS dataset. For the ACL dataset, the number of topics is more consistent across models (CTM: 72, BERTopic: 70, TopicGPT: 66). The Quantum dataset shows moderate variation, with CTM identifying 59 topics, BERTopic 72, and TopicGPT significantly more at 169.

D Evaluation Prompts

We evaluate summaries along four dimensions: relevance, coherence, factuality, and fluency. Each is scored independently using a dedicated prompt, detailed below.

Aspect Definitions

- **Relevance:** The rating measures how well the summary captures the key points of the documents. Consider whether all and only the important aspects are contained in the summary.
- **Coherence:** This rating evaluates how seamlessly the sentences of the summary flow together, creating a unified whole. Assess how smoothly the content transitions from one point to the next, ensuring it reads as a cohesive unit.
- **Factuality:** This rating gauges the accuracy and truthfulness of the information presented in the summary compared to the original documents. Scrutinize the summary to ensure it presents facts without distortion or misrepresentation, staying true to the source content’s details and intent.
- **Fluency:** This rating evaluates the clarity and grammatical integrity of each sentence in the summary. Examine each sentence for its structural soundness and linguistic clarity.

E Gemma-3-27B & Mistral-7B-OpenOrca Results

To complement the main results, we report the MAS obtained using Gemma-3-27B and Mistral-7B-OpenOrca in Table 4. These models provide additional perspectives on the quality of the generated descriptions and help examine the consistency of trends observed with Qwen-2.5-7B-Instruct.

To further examine inter-model agreement, we compute Kendall’s τ_b between the rankings of

TM	Dataset	Coherence	Diversity	N.Topics
CTM	ACL	0.664	0.994	72
	NIPS	0.601	0.949	38
	Quantum	0.692	0.996	59
BERTopic	ACL	0.504	0.972	70
	NIPS	0.458	0.930	24
	Quantum	0.546	0.799	72
TopicGPT	ACL	0.458	0.881	66
	NIPS	0.472	0.963	276
	Quantum	0.526	0.809	169

Table 3: Topic modeling evaluation results across three scientific datasets.

Method	Metric	ACL			NIPS			Quantum		
		CTM	BERTopic	TopicGPT	CTM	BERTopic	TopicGPT	CTM	BERTopic	TopicGPT
SCS	MAS-Mistral	74.913	74.306	76.307	75.497	75.109	77.240	72.077	73.127	76.120
	MAS-Gemma	81.537	85.163	87.411	82.691	83.865	89.400	82.006	85.255	89.395
CCS	MAS-Mistral	73.092	71.207	74.471	66.384	70.771	75.201	65.031	72.758	74.528
	MAS-Gemma	81.483	85.155	86.799	81.194	83.370	88.549	80.204	85.120	89.125

Table 4: MAS scores across methods, datasets, and topic modeling backbones using Mistral and Gemma as judge models.

Evaluation Prompt Template
<p>Instruction: In this task you will evaluate the quality of a summary written for multiple documents. To correctly solve this task, follow these steps: 1. Carefully read the document, be aware of the information it contains. 2. Read the proposed summary. 3. Rate each summary on a scale from 0 (worst) to 100 (best) by its {aspect}. Decimals are allowed.</p> <p># Definition: {definition}</p> <p>Source documents: {source}</p> <p>Summary: {summary}</p> <p>Score:</p>

Prompt 6: Evaluation prompt template used for scoring topic descriptions across relevance, factuality, coherence, and fluency

method–size combinations (i.e., CCS/SCS across the four topic size categories: *Small*, *Medium-Small*, *Medium-Large*, and *Large*) for each pair of judge models. We evaluate agreement across the full 8-item ranking. This provides a single τ_b score per pair, reflecting overall alignment in method preferences across topic sizes. As shown in Table 5, Qwen2.5-7B-Instruct shows moderate to strong alignment with both Gemma-3-27B and Mistral-7B-OpenOrca, while Gemma-3-27B and Mistral-7B-OpenOrca exhibit weaker agreement.

Model Pair	τ_b	p
Qwen & Gemma	0.6183	0.0340
Qwen & Mistral	0.6910	0.0178
Gemma & Mistral	0.2857	0.3988

Table 5: Kendall’s τ_b between full method–size rankings of each model pair.

F Impact of Document Sampling in LLM-Eval

In order to analyze the impact of using subsets of documents of topics as reference documents in the LLM-Eval strategies, we present a detailed visual-

ization of the Quantum dataset results in Figure 4 across all five document draws for each TM and TD approach. Similar patterns are observed for the other datasets.

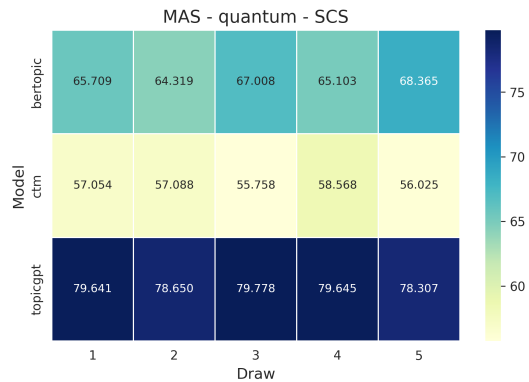
From visual inspection of Figure 4, we observe that scores remain relatively stable across different document draws for the same TM and TD method. When fixing a topic modeling approach and a topic description pipeline, the fluctuations in LLM-Eval MAS are generally small, with most variations remaining within 5 points to the mean on our 100-point scale.

While a comprehensive variance analysis across all datasets would provide further statistical rigor, the consistency observed in the Quantum dataset suggests that our sampling approach may produce reliable evaluations. The observed stability indicates that randomly sampling 10 documents five times appears to provide a reasonable approximation of how a topic description would be evaluated against the full document collection.

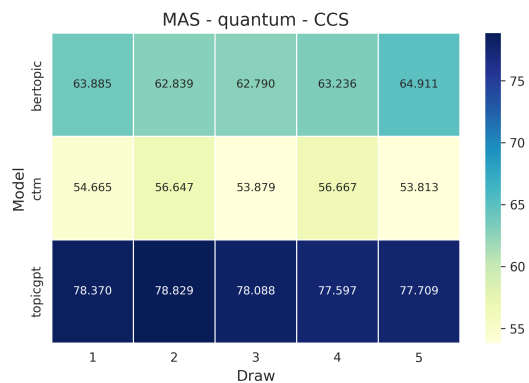
The observed consistency across document draws supports our decision to use this sampling approach as a practical solution to the context window limitations of LLMs. While a more exhaustive analysis would be valuable for future work, the current evidence suggests that our methodology may provide reliable evaluations of topic descriptions despite using only subsets of the complete document collections.

G Effect of Topic Cohesion on Mean Aspect Score (MAS)

To study the impact of topic cohesion on the quality of topic descriptions, We compute the mean cosine embedding distance among all documents for each topic. We call this indicator "Topic Cohesion." Figure 5 shows the MAS distributions for all topics grouped by their topic cohesion quartile. Interestingly, topic cohesion appears to play a relatively minor role in the MAS distributions across all quartiles. There is a slight downward trend indicating some anticipated TD quality degradation towards topics of low cohesion. However, this effect appears minor among all TD approaches, only becoming more pronounced in the low cohesion quartile. Even there, the best topic descriptions of SCS and CCS are competitive with TD's best scores in the more cohesive quartiles.



(a) SCS



(b) CCS

Figure 4: LLM-Eval MAS for every draw of 10 documents per topic.

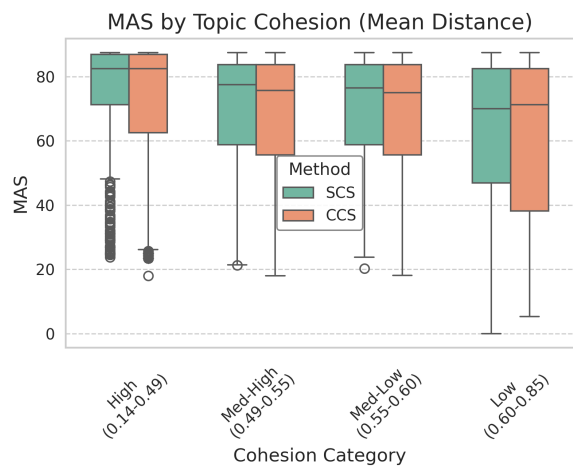


Figure 5: Box-plots of MAS distributions from Qwen conditioned on topic cohesion quartile.

H Effect of Topic Size and Cohesion on SUPERT

Figure 6 shows the distribution of winners per topic size category based on the SUPERT metric. SCS leads in the first and second quartiles, with CCS gaining a slight edge in the third. In contrast to MAS-Qwen, which shows CCS catching up in the largest category, SUPERT continues to favor SCS in the fourth quartile. This suggests that SCS may be more aligned with SUPERT’s relevance-focused evaluation, even as topic size increases.

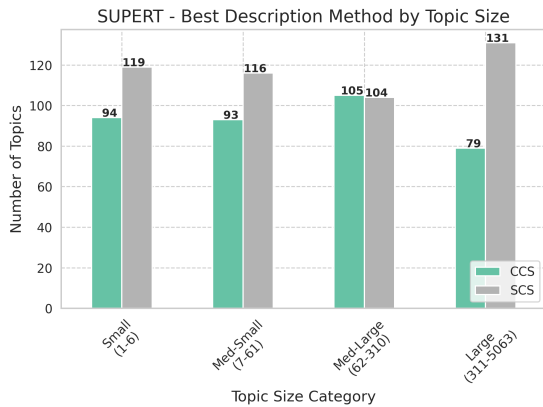


Figure 6: Winner count on SUPERT per topic size quartile over all topic models

Figure 7 shows SUPERT distributions grouped by topic cohesion (mean pairwise cosine distance). Both SCS and CCS appear relatively stable across cohesion levels, with only slight performance degradation from high to low cohesion.

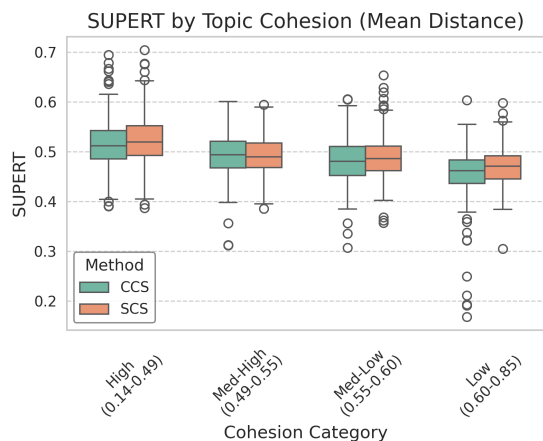


Figure 7: Box-plots of SUPERT distributions conditioned on topic cohesion quartile

The SUPERT experiments on effectiveness nuances across size deciles, to be seen in Figure 8 reveal patterns consistent with the results of LLM-

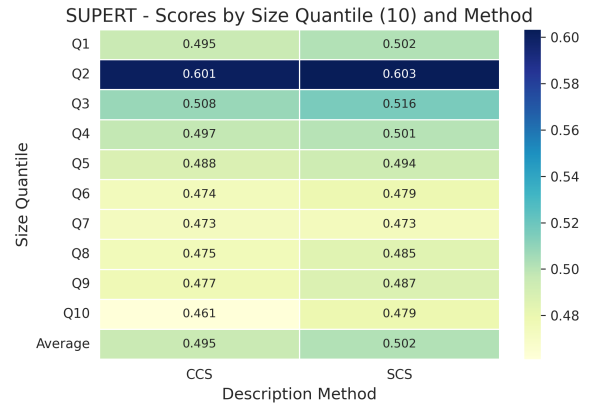


Figure 8: SUPERT Score per topic size decile.

Eval in Section 8.2: method effectiveness appears tied to topic size, with SCS showing advantages for smaller topics (first to sixth decile). The effect of CCS slightly outperforming SCS in LLM-Eval for the larger topics (seventh and eighth deciles) is not to be seen with SCS even achieving minially better scores with SUPERT. Still, the overall trend is replicated from LLM-Eval and this consistency further suggests that scale, besides intrinsic topic quality, drives performance differences and supports the potential viability of adaptive hybrid approaches.

I Examples of generated Topic Descriptions

To demonstrate the results of our topic description pipeline, we present a selection of examples covering different datasets and modelling approaches. Each example provides metadata, the generated summary, and evaluation scores. Together, these elements help contextualize the content and quality of the generated descriptions.

<p>acl-53</p> <p>Dataset: ACL Model: TopicGPT Method: SCS Category: Best Topic Words: data, LOD, digital, linguistics, studies, linked open, humanities, multilingual, modeling SUPERT: 0.7043 LLMEval: 87.5000</p> <p>Topic Summary: Linked Open Data (LOD) in the Humanities Brief Introduction into the Topic Linked Open Data (LOD) is increasingly being integrated into the Humanities to enhance the accessibility, interoperability, and analysis of machine-readable information. Two notable initiatives, LODinG and DigItAnt, exemplify this trend. LODinG explores broad applications across various Humanities disciplines, while DigItAnt focuses on creating, linking, and exploiting LOD lexica for historical linguistics and epigraphy data.</p> <p>Key Research Objects</p> <ul style="list-style-type: none"> - LODinG: Covers digital humanities, computational linguistics, digital lexicography, philology, translatology, cultural and literary studies, media studies, information science, and law. - DigItAnt: Manages LOD lexica and heterogeneous resources in historical linguistics and epigraphy, supporting the H2IOSC Italian infrastructures cluster project. <p>Key Research Methods</p> <ul style="list-style-type: none"> - LODinG: Collects, models, links, releases, and analyzes multilingual data in LOD format, developing a modular, cross-domain Humanities data model. - DigItAnt: Designs architecture for LOD lexica and links Linguistic Linked Open Data (LLOD) with non-LLOD datasets, following FAIR principles. <p>Key Research Purpose</p> <ul style="list-style-type: none"> - LODinG: Improves extraction and modeling of multilingual LOD data, demonstrates LLOD applications in interdisciplinary contexts, and develops a cross-domain data model. - DigItAnt: Aims to build and integrate LOD lexica for historical linguistics and epigraphy, support the H2IOSC project, and contribute to Italian Humanities infrastructure.

Example 1: Example topic description for topic acl-53.

quantum-49**Dataset:** Quantum**Model:** CTM**Method:** SCS**Category:** Ambiguous**Topic Words:** foremost, boast, ineffective, craft, exacerbate, maintains, complication, pack, prevail, upon**SUPERT:** 0.4825**LLMEval:** 62.5**Summary of Quantum Computing and Related Technologies****Brief Introduction into the Topic**

Quantum computing and related technologies focus on advancing fault-tolerant quantum machines, enhancing quantum communication, and exploring novel materials and methods for quantum applications. Research spans from error correction in quantum computing to the development of high-performance quantum platforms, quantum networking, and advanced optical materials.

Key Research Objects

Fault-tolerant quantum machines: Large-scale quantum computers relying on error correction using logical and physical qubits.

Quantum emitters and waveguides: Dipole quantum emitters embedded in dielectric slab waveguides.

Quantum platforms: Diamond with engineered coherent defects for quantum control.

Quantum simulations: SU(3) gauge theories in Quantum Chromodynamics (QCD) using discrete subgroup $S(1080)$.

Quantum networking: Quantum wrapper networking with classical headers and quantum payloads.

Quantum box arrays: Self-assembled porous (metal-)organic networks modified by C60 adsorbates.

Quantum optics: High refractive index polymers for infrared (IR) optics.

Quantum communication: DWDM transmission using ultra-low loss G.654.D fiber and lumped doped fiber amplifiers.

Quantum spin manipulations: Ce^{3+} doped $LiYF_4$ crystals for ENDOR spectroscopy.

Quantum geometry: Desargues and Pappus properties in projective geometry applied to quantum systems.

Key Research Methods

Error correction: Spreading the value of a single logical qubit over multiple physical qubits.

Waveguide analysis: Studying the effective quantum yield of dipole emitters in dielectric slab waveguides.

Materials engineering: Controlling coherent defects in diamond for quantum applications.

Digitization schemes: Practical digitization of SU(3) gauge theories using discrete subgroups.

Noise characterization: Analyzing in-fiber scattering processes in quantum wrapper networking.

Adsorbate modification: Using C60 to modify electronic states in quantum box arrays.

Inverse vulcanization: Synthesizing high refractive index polymers from waste materials.

DWDM transmission: Demonstrating high-speed data transmission using advanced fiber and amplifiers.

ENDOR spectroscopy: Resolving Mims ENDOR spectra for electron-nuclear interactions.

Geometric analysis: Applying Desargues and Pappus properties to quantum systems.

Key Research Purpose To **craft** large, fault-tolerant quantum machines capable of robust computations.

To **boast** the performance of quantum emitters and waveguides by optimizing their properties.

To **maintain** long spin coherence times in diamond-based quantum platforms.

To **pack** the infinite-dimensional gluon field into a minimal qubit representation for QCD simulations.

To **exacerbate** the understanding of noise impacts in quantum networking.

To **prevail** in engineering quantum states in on-surface network architectures.

To **foremost** develop low-cost, high-performance materials for IR optics.

To **complicate** the understanding of high-speed data transmission in optical communication.

To **upon** investigate electron-nuclear interactions for spin manipulations.

To **ineffective** link projective geometry principles to quantum physics and information.

Example 2: Example SCS topic description for topic quantum-49.

<p>acl-20</p> <p>Dataset: ACL Model: BERTopic Method: CCS Category: Ambiguous Topic Words: word2vec, embeddings, vocabulary, wordnet, semantic, semantics, embedding, contextualized, word, linguistic SUPERT: 0.4274 LLMEval: 85.4250</p> <p>Brief Introduction into the Topic The extraction and analysis of health-related information from social media platforms, particularly Twitter, have become critical in pharmacovigilance, public health monitoring, and mental health studies. Advances in Natural Language Processing (NLP) and transformer-based models (e.g., BERT, RoBERTa, BERTweet) have improved the ability to process informal and noisy text, enabling the identification of Adverse Drug Events (ADEs), disease mentions, and psychological conditions such as depression, anxiety, and mood disorders. Social media data, especially tweets, provide valuable insights into public health trends and individual health experiences.</p> <p>Key Research Objects</p> <ul style="list-style-type: none"> - Psychological and Mood-Related Terms: Keywords such as depression, depressed, mood, psychological, and anxiety, relevant to ADE classification and diagnosis. - Tweets: Informal text from Twitter containing colloquial language related to health, drug effects, and psychological conditions. - Disease Mentions: Identification of diseases such as depression and anxiety in Spanish tweets and other social media content. - Adverse Drug Events (ADEs): Mentions of adverse effects from medications and dietary supplements in tweets. - Social Media Platforms: Twitter, Reddit, and health forums as primary sources of health-related data. <p>Key Research Methods</p> <ol style="list-style-type: none"> 1. Transformer-Based Models <ul style="list-style-type: none"> - BERT, RoBERTa, BERTweet, and CT-BERT are fine-tuned for tasks such as classification, Named Entity Recognition (NER), and information extraction. - Ensemble Methods: Combining multiple BERT variants to improve accuracy in classifying health-related tweets. 2. Large Language Models (LLMs) <ul style="list-style-type: none"> - Data Augmentation: Generating synthetic data to address data imbalance in ADE extraction and disease mention detection. - Retrieval-Augmented Generation (RAG): Mapping informal tweet language to standardized terms like MedDRA Preferred Terms. 3. Advanced NLP Techniques <ul style="list-style-type: none"> - Transfer Learning: Fine-tuning multilingual models like mBERT for disease mention extraction in Spanish tweets. - Knowledge Graph Integration: Enhancing models with external evidence from medical gazetteers and ontologies. - Flair-NER Framework: Evaluating embeddings for capturing complex disease mentions in social media text. 4. Preprocessing and Optimization <ul style="list-style-type: none"> - Domain-Specific Preprocessing: Cleaning and preparing social media text for analysis. - Hyperparameter Optimization: Enhancing model performance through tuning. <p>Key Research Purpose</p> <ul style="list-style-type: none"> - Enhance ADE Discovery: Improve the accuracy and robustness of ADE extraction and normalization from social media data. - Improve Disease Mention Extraction: Develop advanced NER systems for identifying diseases in informal text, particularly in non-English languages like Spanish. - Standardize Health-Related Terms: Map colloquial language to standardized medical terms (e.g., MedDRA Preferred Terms) for better pharmacovigilance and health monitoring. - Promote Open Research: Provide open-source code and methodologies to foster collaboration and further innovation in the field. - Monitor Public Health Trends: Leverage social media data for real-time monitoring of public health issues, including COVID-19 and psychological conditions.
--

Example 3: Example topic description for topic acl-20.

<p>quantum-49</p> <p>Dataset: Quantum Model: CTM Method: CCS Category: Worst Topic Words: foremost, boast, ineffective, craft, exacerbate, maintains, complication, pack, prevail, upon SUPERT: 0.1685 LLMEval: 7.1150</p> <p>Brief Introduction into the Topic The topic revolves around the analysis of certain systems or strategies that are deemed foremost in their field but may boast features that are ultimately ineffective. These systems often craft solutions that inadvertently exacerbate existing issues, leading to further complication. Despite these challenges, the systems maintain their prominence, as they are designed to pack significant functionality. The research aims to explore why such systems prevail and the factors that drive their continued use upon further scrutiny.</p> <p>Key Research Objects The primary objects of research include systems or strategies that are considered leading in their domain. These objects are characterized by their advanced features, which are often highlighted as strengths but may contribute to inefficiencies or unintended consequences. The study focuses on understanding the dynamics of these systems and their impact on the broader context in which they operate.</p> <p>Key Research Methods The research employs a combination of qualitative and quantitative methods to evaluate the effectiveness of the systems under study. This includes analyzing case studies, conducting surveys, and performing comparative assessments to identify patterns and outcomes. The methods aim to uncover the reasons behind the systems' continued use despite their potential drawbacks.</p> <p>Key Research Purpose The purpose of the research is to critically examine the systems that are widely regarded as top-tier in their field. It seeks to identify the factors that contribute to their perceived success, as well as the unintended consequences that may arise from their implementation. The study aims to provide a comprehensive understanding of why these systems prevail and how they impact their respective domains.</p>
--

Example 4: Example CCS topic description for topic quantum-49.

HalluTree: Explainable Multi-Hop Hallucination Detection for Abstractive Summarization

Oskar Oomen*, Daniel Orshansky*, Naaisha Agrawal, Ryan Lagasse

AlgoVerse

ryan@algoverseairesearch.com

Abstract

Black-box verifiers for abstractive summaries often struggle with complex claims that require multi-hop reasoning, and they typically provide a single verdict without an interpretable rationale. As a result, it becomes difficult to understand or audit their failures. We address this with HalluTree, a framework that models verification as an interpretable claim tree. HalluTree first decomposes summaries into subclaims, classifying each into two types – extractive (directly verifiable against evidence) or inferential (requiring reasoning) – which follow distinct verification paths. Extractive claims are robustly verified against evidence using an ensemble of lightweight NLI models. Crucially, inferential claims trigger a process that generates a natural program – an explicit reasoning chain that integrates supporting evidence and logical steps – which is then executed to determine the claim’s validity. Evaluation on the LLM-AggreFact benchmark demonstrates HalluTree’s effectiveness: it achieves performance competitive with top-tier black-box models, including Bespoke-MiniCheck, while providing transparent and auditable reasoning programs for every inferential judgment. This combination of competitive accuracy and high interpretability offers a significant advance over opaque, single-classification verifiers.

1 Introduction

Large language models (LLMs) frequently hallucinate, producing content that is factually unsupported or incorrect (Dmonte et al., 2025; Huang et al., 2023, 2025). Even when grounded in source documents, LLM-generated summaries may contain contradictions or unverifiable statements, which can mislead readers and contribute to the spread of misinformation (Huang et al., 2025; Scirè et al., 2024). Ensuring the factual consistency of such outputs is therefore critical, particularly in

domains where accuracy is paramount. In addition to raw accuracy, the explainability of these classifiers is increasingly important for transparency and human validation (Wang and Shu, 2023; Dammu et al., 2024). Without clear rationales, even correct predictions may be difficult to trust, and incorrect ones may be difficult to diagnose.

Existing work on grounded factuality verification spans a variety of strategies, including entailment-based classification, question-answering formulations, and more recent LLM-driven verification pipelines (Dmonte et al., 2025; Huang et al., 2023). While these approaches have achieved strong results in certain settings, they often provide limited transparency into the inferences behind complex judgments, and struggle with claims that require multi-hop reasoning across dispersed evidence (Belém et al., 2025). Additionally, despite the lack of fine-grained classification and weak interpretability, frontier LLMs with few-shot prompting can achieve top-tier performance even compared to the strongest specialized baselines, but still have room for improvement in complex reasoning tasks (Seo et al., 2025). This highlights the need for strong-performing and explainable verification methods that are robust to challenging multi-hop reasoning.

We propose a dual-path verification framework that decomposes a generated summary into subclaims and organizes verification results in an interpretable claim tree, with the summary as the root. The summary is first decomposed and decontextualized with an LLM to preserve original, ensuring potential hallucinations are not either introduced or inadvertently corrected away. The system then filters out unverifiable subclaims (advice, opinions, or other statements not containing factually verifiable assertions) before classifying each subclaim as extractive (directly checkable against the source) or inferential (requiring multi-hop reasoning over evidence to verify).

*Equal contribution

Extractive subclaims are verified using two existing lightweight NLI-based hallucination detectors, LettuceDetect (Ádám Kovács and Recski, 2025) and MiniCheck-FT5 (Tang et al., 2024a), followed by evidence retrieval for ease of understanding.

Inferential subclaims trigger a reasoning pathway that gathers supporting facts from the source and beyond, which may include textual evidence, mathematical reasoning, logical inference, or unrelated elementary knowledge that need not be verified. We attach the chain-of-thought trace from the LLM when verifying the subclaim based on the supporting facts in the tree to boost auditability and ease of understanding. More importantly, these supporting facts are developed by the LLM into a natural program – a natural language-based chain of reasoning which explicitly sets out premises and rigorously demonstrates how they are composed to support the claim (Ling et al., 2023). The reasoning of this natural program is then executed and validated step-by-step by an LLM for verification, with dynamic error correction applied to detected mistakes in the program.

The verification process is represented as a hierarchical claim-tree for easy visualization, where the root corresponds to the full summary, subclaims are the children of the root, and each subclaim’s children are either the evidence or supporting facts which motivate its classification, making the motivation for each subclaim’s classification clear. The summary is deemed supported only if all verifiable subclaims (extractive and inferential) are supported, ensuring conservative, evidence-grounded judgments that expose the full chain of evidence and reasoning behind complex decisions, yielding stronger reliability and greater interpretability than prior single-pass verifiers.

Our primary contributions are:

- **Typed, dual-path verification.** A framework that separates extractive and inferential subclaims, verifying the former with lightweight NLI models and the latter through natural programs that rigorously combine evidence and reasoning, yielding stronger multi-hop performance.
- **Interpretable claim tree.** A hierarchical structure that links each subclaim to evidence or inferences, with natural programs making inferential reasoning rigorous and auditable rather than opaque.
- **Empirical gains.** Evaluation on AggreFact-CNN (Hermann et al., 2015), AggreFact-XSUM (Narayan et al., 2018), TofuEval (Tang et al., 2024b), and WiCE (Kamoi et al., 2023) using balanced accuracy, showing improvements over strong baselines, including GPT-4o, while also providing enhanced human-interpretability and also setting a new state-of-the-art on WiCE.

2 Related Work

Entailment-Based Methods. A major line of work frames grounded factuality classification as natural language inference (NLI) (Dmonte et al., 2025; Huang et al., 2023). Several methods fine-tune an NLI model to predict the faithfulness of a generated claim, sentence, or summary against the ground-truth source documents (Goyal and Durrett, 2020; Tang et al., 2024a; Laban et al., 2022; Kryscinski et al., 2020; Zha et al., 2023). These approaches are very efficient compared to methods involving LLMs but often offer poor interpretability due to the low granularity of the classification and the lack of a provided rationale. This can often leave what is truly causing the classification ambiguous to humans. To combat this, some methods operate at a token or span level (Ádám Kovács and Recski, 2025; Rawte et al., 2025; Belyi et al., 2025). These provide a finer granularity of classification, often helping pinpoint the precise subclaims which are not faithful. However, since these methods only operate on spans of the generated text, when multi-hop reasoning is involved, they may still fail to reveal the intermediate inferences or supporting facts that connect the evidence to the claim, as such implicit inferences do not appear in the generated text.

QA-Based Metrics. Another prominent class of faithfulness evaluation methods frames factuality checking as a question-answering (QA) problem (Scialom et al., 2021; Fabbri et al., 2022; Wang et al., 2020). In these approaches, a set of questions is typically generated from the summary. A QA model is then used to answer these questions given the source documents, and the answers are compared to the corresponding content in the summary (Huang et al., 2023). High overlap or semantic similarity indicates factual consistency, whereas discrepancies signal potential hallucinations. While QA-based metrics offer the advantage of explicitly tying verification to discrete factual questions,

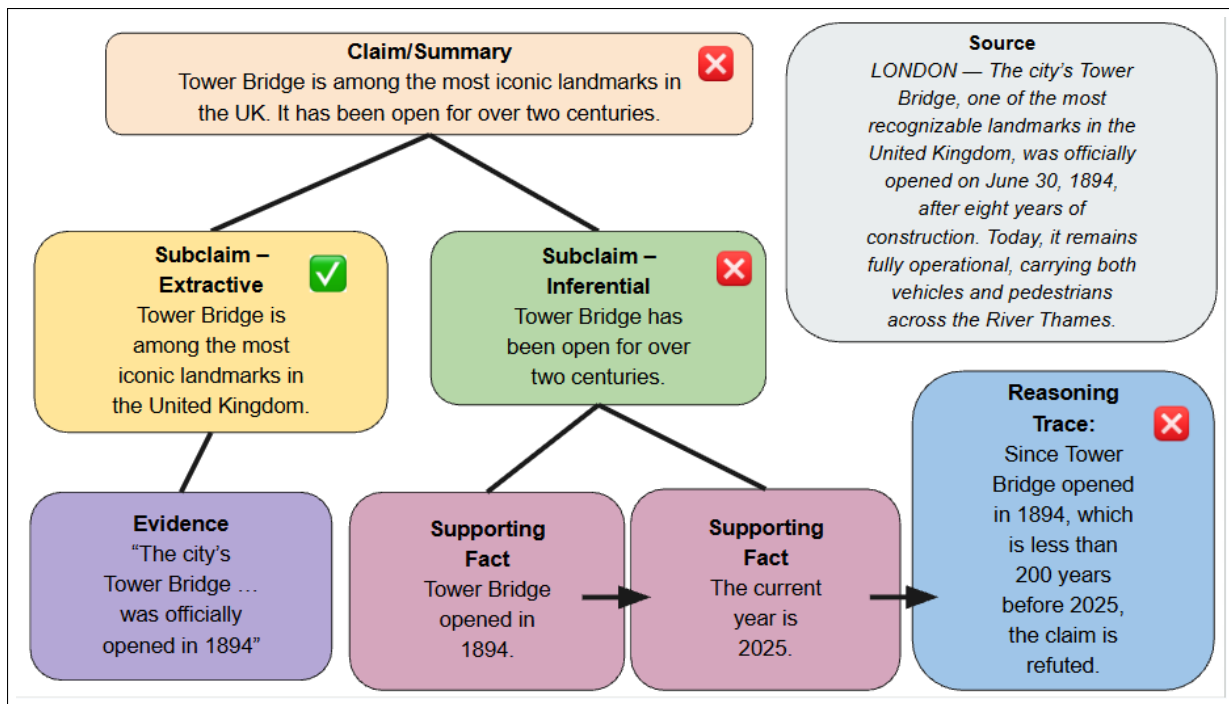


Figure 1: This diagram shows a visualization of our method’s claim tree structure as well as a demonstration of how our chain-of-thought reasoning traces enhance explainability.

multi-hop reasoning may still be unexplored or left implicit as the reasoning required to support or refute a claim is often only able to emerge indirectly through evidence from the source rather than being explicitly modeled from the surface-level claims. Additionally, effectively verifying how inferences over evidence connect to multi-hop claims may require reasoning over both the source and the summary, rather than relying on isolated question–answer pairs.

LLM-based Metrics. A growing body of work leverages LLMs for factuality evaluation by decomposing outputs into smaller, verifiable units (Huang et al., 2025). A common approach, often termed *Decompose-then-Verify*, prompts an LLM to split a generated summary or claim into atomic statements, assess each against source evidence, and then aggregate the results into an overall factuality judgment (Hu et al., 2025; Lu et al., 2025; Zheng and Lee, 2025; Akbar et al., 2024). This approach improves interpretability over coarse-grained entailment scoring by providing statement-level judgements. While the faithfulness judge is typically an LLM, FENICE (Scirè et al., 2024) instead applies an NLI model to each subclaim.

Our method builds on this paradigm but introduces several extensions. First, we propose a dual-typed classification of subclaims into extractive

and inferential, with separate verification pathways. Like FENICE, extractive claims are handled with lightweight NLI-based verifiers. In contrast, inferential claims trigger a reasoning process that collects supporting facts—which may include source evidence or logical and mathematical inferences—and evaluates whether they logically support the claim through a natural program, a structured reasoning sequence executed and validated step by step, marking its first use in factuality verification for summarization.

3 Methodology

Our claim verification framework operates through a multi-stage process that decomposes generated text into subclaims, verifies each subclaim according to its relation to the evidence, and organizes the verification results in an interpretable tree structure. Given a summary, the framework constructs a tree where the root node represents the full summary, intermediate nodes represent subclaims, and leaf nodes contain evidence snippets or supporting facts.

3.1 Claim Decomposition

We begin by prompting the LLM (we use GPT-4o for our purposes) to break the summary into decontextualized subclaims. Rather than decom-

posing into atomic claims, we decompose the summary into subclaims mirroring the structure of a sentence or complete clauses as closely as possible, to preserve semantic fidelity and avoid over-decomposition. For decontextualization, pronouns and ambiguous references are replaced by the LLM with explicit entity mentions, making each subclaim interpretable in isolation.

3.2 Subclaim Classification

The LLM then classifies each subclaim into one of three categories, determining the verification pathway that follows.

Extractive: Subclaims that can be directly supported or refuted by textual evidence from the source without requiring additional reasoning steps.

Inferential: Subclaims that are not directly supported or refuted by evidence in the documents and may require multi-hop reasoning or logical inference to judge their veracity.

Unverifiable: Subclaims that represent opinions, judgements, or unrelated elementary knowledge not about the subject(s) of the source. These are excluded from further processing.

3.3 Verification Process

Depending on the classification of the subclaim, the verification process differs.

Subclaims classified as extractive are evaluated on LettuceDetect (Ádám Kovács and Recski, 2025) and MiniCheck-FT5 (Tang et al., 2024a) for verification, lightweight NLI models for groundedness classification. We aggregate the results of these models by deeming the claim unsupported if both models find it unsupported; otherwise, we deem the claim supported.

For subclaims classified as inferential, the system initiates a reasoning pathway. The LLM first proposes a set of supporting facts—drawn from the source text, logical or mathematical reasoning, or elementary knowledge not requiring verification—and orders them so they can form a coherent reasoning chain. Finally, we prompt an LLM to judge the groundedness of the claim given the supporting facts with chain-of-thought reasoning. These facts are added as children of the subclaim node in the verification tree, and additionally, to increase explainability, we attach the chain-of-thought reasoning from the LLM’s verification to provide the rationale and logical connection between the supporting facts and the claim.

After initial judgment, for each inferential subclaim, the LLM constructs a natural program based off its supporting facts using few-shot prompting adapted from Ling et al. (2023) in order to verify the judgment rigorously. A natural program is a natural language-based chain of reasoning that explicitly lays out premises and demonstrates how they compose to support the claim. Crucially though, this does not simply verify the judgment rigorously but provides an interpretable demonstration of the underlying reasoning to humans, allowing even the rigorous reasoning between the premises and the claim to be audited. This reasoning for the natural program is executed and validated step-by-step by the LLM, with mistakes detected within the original classification being dynamically corrected.

3.4 Evidence Retrieval for Extractive Subclaims

To collect the relevant evidence snippets for extractive subclaims, we first segment the source into manageable chunks. The LLM is then prompted iteratively, selecting the most relevant chunk with respect to verifying the subclaim and decides whether additional evidence is needed. This process continues until the model judges the gathered evidence sufficient for verification. Finally, we add all collected snippets as children of the respective subclaim.

3.5 Final Verification

After the independent verification of subclaims, the system considers that the original claim is supported only if all verifiable subclaims (extractive and inferential) are individually supported.

4 Experiments and Results

4.1 Datasets

To evaluate our method, we evaluate the balanced accuracy of our model on binary factuality verification tasks from several established datasets from the LLM-AggreFact benchmark (Tang et al., 2023) that are established in faithfulness verification for abstractive summarization.

AggreFact-CNN includes generated summaries of CNN/DailyMail articles from the CNN/DM dataset (Hermann et al., 2015). The dataset consists of source news articles from the CNN/DailyMail corpus, generated summaries produced by various summarization models, and binary hallucination

Model	AggreFact		TofuEval		WiCE	AVG
	CNN	XSUM	MediaS	MeetingB		
GPT-4o-2024-08-06	67.5	73.9	66.0	81.1	74.3	72.6
AlignScore	73.2	72.4	67.1	76.5	69.6	71.8
LettuceDetect-large-v1	58.3	67.7	65.8	69.6	79.2	68.1
MiniCheck-FT5	<u>69.9</u>	74.3	<u>73.6</u>	77.6	72.4	73.5
Bespoke-MiniCheck-7B	65.5	77.8	76.0	78.3	<u>83.0</u>	76.1
HalluTree	68.5	<u>74.5</u>	66.4	<u>79.8</u>	83.7	<u>74.6</u>

Table 1: Balanced accuracy (%) on datasets from LLM-AggreFact. Highest score is bolded. Second highest is underlined.

Model	AggreFact		TofuEval		WiCE	AVG
	CNN	XSUM	MediaS	MeetingB		
HalluTree (Dual-Pathed)	68.5	74.5	66.4	79.8	83.7	74.6
Treat All Extractive	58.4	71.0	57.8	68.6	80.1	67.2
Treat All Inferential	<u>63.0</u>	<u>71.9</u>	<u>65.8</u>	<u>79.2</u>	<u>83.0</u>	<u>72.6</u>

Table 2: Results of ablation study on dual-paths for subclaims.

labels indicating whether summaries contain factual inconsistencies with respect to their source articles. We evaluate on the 558 examples in LLM-AggreFact.

AggreFact-XSum contains generated summaries of BBC articles from the XSum corpus (Narayan et al., 2018). Like AggreFact-CNN, it provides binary hallucination labels for summaries generated by various models. We evaluate on the 558 examples in LLM-AggreFact.

WiCE is a fine-grained textual entailment dataset built on claim and evidence pairs extracted from Wikipedia (Kamoi et al., 2023). The data set uses real-world examples extracted from Wikipedia sentences, evidence articles to which the claims refer, fine-grained entailment judgments over sub-sentence units, and minimal subsets of evidence sentences supporting each sub-claim. WiCE includes challenging verification and retrieval problems involving multi-sentence reasoning. We evaluate on the 358 examples from this dataset in LLM-AggreFact.

TofuEval contains two factuality evaluation tasks – **MediaS** and **MeetingB** – drawn from the TofuEval benchmark (Tang et al., 2024b), which was designed to assess LLM factual consistency across multiple domains. MediaS consists of summaries of news and media sources with binary factuality annotations, while MeetingB consists of gener-

ated summaries of meeting transcripts, annotated for consistency with the meeting records. These datasets broaden evaluation coverage to conversational and multi-speaker domains, providing a more diverse testbed for factual verification methods.

4.2 Baselines

We compare our method against strong baselines spanning both NLI-based and LLM-based verification approaches. On the NLI side, we include **LettuceDetect** (Ádám Kovács and Recski, 2025), **MiniCheck-FT5** (Tang et al., 2024a), and **AlignScore** (Zha et al., 2023), which use lightweight natural language inference models to detect hallucinations and assess faithfulness. Among LLM-based systems, we consider chain-of-thought prompting **GPT-4o**, a state-of-the-art model frequently used for faithfulness assessment, as well as **Bespoke-MiniCheck-7B** (Tang et al., 2024a), a state-of-the-art finetuned model which outperforms frontier models on LLM-AggreFact.

4.3 Main Results

Table 1 presents balanced accuracy across the datasets selected from LLM-AggreFact. HalluTree achieves the second-highest average accuracy (74.6%), outperforming all baselines we tested except for Bespoke-MiniCheck. While Bespoke-MiniCheck attains a slightly higher average accuracy, HalluTree offers a key advantage in trans-

Snippet from Natural Program Output for Inferential Subclaim (WiCE)

Premises

- #1. ROOT (R4): “Currently, SriLankan operates an all-Airbus fleet with the exception of its discontinued Air-Taxi services.”
- #2. S4 (INFERENTIAL): “SriLankan Airlines currently operates an all-Airbus fleet with the exception of SriLankan Airlines' discontinued Air-Taxi services.”
- #3. E4A: The context mentions a fiasco involving the launch of an air taxi service which was eventually abandoned causing millions of dollars in losses to SriLankan Airlines.
- #4. E4B: The context does not provide any specific information about the current fleet composition of SriLankan Airlines being exclusively Airbus.

Reasoning

- #5. (by #3) The air taxi service of SriLankan Airlines was discontinued.
- #6. (by #4) There is no evidence in the context to confirm that SriLankan Airlines operates an all-Airbus fleet currently.
- #7. (by #5, #6) While the air taxi service is confirmed to be discontinued, the claim that the current fleet is all-Airbus is not supported by the provided evidence; S4 is not supported.

Subclaim Status

- S4: Not Supported – The context confirms the discontinuation of the air taxi service but does not confirm that the current fleet is exclusively Airbus.

Figure 2: An example natural program generated during verification of an inferential claim from the WiCE dataset.

parency. The finer-grained decomposition into subclaims, coupled with hierarchical verification trees, makes the reasoning process auditable and interpretable.

Importantly, HalluTree outperforms all NLI-based methods such as AlignScore, MiniCheck-FT5, and LettuceDetect, demonstrating that structured decomposition paired with specialized verification pathways can yield stronger performance than flat entailment classification. This shows that HalluTree narrows the gap with Bespoke-MiniCheck while introducing interpretable reasoning, enabling both competitive accuracy and improved transparency.

4.4 Explainability in Practice

Unlike black-box verifiers that surface only a final label, HalluTree exposes the full reasoning trail for each decision. Consider the Natural Program excerpt for the WiCE subclaim about SriLankan Airlines’ fleet: the system (i) lists concrete premises (#1–#4), separating source evidence from assumptions; (ii) derives intermediate conclusions (#5–#6)

with provenance (e.g., “by #3”); and (iii) composes these steps into a final inference (#7) that justifies the verdict of “Not Supported” because the context confirms the air-taxi discontinuation but lacks evidence that the current fleet is all-Airbus. This structured trace makes the decision auditable: a reviewer can pinpoint exactly which premise would need revision to flip the outcome.

Empirically, HalluTree matches or exceeds most state-of-the-art baselines while providing superior transparency. For extractive claims, span-level evidence highlights show where the text is (or isn’t) supported; for inferential claims, natural programs show why—linking premises to conclusions via explicit, checkable steps. The result is a verifier that not only performs competitively but also turns factuality judgments into explanations that users can inspect, contest, and improve.

4.5 Ablations

To better understand the effect of our dual-path routing, we conduct two ablations. First, we evaluate variants that route all subclaims through the extrac-

tive pathway (*All-Extractive*) or through the inferential pathway (*All-Inferential*). Second, we analyze the distribution of claim types across datasets along with our method’s performance on those datasets considering the proportion of extractive and inferential claims.

Routing Variants. Table 2 shows balanced accuracy for the ablated models. Both constrained settings degrade performance: *All-Extractive* struggles on inference-heavy datasets, while *All-Inferential* incurs extra reasoning cost while having worse performance on extractive-heavy datasets. These results demonstrate that routing based on subclaim type enables our method to selectively apply natural program reasoning where it matters, improving accuracy on complex inferential claims, while avoiding unnecessary overhead on extractive ones.

Claim-Type Distribution. We also measure the proportion of extractive vs. inferential subclaims in each dataset (Table 3). Comparing our results from Table 1, our method performed strongly compared to other baselines on datasets with heavy inferential subclaim ratios, such as WiCE and AggreFact-XSUM, while generally maintaining more average performance on extractive-heavy datasets. This distribution provides an explanation for where type-aware routing yields the largest gains.

Dataset	Extractive	Inferential
AggreFact-CNN	96.9	2.1
AggreFact-XSUM	50.4	49.6
TofuEval-MediaS	85.6	14.4
TofuEval-MeetingB	81.2	18.8
WiCE	42.0	58.0

Table 3: Subclaim type distribution (% of verifiable subclaims).

Summary. These ablations highlight that type-aware decomposition and routing are not only interpretable but also empirically necessary: forcing all claims into a single pathway reduces accuracy, while claim-type distributions explain why balanced routing achieves consistent gains.

5 Conclusion

We present a hierarchical claim verification framework that advances the state-of-the-art in hallucination detection by providing both accurate classifica-

tion and human-interpretable explanations. Unlike black-box approaches that output only binary classifications, our framework makes the verification process transparent through a tree-based structure that traces the pipeline from claim decomposition to evidence gathering and reasoning.

Our method provides several key advantages. It offers fine-grained explainability by attaching concrete evidence or inferences to each subclaim and by generating Natural Programs—explicit, natural language reasoning chains that demonstrate how inferential claims are logically supported. This hybrid verification design combines lightweight transformer-based models for extractive verification with LLM-based reasoning for more complex inferential claims, organized in a natural tree structure for clarity.

Experimental evaluation across four diverse datasets—AggreFact-CNN, AggreFact-XSUM, TofuEval, and WiCE—demonstrates the effectiveness of our approach. Our method achieves competitive performance with an average balanced accuracy of 74.6%, exceeding GPT-4o while providing detailed reasoning traces that enhance transparency. This combination of strong performance with explicit reasoning via Natural Programs represents a significant advance over existing black-box approaches.

Such transparency is crucial for practical deployment, where understanding not just whether a claim is supported but also why it is supported or refuted can help identify weaknesses, build trust, and improve reliability in real-world applications.

Limitations

Performance Tradeoffs. HalluTree improves on inference-heavy datasets but underperforms on extractive-heavy ones, where some simpler entailment-based baselines remain stronger. This reflects that our specialized reasoning pathway benefits complex claims, but introduces unnecessary overhead and noise when most claims can be directly verified against the source.

Granularity of Judgments. Our framework outputs binary faithful/unfaithful decisions at the claim level. While subclaims are verified individually, the final aggregation does not capture intermediate degrees of support or uncertainty, which could limit usefulness in downstream applications that require nuanced reliability scores.

Computational Overhead. Compared to single-pass verification methods, HalluTree incurs significantly higher cost. Each stage—decomposition, classification, evidence selection, and verification—requires separate LLM calls. This overhead grows with claim length and makes the method less practical for large-scale or latency-sensitive deployments.

Reliance on LLM Quality. Errors in early stages, such as decomposition or classification, propagate through the pipeline and can compromise verification accuracy. In particular, misclassification between extractive and inferential claims can route subclaims through an inappropriate verification pathway, lowering performance.

Evaluation Scope. Our experiments are limited to benchmark datasets that primarily focus on factual consistency in summarization. Broader domains—such as multimodal sources, conversational data, or more diverse factuality errors—may expose different challenges not addressed by our current framework.

References

- Shayan Ali Akbar, Md Mosharaf Hossain, Tess Wood, Si-Chi Chin, Erica M Salinas, Victor Alvarez, and Erwin Cornejo. 2024. [HalluMeasure: Fine-grained hallucination measurement using chain-of-thought reasoning](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15020–15037, Miami, Florida, USA. Association for Computational Linguistics.
- Catarina G Belém, Pouya Pezeshkpour, Hayate Iso, Seiji Maekawa, Nikita Bhutani, and Estevam Hruschka. 2025. [From single to multi: How LLMs hallucinate in multi-document summarization](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5276–5309, Albuquerque, New Mexico. Association for Computational Linguistics.
- Masha Belyi, Robert Friel, Shuai Shao, and Atindriyo Sanyal. 2025. [Luna: A lightweight evaluation model to catch language model hallucinations with high accuracy and low cost](#). In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 398–409, Abu Dhabi, UAE. Association for Computational Linguistics.
- Preetam Prabhu Srikar Dammu, Himanshu Naidu, Mouly Dewan, YoungMin Kim, Tanya Roosta, Aman Chadha, and Chirag Shah. 2024. [ClaimVer: Explainable claim-level verification and evidence attribution of text through knowledge graphs](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13613–13627, Miami, Florida, USA. Association for Computational Linguistics.
- Alphaeus Dmonte, Roland Oruche, Marcos Zampieri, Prasad Calyam, and Isabelle Augenstein. 2025. [Claim verification in the age of large language models: A survey](#). *Preprint*, arXiv:2408.14317.
- Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. [QAFactEval: Improved QA-based factual consistency evaluation for summarization](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.
- Tanya Goyal and Greg Durrett. 2020. [Evaluating factuality in generation with dependency-level entailment](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3592–3603, Online. Association for Computational Linguistics.
- Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. [Teaching machines to read and comprehend](#). In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.
- Qisheng Hu, Quanyu Long, and Wenya Wang. 2025. [Decomposition dilemmas: Does claim decomposition boost or burden fact-checking performance?](#) In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6313–6336, Albuquerque, New Mexico. Association for Computational Linguistics.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. [A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions](#). *ACM Trans. Inf. Syst.*, 43(2).
- Yichong Huang, Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2023. [The factual inconsistency problem in abstractive text summarization: A survey](#). *Preprint*, arXiv:2104.14839.
- Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. [WiCE: Real-world entailment for claims in Wikipedia](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7561–7583, Singapore. Association for Computational Linguistics.
- Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. [Evaluating the factual consistency of abstractive text summarization](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

- Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. [SummaC: Re-visiting NLI-based models for inconsistency detection in summarization](#). *Transactions of the Association for Computational Linguistics*, 10:163–177.
- Zhan Ling, Yunhao Fang, Xuanlin Li, Zhiao Huang, Mingu Lee, Roland Memisevic, and Hao Su. 2023. [Deductive verification of chain-of-thought reasoning](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 36407–36433. Curran Associates, Inc.
- Yining Lu, Noah Ziemis, Hy Dang, and Meng Jiang. 2025. [Optimizing decomposition for optimal claim verification](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5095–5114, Vienna, Austria. Association for Computational Linguistics.
- Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. [Don’t give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.
- Vipula Rawte, S.m Towhidul Islam Tonmoy, Shrivani Nag, Aman Chadha, Amit Sheth, and Amitava Das. 2025. [FACTOID: FActual enTailment fOr hallucination detection](#). In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, pages 599–617, Albuquerque, New Mexico. Association for Computational Linguistics.
- 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.
- Alessandro Scirè, Karim Ghonim, and Roberto Navigli. 2024. [FENICE: Factuality evaluation of summarization based on natural language inference and claim extraction](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14148–14161, Bangkok, Thailand. Association for Computational Linguistics.
- Wooseok Seo, Seungju Han, Jaehun Jung, Benjamin Newman, Seungwon Lim, Seungbeen Lee, Ximing Lu, Yejin Choi, and Youngjae Yu. 2025. [Verifying the verifiers: Unveiling pitfalls and potentials in fact verifiers](#). *Preprint*, arXiv:2506.13342.
- Liyan Tang, Tanya Goyal, Alex Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryscinski, Justin Rousseau, and Greg Durrett. 2023. [Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11626–11644, Toronto, Canada. Association for Computational Linguistics.
- Liyan Tang, Philippe Laban, and Greg Durrett. 2024a. [MiniCheck: Efficient fact-checking of LLMs on grounding documents](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8818–8847, Miami, Florida, USA. Association for Computational Linguistics.
- Liyan Tang, Igor Shalyminov, Amy Wong, Jon Burnsky, Jake Vincent, Yu’an Yang, Siffi Singh, Song Feng, Hwanjun Song, Hang Su, Lijia Sun, Yi Zhang, Saab Mansour, and Kathleen McKeown. 2024b. [TofuEval: Evaluating hallucinations of LLMs on topic-focused dialogue summarization](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4455–4480, Mexico City, Mexico. Association for Computational Linguistics.
- Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. [Asking and answering questions to evaluate the factual consistency of summaries](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.
- Haoran Wang and Kai Shu. 2023. [Explainable claim verification via knowledge-grounded reasoning with large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6288–6304, Singapore. Association for Computational Linguistics.
- Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. [AlignScore: Evaluating factual consistency with a unified alignment function](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.
- Zhi Zheng and Wee Sun Lee. 2025. [Reasoning-cv: Fine-tuning powerful reasoning llms for knowledge-assisted claim verification](#). *Preprint*, arXiv:2505.12348.
- Ádám Kovács and Gábor Recski. 2025. [Lettucedetect: A hallucination detection framework for rag applications](#). *Preprint*, arXiv:2502.17125.

A Appendix

A.1 Prompts

Decomposition and Decontextualization Prompt

You are given a summary text. Your task is to decompose it into subclaims that mirror the sentence-like structure of the original as closely as possible. Each subclaim should

be decontextualized, meaning it must stand on its own and be understandable without reference to the surrounding text.

Guidelines:

Preserve sentence alignment: Each subclaim should correspond to one sentence in the original summary wherever possible, or the closest equivalent if sentences are not present.

Minimal splitting: Do not overdecompose by introducing claims not directly stated in the text.

No merging or compression: Each subclaim should stay as close as possible to its original sentence(s).

Decontextualize: Rewrite each subclaim so that it is fully interpretable in isolation, avoiding pronouns or vague references.

Output format: Output sentences in a numbered list (1. 2. 3. etc) with each sentence on its own line.

Source document for context: { }

Summary: { }

Subclaims:

Subclaim Classification Prompt

You are an expert at classifying sentences based on their relationship to the provided context and the subjects (the main entities or events the context is about) of that context.

Classifications:

UNVERIFIABLE: Contains opinions or judgments, and background/common-knowledge or definite-truth statements not about the subject(s) of the source. Includes math or logic truths, calendar arithmetic, unit conversions, definitional or taxonomic facts, and geographic containment that do not need verification against the context. These are often bridge facts used to connect evidence.

EXTRACTIVE: Contains claims that are directly supported or directly refuted by explicit spans in the context without reasoning.

INFERENTIAL: Contains claims about the subject(s) of the source that are not directly supported or refuted by the context and require multi-hop reasoning over the provided evidence. They may rely on UNVERIFIABLE background facts as bridges, but the claim itself is about the subject(s).

Rubric:

- 1) Identify the subject(s) of the source.
- 2) If the ENTIRE claim is a background or definite-truth proposition not about the subject(s) of the source, classify as UNVERIFIABLE.
- 3) Else, if explicit context spans support or refute the claim, classify as EXTRACTIVE.
- 4) Else, classify as INFERENTIAL.

Tie-breakers:

- Prefer UNVERIFIABLE for math, logic, calendar arithmetic, unit conversions, definitional or lexical truths, and geography containment that are not about the subject(s).
- Do not mark as UNVERIFIABLE if the statement asserts a property or relation of the subject(s), even if widely known; that is INFERENTIAL unless directly supported by the context.

- If deciding requires external, subject-specific facts not in the context, classify as INFERENTIAL.

1. First reason toward your decision. Do not decide until after you have reasoned.
2. After reasoning, output exactly one label from UNVERIFIABLE, EXTRACTIVE, INFERENTIAL on a new line and nothing else.

Context: { }

Claim: { }

Let's think step by step:

Evidence Collection Prompt

You are an expert at extracting evidence from context to support or refute a subclaim.

Critical Rules:

1. If possible, extract the span of evidence that is most directly relevant to the sub-

claim.

2. Don't repeat evidence that has already been collected.
3. If there is truly no additional relevant evidence in the context, output the token <NO_MORE_EVIDENCE>

Context: {}

Subclaim: {}

Already collected evidence (do not repeat):

Next evidence:

Supporting Fact Proposal Prompt

You are an expert at constructing logical bridges between evidence and an inferential subclaim to either support or refute the subclaim.

Terminology:

- EVIDENCE fact: directly supported by explicit spans in the context.
- BACKGROUND fact: elementary common knowledge or definite truth (math, logic, calendar arithmetic, definitions, geography containment) that is not about the subject(s) of the source and does not require verification against the context. Use only if needed to connect evidence to the subclaim.

Critical Rules:

1. Include EVIDENCE facts only if they are explicitly supported by the context. Closely paraphrase or directly copy the supporting span.
2. You may include BACKGROUND facts that are not about the subject(s) and are necessary to form the reasoning chain. Do not introduce subject-specific facts that are absent from the context.
3. Order the facts so they form a minimal, coherent chain that best supports or refutes the subclaim.
4. Do not add new, subject-specific information. If the context provides nothing usable, output the token <NO_SUPPORTING_FACTS>.
5. Reason first, then output the FACTS.

Example:

Context: "Aspirin was first synthesized in 1897 by chemist Felix Hoffmann at Bayer."
End of Context

Inferential subclaim: "Aspirin was synthesized over a century ago"

Let's think step by step: From the context we know the synthesis year is 1897. Using current-year arithmetic, 1897 is more than 100 years before 2025, so the subclaim is supported.

FACTS:

1. Aspirin was first synthesized in 1897
2. The current year is 2025.
3. 1897 is more than 100 years before 2025.

Context: {}

Inferential subclaim: {}

Let's think step by step:

Inferential Subclaim Verification Prompt

You are an expert at judging whether a set of proposed supporting facts logically supports an inferential subclaim.

Critical Rules:

1. Use only the facts provided; do not rely on any external knowledge or assumptions except for cases of common knowledge or facts that need not be verified.
2. The supporting facts should be able to form a coherent reasoning chain that directly supports the subclaim.
3. Output sections in this order: Reasoning, then final judgment ("YES" or "NO"). YES for supported, NO for refuted.
4. Don't be pedantic in your judgments, direct contradictions or completely unfounded statements are mainly what we seek to prevent. Refuted claims should be clearly, strongly refutable.

Example:

Context:

"Aspirin was first synthesized in 1897 by

chemist Felix Hoffmann at Bayer."
End of Context

Supporting facts:

- 1) [EVIDENCE] "Aspirin was first synthesized in 1897 ..."
- 2) [BACKGROUND] The current year is 2025.
- 3) [BACKGROUND] 1897 is more than 100 years before 2025.

Inferential subclaim: "Aspirin was synthesized over a century ago"

Let's think step by step: The facts provide the synthesis year, the current year, and the difference being more than 100 years. This supports the the subclaim.

Is the claim supported: YES

Context: {}

Supporting facts: {}

Inferential subclaim: {}

Let's think step-by-step:

GPT-4o Baseline Faithfulness Verification Prompt

Your task is to check if the Summary is accurate to the Evidence.

Generate 'Supported' if the Summary is supported when verified according to the Evidence, or 'Unsupported' if the Summary is inaccurate (contradicts the evidence) or cannot be verified.

Evidence: {}

End of Evidence

Summary: {}

End of Summary

Classification ('Supported' or 'Unsupported'):

Multi²: Multi-Agent Test-Time Scalable Framework for Multi-Document Processing

Juntai Cao^{1*}, Xiang Zhang^{1*}, Raymond Li¹, Jiaqi Wei², Chuyuan Li¹, Shafiq Joty³, Giuseppe Carenini¹

¹ University of British Columbia

² Zhejiang University

³ Salesforce Research

{jtcao7, raymond1, chuyuan.li, carenini}@cs.ubc.ca

xzhang23@ualberta.ca, cyou@cs.stonybrook.edu, sjoty@salesforce.com

Abstract

Recent advances in test-time scaling have shown promising results in improving large language model performance through strategic computation allocation during inference. While this approach has demonstrated strong improvements in reasoning tasks, its application to natural language generation tasks, particularly summarization, remains unexplored. Among all of the generation tasks, multi-document summarization (MDS) presents unique challenges by requiring models to extract and synthesize essential information across multiple lengthy documents. Unlike reasoning tasks, MDS demands a more complicated approach to prompt design and ensemble methods, as no single “best-overall” prompt can satisfy diverse summarization requirements. The inherent diversity in summarization needs necessitates exploring how different prompting strategies can be systematically combined to improve performance. We propose a novel framework that harnesses prompt diversity to enhance MDS performance. Our approach generates multiple candidate summaries using carefully designed prompt variations, then ensemble them through sophisticated aggregation methods to produce refined summaries. This prompt diversity enables models to capture different aspects and perspectives of the source documents, leading to more comprehensive and higher-quality summaries. To evaluate our method effectively, we also introduce two new LLM-based metrics: the Preference Alignment Score (PAS) and LLM Atom-Content-Unit score (LLM-ACU), which assess summary quality while addressing the positional bias inherent in automatic evaluations performed by LLMs. Our experiments demonstrate that leveraging prompt diversity significantly enhances summary quality, while also revealing the practical scaling boundaries for MDS tasks.

*Equal contribution.

Distribution of Preference Alignment Score
c = 0.5

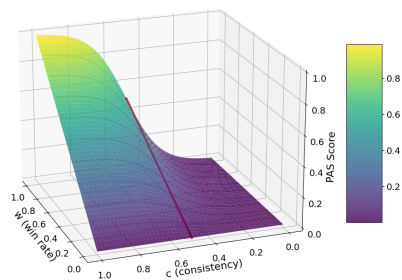


Figure 1: Visualization of the distribution Preference Alignment Score. Applying LLMs’ strong language understanding ability, PAS assign higher score to the content which *consistently* gets preferred by the LLM.

1 Introduction

Test-time scaling (or inference-time scaling) has emerged as a promising approach for enhancing LLM’s performance beyond traditional architectural or data improvements (OpenAI, 2024). While earlier work focused on relationships between models’ capabilities, size, and training resources, recent research demonstrates that strategic compute allocation during inference can yield substantial performance gains. For instance, studies show that increased inference computation produces better results than equivalent investments in pretraining (Snell et al., 2024; Agarwal et al., 2024; Muenighoff et al., 2025).

Research on test-time scaling has largely centered on logical and math reasoning tasks, leaving traditional natural language generation (NLG) tasks relatively unexplored. This gap is particularly notable in summarization, a domain where LLMs have already demonstrated significant advances, generating summaries competitive with human performance (Xiao et al., 2024; Zhang et al.,

2024c; Pu et al., 2023). Beyond text generation, LLMs have also been proven effective as judges when guided by well-designed evaluation protocols (Liu et al., 2024b,c). Recent expansions in context window sizes have created new opportunities to study scaling effects on length-constrained tasks like summarization (Liu et al., 2022). However, LLMs still struggle with key challenges including hallucination, incomplete coverage, language inconsistency, and verbosity (Liu et al., 2024b; Belem et al., 2024).

In this paper, we aim to examine LLMs’ summarization capabilities and their scaling properties by focusing on the multi-document summarization (MDS) task. MDS requires synthesizing and linking information across lengthy documents, handling information redundancy, maintaining factual consistency, and generating coherent and concise summaries while preserving key details. In addition, MDS demands effective reasoning to determine relevance and priority among diverse pieces of information. These characteristics make MDS particularly time- and labor-intensive (Van Veen et al., 2024). To tackle these challenges, we propose a multi-agent approach that leverages prompt ensemble to scale summarization at test time. While traditional prompt ensemble methods exist - such as (a) applying different sampling strategies to a single prompt (Li et al., 2023), or (b) varying few-shot examples within prompts (Arora et al., 2022), their direct application to summarization presents notable limitations. The first approach merely explores variations in the output space, while the second heavily relies on example-based learning, which is better suited for reasoning tasks. Furthermore, summarization differs fundamentally from reasoning tasks, where specific prompts like “Let’s think step by step” (Kojima et al., 2022) can effectively guide models through predetermined reasoning patterns (Zhang et al., 2024d). In contrast, no single “optimal” prompt exists for generating summaries that satisfy diverse requirements. Given these distinctions, summarization demands a more sophisticated approach to prompt ensemble techniques.

Therefore, we propose Multi² framework (Fig. 2) to address this challenge. After generating multiple summaries through diverse prompts while maintaining consistent requirements, we employ aggregation to construct a comprehensive final summary that leverages the strengths of each summary candidate. While increased inference-time com-

putation generally improves performance, recent studies have also identified an *inverse scaling* phenomenon, where excessive computation at test-time can paradoxically degrade performance (Gao et al., 2022; Stroebel et al., 2024). We also investigate this phenomenon by systematically varying the number of samples and examining its boundaries.

Another challenge in MDS is the reliability of automatic evaluation metrics. Traditional metrics like ROUGE (Lin, 2004) have proven insufficient for capturing summary quality, while more recent LLM-based metrics such as Auto-ACU (Liu et al., 2023b), LLMCompare (Liu et al., 2024b), and LLMRank (Liu et al., 2024c) show limitations, including constraints in contextual understanding for smaller models and persistent positional biases (Wang et al., 2024c). We specifically highlight **positional bias**, where LLMs tend to favor summaries appearing in a particular position (first or second in a pairwise comparison), leading to inconsistencies in evaluation, particularly during test-time scaling. To improve evaluation consistency, we propose two novel metrics: Preference Alignment Score (PAS) and LLM Atom-Content-Unit (LLM-ACU) score. These metrics aim to leverage LLMs’ contextual understanding while incorporating mechanisms to mitigate positional bias, ensuring more reliable and robust summary assessment.

In summary, (1) We present the first comprehensive investigation of test-time scaling laws in text summarization, extending the analysis beyond traditionally explored reasoning tasks; (2) We introduce a new framework Multi² that enhances summarization performance through prompt ensemble at test time; (3) We enhance two existing evaluation protocols for summarization through strategic modifications and incorporating LLMs, improving quantitative assessment of summary quality and advancing automatic evaluation methodologies for summarization tasks.

2 Prompt Ensemble: A Formal Formulation

Let x denote the input text and $\mathcal{P} = \{p_1, p_2, \dots, p_N\}$ be a collection of prompts designed to elicit different aspects of information from the underlying language model. For each prompt $p_i \in \mathcal{P}$, the model produces an output y_i according to a generation function f :

$$y_i = f(x, p_i).$$

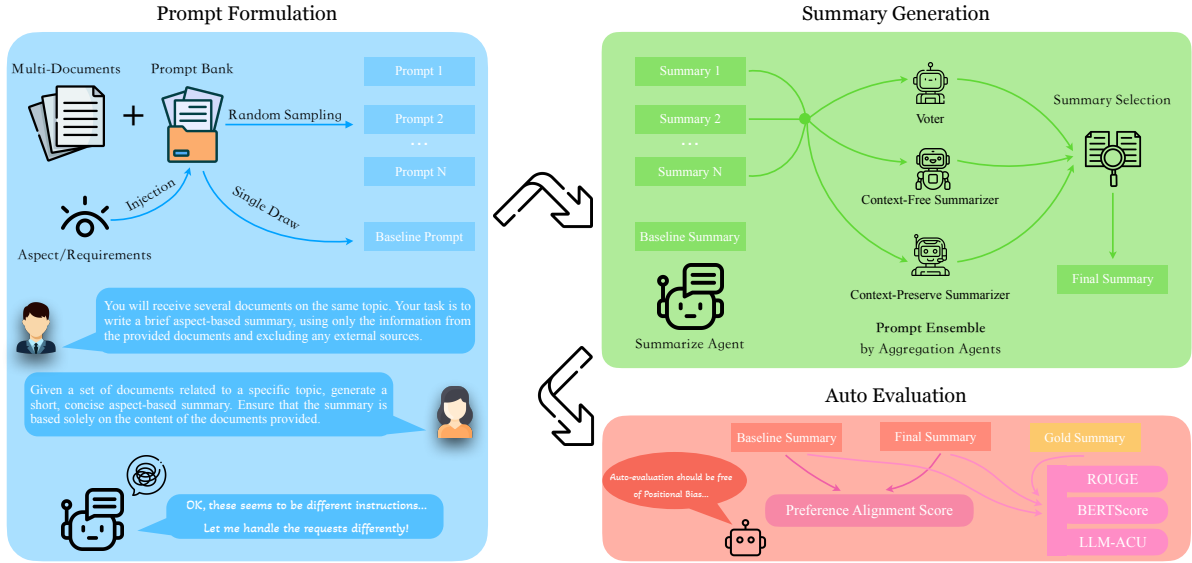


Figure 2: Overview of Multi² summarization inference-time scaling framework. Documents are first summarized by independent LLM agents, each guided by a different prompt from a curated prompt bank and constrained by user requirements. The resulting summaries are then processed by an aggregator (Voter, Context-Preserving Summarizer, or Context-Independent Summarizer) to generate the final consolidated summary.

The intuition behind this methodology is that different prompts p_i induce the model to focus on distinct features or details in the input x , thereby generating complementary outputs.

To combine these outputs, we define an aggregation function $g : \mathcal{Y}^N \rightarrow \mathcal{Y}$ that fuses the individual outputs $\{y_1, y_2, \dots, y_N\}$ into a final output y :

$$y = g(y_1, y_2, \dots, y_N).$$

The aggregation function g can take various forms depending on the specific application, with weighted averaging and majority voting being common implementations. For our MDS task, we implement three distinct formulations of g : content-independent summarization, content-preserving summarization, and voting-based aggregation. The overall system can therefore be formalized as:

$$y = g(f(x, p_1), f(x, p_2), \dots, f(x, p_N)).$$

This formulation ensures that the final generated text y benefits from the diverse perspectives provided by the prompt ensemble. Empirical results indicate that the ensemble method consistently outperforms individual prompt-based generations, as it effectively mitigates the shortcomings of any single prompt by incorporating a broader range of contextual insights from the input x .

2.1 Prompt Space Theory

In this section, we formalize the notion of the *prompt space* and analyze its complexity in the context of Chain-of-Thought (CoT) reasoning. The prompt space, denoted as \mathcal{P} , represents the set of all possible step templates that a language model (LM) may generate or be guided to generate during the reasoning process. Each template $p \in \mathcal{P}$ is a discrete instruction that dictates how information is to be extracted from the latent representation $h \in R^d$ and subsequently discretized into a sequence of tokens $o = (o_1, o_2, \dots, o_k)$. In effect, the prompt space forms the interface between the continuous latent space and the discrete textual output (Zhang et al., 2024d).

The latent vector h is assumed to encode m bits of information relevant to the task at hand. When the model follows a given prompt template p , it extracts up to s bits of information per reasoning step. Thus, each template can be viewed as a function

$$p : h \rightarrow o, \quad o \in \{0, 1\}^s,$$

where the mapping is constrained by the model’s capacity to “read out” a subset of the information encoded in h . The total number of unique ways to extract s bits from m bits is given combinatorially by

$$C(m, s) = \binom{m}{s} = \frac{m!}{s!(m-s)!}.$$

This expression characterizes the *prompt space complexity*, as it represents the number of potential step templates available to the model at each CoT step.

In practice, the prompt space is not uniformly sampled; instead, the LM employs learned heuristics to navigate this enormous space. That is, while the theoretical upper bound $C(m, s)$ may be astronomically high, the effective search space is significantly reduced through task-specific training and, in many cases, human supervision. In an unsupervised setting, the model’s intrinsic biases might lead it to select suboptimal templates, thereby increasing the difficulty of navigating the subsequent *answer space* \mathcal{S} – the space of all possible reasoning paths and final outputs.

More formally, let ϕ denote the underlying computation that updates the hidden state:

$$h_{t+1} = \phi(h_t, p),$$

For brevity, we summarize the CoT process as follows: for $t = 1, \dots, T$,

$$o_t = p_t(h_{t-1}), \quad h_t = \phi(h_{t-1}, p_t).$$

This compact notation encapsulates the iterative extraction of output tokens o_t and the recurrent update of the hidden state h_t via the chosen prompt p_t .

Here, the selection of each $p_t \in \mathcal{P}$ not only determines the immediate output o_t but also has a cascading effect on the evolution of the hidden state h_t and, consequently, the trajectory within the answer space \mathcal{S} .

This intricate relationship between the prompt space and the answer space can be seen as a two-tier search problem: first, the model must identify a suitable template p from the high-dimensional prompt space \mathcal{P} , and then it must effectively navigate the answer space \mathcal{S} defined by the recurrence $h_t \rightarrow h_{t+1}$. Empirical evidence shows that even small deviations in the chosen template p can lead to exponentially larger errors in the final answer, underscoring the sensitivity of the overall reasoning process to prompt selection.

In summary, the prompt space theory emphasizes that the effectiveness of CoT reasoning hinges on the model’s ability to manage the combinatorial complexity inherent in extracting relevant information from its latent space. Supervised methods, which incorporate task-specific guidance, can significantly reduce the search complexity from

the theoretical bound $C(m, s)$ by constraining the model to a subset of high-quality prompts. This not only simplifies the navigation of the answer space but also enhances the overall reliability of the reasoning process.

3 Multi² Framework

3.1 Multi-Agent Summarization

Our Multi² test-time scaling framework for MDS is illustrated in Figure 2. The framework operates in two main stages: candidate generation and summary aggregation. In the first stage, input documents are processed by multiple independent LLM agents using randomly selected prompts from a curated prompt bank, simulating real-world summarization scenarios. The generated candidate summaries, along with the original requirements, are then passed to the aggregator module. The aggregator module implements three distinct approaches: vote, context-preserving summarizer (CPS), and context-free summarizer (CFS).

The **vote** agent evaluates all candidate summaries against the original input documents and provides a detailed explanation before selecting the best summary. We explicitly require the agent to complete its reasoning before indicating its final selection, ensuring the choice is constrained by the documented rationale. Instead of selecting the best candidate summary, CPS and CFS aggregate the candidate summaries into a final summary. The **CPS** agent generates a refined summary by consulting both the original documents and the candidate summaries, aiming for completeness and conciseness. In contrast, the **CFS** agent focuses solely on the candidate summaries without access to the original documents, producing a consolidated summary through reference-free synthesis.

3.2 Automatic Evaluation

3.2.1 Positional Bias and Motivation

Recent approaches to automatic evaluation have increasingly leveraged LLMs, either through comparative (pairwise) assessment or direct scoring mechanisms. However, both approaches face challenges. Comparative methods struggle with positional bias, an inherent limitation of LLM judges. While previous research (Liu et al., 2024c) suggested that advanced models (like gpt-4o) might mitigate this issue, our experiments in Appendix demonstrate that LLM evaluations remain extremely susceptible to position-dependent variations, especially

on contextual tasks like MDS. Direct scoring approaches face different challenges: defining clear scoring guidelines could be difficult, and ensuring consistent application of grading rubrics across different generations remains challenging. Moreover, the complexity of nuanced scoring - a task challenging even for human evaluators who struggle more with five-point Likert scales than binary preferences makes it particularly difficult for LLMs to provide reliable quantitative assessments.

To address these limitations and enable reliable large-scale evaluation of generated summaries, we propose two novel metrics Preference Alignment Score (PAS) and LLM-ACU score. These metrics are specifically designed to mitigate positional bias, while providing repeatable quantitative measurements for systematic comparison of summary quality.

3.2.2 Preference Alignment Score

We develop the Preference Alignment Score (PAS) as an enhancement to the LLMCompare (Liu et al., 2024b) method for quantitatively evaluating preference rates of summaries compared to a baseline. LLMCompare employs an LLM judge to evaluate two summaries against the source documents, determining which is superior (1 or 2) or if they are equivalent (tie). The pairwise comparative setup offers utility to practitioners (e.g., evaluation for A/B testing) while eliciting evaluations better aligned with human judgment from automatic evaluators (Wang et al., 2023a; Liu et al., 2024a). To address the inherent positional bias, we implement the metric with two-phase comparison process. First, we use an LLM as judge to obtain preferences with summaries (target and baseline) in their original positions. Then, we swap the positions of the two summaries and obtain a second set of preferences, relabeling them based on their new positions to eliminate labeling bias. From this two-step comparison, we compute the *win rates* (w_1, w_2) of the target summarization method against the baseline in each step, and the *consistency rate* (C) of predictions across both orderings (Figure 3).

Importantly, when evaluating consistency, if either comparison (i.e., before or after the swapping) results in a tie, we consider it consistent with any outcome in the other comparison to avoid overpenalizing borderline cases. The final PAS score is computed as follows:

$$\text{PAS} = W_{\text{pref}} \frac{1}{1 + \exp^{-k(C-0.5)}}, \quad (1)$$

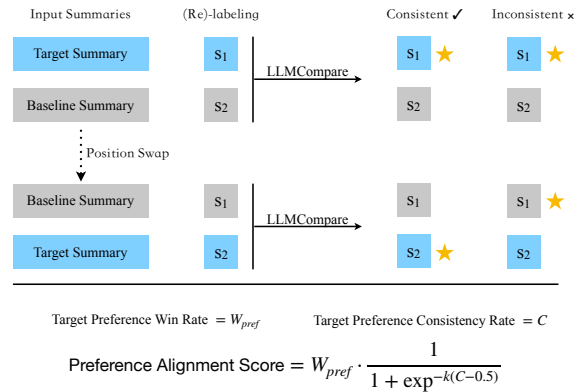


Figure 3: Two-step calculation of PAS based on LLMCompare.

where W_{pref} refers to *preference rate* calculated from *win rates* (w_1 and w_2); C refers to *consistency score*; k controls sensitivity to consistency variations (default to 10 according to our experiments on a validation set). In practice, the preference weight W_{pref} can be determined using either max-pooling or averaging:

$$W_{\text{pref}}^{\text{max}} = \max(w_1, w_2) \quad (2)$$

$$W_{\text{pref}}^{\text{avg}} = \frac{(w_1 + w_2)}{2} \quad (3)$$

Figure 1 illustrates the distribution of PAS score across different preference weights W and consistency values C .

The PAS score integrates both *preference rate* and *consistency* to ensure robust evaluation. A high PAS score requires both factors to be high, indicating consistent preference for the same summary. When model predictions remain stable, the PAS score correlates directly with the *preference rate*. However, inconsistent predictions yield low PAS scores regardless of preference outcomes, as the metric deliberately penalizes unreliable evaluations.

A low (near-zero) PAS score can result from either (a) summaries that consistently underperform the baseline, or (b) unreliable evaluations due to positional bias. Our framework effectively distinguishes between these scenarios. For instance, if evaluations consistently favor Position 1 regardless of content, the win rate might reach 100%, but the consistency score would approach 0, yielding a very low PAS score (0.06) to correctly identify unreliable evaluation. Conversely, with high consistency, the same win rate produces a PAS score near 1, indicating clear, reliable preference.

By design, PAS scores above 0.5 indicate performance better than baseline, while lower scores signal inferior performance or evaluation inconsistency. PAS deliberately employs a conservative approach to ambiguous cases, assigning low scores when no clear winner emerges due to tied quality or inconsistent judgments. This design choice prioritizes robustness and interpretability over sensitivity, treating both “tie with baseline” and “worse than baseline” scenarios similarly, as both indicate failure to establish consistent advantage.

3.2.3 LLM-ACU Score

Inspired by the Atomic Content Unit (ACU) score (Liu et al., 2023a,b), we propose an LLM-based ACU metric to quantitatively measure the completeness of summaries. The process consists of two phases. First, using few-shot prompting, we guide an LLM to extract ACUs from reference summaries. These ACUs are designed to capture essential factual units that are independently interpretable without references. In the evaluation phase, we present the extracted ACUs alongside the model generated summary and ask an LLM to determine which of the ACUs are entailed in the generated summary. The final score f for a set of summaries S and their corresponding ACU sets \mathcal{A} is computed as the average unnormalized ACU score:

$$f(S, \mathcal{A}) = \frac{1}{|S|} \sum_{s \in S} \frac{e_s}{|\mathcal{A}_s|}, \quad (4)$$

where e_s represents the number of ACUs in the system output that are entailed by the gold standard ACUs \mathcal{A}_s determined by the LLM. Recent work has suggested that fine-tuning primarily enables format adaptation rather than information acquisition in language models (Allen-Zhu and Li, 2024). Therefore, we do not finetune models for extracting ACUs and checking entailment, but instead leverage the advanced language understanding capabilities of LLMs directly for both steps. Therefore, we adopt gpt-4o for both ACU extraction and entailment verification.

4 Main Results

Our experimental results are presented in Tables 1 and 2 for preference metric (PAS scores), Tables 3, 4, and 5 for completeness metrics (LLM-ACU, ROUGE and BERTScore), across both MultiNews and OpenASP datasets. We also attach a detailed analyses examining the relationship between summary length and quality in the Appendix.

4.1 Effectiveness of Test-Time Scaling and Metrics Alignment

Our experiments demonstrate significant improvements through test-time scaling across both preference and completeness metrics. On MultiNews, starting from a low preference baseline, all scaling methods show substantial gains in overall quality. For LLM-ACU score specifically, CPS aggregator achieves the strongest performance in information coverage, with gpt-4o-mini showing substantial gains from a baseline of 47.13 to 54.64 with 6 samples. Similarly for OpenASP, despite beginning from a stronger preference baseline, scaling with prompt ensemble still provides notable improvements in overall quality. The LLM-ACU score show comparable trends, with CPS improving gpt-4o-mini’s coverage from 42.35 to 47.82 using 5 samples. These results consistently demonstrate that scaling at test time can effectively enhance both summarization quality and information coverage across different datasets.

Furthermore, Table 5 demonstrates that ROUGE scores consistently improve as the number of ensembled samples increases across both datasets, while maintaining similar degrees of BERTScore demonstrates the robustness of our approach in scaling summarization performance while preserving semantic fidelity. This trend not only reinforces the effectiveness of our test-time scaling approach from the perspective of traditional metrics, but also validates that our new metrics PAS and LLM-ACU score align well with the established evaluation frameworks.

Analysis of the results reveals two key patterns. First, CPS consistently outperforms both CFS and voting approaches across all experimental conditions, suggesting that **access to source documents during ensemble is crucial for maintaining comprehensive coverage and generating more preferred summaries**. Second, completeness improvements are more pronounced on MultiNews compared to OpenASP, indicating that **general-purpose summarization may benefit more from diverse prompt sampling for information capture**.

4.2 Scaling Boundaries and Inverse Scaling

The scaling limitations manifest differently across ensemble methods. In terms of completeness, voting shows minimal improvement across all sample sizes, suggesting that simple selection-based

# Samples	Baseline		gpt-4o				gpt-4o-mini				Vote	
	Max	Avg	CFS		CPS		CFS		CPS		Max	Avg
			Max	Avg	Max	Avg	Max	Avg	Max	Avg		
2			0.69	0.51	0.82	0.62	0.67	0.49	0.80	0.61	0.37	0.23
3			0.73	0.55	0.79	0.62	0.72	0.53	0.72	0.54	0.27	0.16
4	0.25	0.15	0.68	0.50	0.82	0.64	0.73	0.55	0.80	0.60	0.27	0.16
5			0.71	0.52	0.85	0.69	0.81	0.62	0.78	0.60	0.28	0.17
6			0.79	0.60	0.81	0.63	0.77	0.57	0.77	0.60	0.37	0.23

Table 1: PAS scores on Multinews dataset using gpt-4o and gpt-4o-mini models with context-free summarizer (CFS) and context-preserving summarizer (CPS). The aggregator using Vote is model-invariant. We report PAS score with max-pooled (“Max”) and average (“Avg”) preference scores (W_{pref}). Baseline shows both max-pooled and average PAS across all samples. Best scores per column are shown in **bold**.

# Samples	Baseline		gpt-4o				gpt-4o-mini				Vote	
	Max	Avg	CFS		CPS		CFS		CPS		Max	Avg
			Max	Avg	Max	Avg	Max	Avg	Max	Avg		
2			0.63	0.50	0.70	0.55	0.73	0.57	0.79	0.63	0.61	0.45
3			0.72	0.57	0.76	0.59	0.75	0.60	0.83	0.69	0.64	0.48
4	0.51	0.36	0.72	0.55	0.74	0.59	0.77	0.62	0.83	0.71	0.66	0.51
5			0.74	0.59	0.76	0.61	0.82	0.67	0.86	0.72	0.64	0.48
6			0.74	0.60	0.77	0.60	0.81	0.66	0.85	0.72	0.56	0.42

Table 2: PAS scores on OpenASP dataset under the same settings as in Table 1.

LLM-ACU (MultiNews)		gpt-4o		gpt-4o-mini		Vote
# Samples	Baseline	CFS	CPS	CFS	CPS	
2		48.75	51.00	49.14	52.35	47.44
3		49.25	51.11	50.03	52.88	48.31
4	47.13	49.69	51.96	51.02	54.17	48.29
5		50.86	52.70	50.95	53.90	47.65
6		50.35	52.40	51.70	54.64	48.34

Table 3: Comparison of LLM-ACU scores on MultiNews dataset using different ensemble methods. The vote scores are model-invariant and apply to both models. Baseline indicates single sample performance without prompt ensemble. Best score for each model and aggregation agent is shown in **bold**.

ensemble may be insufficient for maintaining comprehensive information coverage. The impact of document context during ensemble emerges as a crucial factor. While CFS performs better than voting, it consistently achieves lower completeness scores than CPS, indicating that losing document context during ensemble creates a ceiling on information preservation.

For preference scores, both datasets exhibit satu-

LLM-ACU (OpenASP)		gpt-4o		gpt-4o-mini		Vote
# Samples	Baseline	CFS	CPS	CFS	CPS	
2		43.05	44.16	44.36	46.07	43.86
3		44.00	45.00	45.04	47.35	44.03
4	42.35	43.64	45.51	45.05	47.55	44.47
5		44.07	46.47	46.13	47.82	44.47
6		44.66	46.30	46.35	47.46	45.00

Table 4: Comparison of LLM-ACU scores on OpenASP dataset under same settings as Table 3.

ration points at approximately 5 samples, beyond which additional scaling yields diminishing returns. This inverse scaling phenomenon is particularly evident in MultiNews, where CPS performance peaks at 5 samples before declining at 6 samples, with the preference score nearly dropping to the same level as CFS. Completeness metrics follow a similar pattern, with gpt-4o’s scores using CPS plateauing around 5 samples, and gpt-4o-mini demonstrating comparable saturation behavior.

These observations suggest that excessive ensemble sizes may introduce noise rather than improvements, and that the choice of ensemble method

Dataset	Model	# Samples	CPS					CFS				
			R1	R2	RL	RLsum	BERTScore	R1	R2	RL	RLsum	BERTScore
MultiNews	Baseline	1	36.29	10.57	18.57	19.23	63.27	36.29	10.57	18.57	19.23	63.27
		2	37.56	10.50	18.44	18.97	63.29	36.33	10.09	18.15	18.38	63.09
	gpt-4o	3	37.74	10.57	18.69	19.20	63.26	36.95	10.23	18.36	18.64	63.19
		4	37.81	10.64	18.67	19.29	63.33	36.98	10.33	18.40	18.74	63.11
		5	38.22	10.88	18.85	19.60	63.35	37.09	10.35	18.43	18.71	63.24
		6	38.17	10.83	18.88	19.56	63.39	37.34	10.45	18.43	18.75	63.27
	gpt-4o-mini	2	39.04	10.78	18.83	20.91	63.22	37.07	10.14	18.25	18.90	63.18
		3	39.28	10.87	18.88	21.15	63.26	37.53	10.16	18.40	19.29	63.12
		4	39.42	10.86	18.87	21.39	63.19	37.81	10.21	18.40	19.61	63.06
		5	39.45	10.89	18.88	21.46	63.14	38.08	10.39	18.49	19.86	62.98
		6	39.67	11.04	18.93	21.56	63.02	38.34	10.49	18.68	20.22	63.10
		Baseline	1	32.47	7.89	15.77	17.11	60.21	32.47	7.89	15.77	17.11
OpenASP	Baseline	2	33.37	7.83	15.94	17.54	60.46	32.19	7.48	15.60	16.49	60.31
		3	33.37	7.87	15.91	17.54	60.50	32.40	7.43	15.63	16.70	60.19
	gpt-4o	4	33.66	7.95	16.04	17.86	60.51	32.42	7.49	15.70	16.89	60.27
		5	33.74	8.06	16.02	17.90	60.54	32.67	7.67	15.70	16.95	60.26
		6	33.98	8.08	16.08	18.00	60.51	32.71	7.61	15.67	17.03	60.25
		2	35.37	8.14	16.20	19.56	60.10	33.19	7.56	15.76	17.79	60.00
	gpt-4o-mini	3	35.77	8.32	16.23	19.84	60.16	33.91	7.65	15.92	18.37	59.94
		4	35.73	8.29	16.26	19.92	60.08	34.30	7.83	15.96	18.66	59.93
		5	35.95	8.30	16.37	20.07	60.12	34.53	7.83	16.01	18.94	59.91
		6	36.04	8.39	16.30	20.15	60.10	34.52	7.79	15.95	18.87	59.84

Table 5: Comparison of ROUGE and BERTScore scores on MultiNews and OpenASP datasets using different models and ensemble sizes. The BERTScore is computed by DEBERTA-XLARGE-MNLI. Best score for each dataset, model and aggregation method is shown in **bold**.

significantly affects both quality and coverage outcomes. **This highlights the importance of identifying optimal scaling thresholds and maintaining document context throughout the ensemble process.**

4.3 Scaling Effect across Model Sizes

Our experiments with gpt-4o and gpt-4o-mini reveal interesting patterns in how model size interacts with scaling benefits. In terms of completeness scores, gpt-4o-mini often achieves larger relative improvements compared to gpt-4o when scaled through prompt ensemble. This suggests that **prompt ensemble can partially compensate for model size limitations in terms of information capture.**

Regarding preference scores, the relationship between model size and performance is more nuanced. While gpt-4o generally outperforms gpt-4o-mini on MultiNews when using CPS, the smaller model achieves competitive results with CFS. More surprisingly, on OpenASP, gpt-4o-mini consistently outperforms its larger version across both CFS and CPS aggregators. This suggests that the **benefits of model scale are not uniform across different summarization tasks, and that scaling smaller models, when combined with appropriate scaling strategies, may sometimes be more effective.** These findings challenge

the assumption that larger models necessarily benefit more from inference-time scaling and emphasize the importance of considering both model size and ensemble size in optimization strategies.

5 Conclusion

In this work, we introduced the Multi² framework to scale MDS through prompt ensemble, showing that we can leverage computational resources at test time to produce more comprehensive and accurate summaries. Our metrics, PAS score and LLM-ACU score also provide more reliable assessments by effectively mitigating positional bias in summary evaluation. Through systematic analysis, we identified specific scaling boundaries in summarization tasks, offering valuable insights into scaling summarization. Our findings suggest two promising research directions: (1) incorporating test-time search algorithms to dynamically guide prompt ensemble optimization, and (2) extending our evaluation metrics to assess model performance in reasoning tasks. These directions highlight the potential of optimizing LLMs’ inference-time behavior across applications where both factual accuracy and logical consistency are crucial.

Limitations

Despite demonstrating that test-time scaling improves summarization quality, our work has several

limitations. First, we restricted our experimental scope to larger general-purpose commercial LLMs rather than including smaller open-source LLMs. This decision was guided by two considerations: (1) our primary objective was to validate the Multi² framework’s general effectiveness rather than comprehensively benchmarking various LLMs’ scaling capabilities, and (2) MDS tasks require robust context understanding typically found in general-purpose, market-proven models rather than smaller models with limited contextual processing ability. Second, we did not conduct human evaluations to compare alignment between our metrics and previous ones. This decision reflects that the baseline metrics we sought to improve have already undergone comprehensive human evaluation and peer review, making additional human studies redundant for our specific research questions.

Acknowledgments

The authors thank the anonymous reviewers for their valuable feedback and suggestions. The authors acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC). Nous remercions le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG) de son soutien.

References

- David H. Ackley, Geoffrey E. Hinton, and Terrence J. Sejnowski. 1985. [A learning algorithm for boltzmann machines](#). *Cognitive Science*, 9(1):147–169.
- Rishabh Agarwal, Avi Singh, Lei M. Zhang, Bernd Bohnet, Luis Rosias, Stephanie Chan, Biao Zhang, Ankesh Anand, Zaheer Abbas, Azade Nova, John D. Co-Reyes, Eric Chu, Feryal Behbahani, Aleksandra Faust, and Hugo Larochelle. 2024. [Many-shot in-context learning](#). *Preprint*, arXiv:2404.11018.
- Zeyuan Allen-Zhu and Yuanzhi Li. 2024. [Physics of language models: Part 3.1, knowledge storage and extraction](#). *Preprint*, arXiv:2309.14316.
- Shmuel Amar, Liat Schiff, Ori Ernst, Asi Shefer, Ori Shapira, and Ido Dagan. 2023. [OpenAsp: A benchmark for multi-document open aspect-based summarization](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1967–1991, Singapore. Association for Computational Linguistics.
- Simran Arora, Avani Narayan, Mayee F. Chen, Laurel Orr, Neel Guha, Kush Bhatia, Ines Chami, Frederic Sala, and Christopher Ré. 2022. [Ask me anything: A simple strategy for prompting language models](#). *Preprint*, arXiv:2210.02441.
- Catarina G. Belem, Pouya Pezeskhpour, Hayate Iso, Seiji Maekawa, Nikita Bhutani, and Estevam Hruschka. 2024. [From single to multi: How llms hallucinate in multi-document summarization](#). *Preprint*, arXiv:2410.13961.
- Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, and Pengfei Liu. 2020. [Metrics also disagree in the low scoring range: Revisiting summarization evaluation metrics](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5702–5711, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. 2024. [Large language monkeys: Scaling inference compute with repeated sampling](#). *Preprint*, arXiv:2407.21787.
- Lingjiao Chen, Jared Quincy Davis, Boris Hanin, Peter Bailis, Ion Stoica, Matei Zaharia, and James Zou. 2024. [Are more LLM calls all you need? towards the scaling properties of compound AI systems](#). In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. [Training verifiers to solve math word problems](#). *arXiv preprint arXiv:2110.14168*.
- Yann Dubois, Percy Liang, and Tatsunori Hashimoto. 2024. [Length-controlled alpacaeval: A simple debiasing of automatic evaluators](#). In *First Conference on Language Modeling*.
- Günes Erkan and Dragomir R. Radev. 2004. [Lexrank: Graph-based lexical centrality as salience in text summarization](#). *ArXiv*, abs/1109.2128.
- Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. [Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.
- Leo Gao, John Schulman, and Jacob Hilton. 2022. [Scaling laws for reward model overoptimization](#). *Preprint*, arXiv:2210.10760.
- Shima Gerani, Yashar Mehdad, Giuseppe Carenini, Raymond T. Ng, and Bitan Nejat. 2014. [Abstractive summarization of product reviews using discourse structure](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1602–1613, Doha, Qatar. Association for Computational Linguistics.
- John Giorgi, Luca Soldaini, Bo Wang, Gary Bader, Kyle Lo, Lucy Lu Wang, and Arman Cohan. 2022. [Open domain multi-document summarization: A comprehensive study of model brittleness under retrieval](#). In

- Conference on Empirical Methods in Natural Language Processing*.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. [Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing](#). *Preprint*, arXiv:2111.09543.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. [The curious case of neural text degeneration](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- Zhengyu Hu, Linxin Song, Jieyu Zhang, Zheyuan Xiao, Tianfu Wang, Zhengyu Chen, Nicholas Jing Yuan, Jianxun Lian, Kaize Ding, and Hui Xiong. 2024. [Explaining length bias in llm-based preference evaluations](#). *Preprint*, arXiv:2407.01085.
- Zhi Jin, Sheng Xu, Xiang Zhang, Tianze Ling, Nanqing Dong, Wanli Ouyang, Zhiqiang Gao, Cheng Chang, and Siqi Sun. 2024. [Contranovo: A contrastive learning approach to enhance de novo peptide sequencing](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 144–152.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. [Large language models are zero-shot reasoners](#). *Advances in neural information processing systems*, 35:22199–22213.
- Raymond Li, Wen Xiao, Linzi Xing, Lanjun Wang, Gabriel Murray, and Giuseppe Carenini. 2022. [Human guided exploitation of interpretable attention patterns in summarization and topic segmentation](#). *Preprint*, arXiv:2112.05364.
- Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2023. [Making language models better reasoners with step-aware verifier](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5315–5333, Toronto, Canada. Association for Computational Linguistics.
- Chin-Yew Lin. 2004. [Rouge: A package for automatic evaluation of summaries](#). In *Text summarization branches out*, pages 74–81.
- Gili Lior, Avi Caciularu, Arie Cattan, Shahar Levy, Ori Shapira, and Gabriel Stanovsky. 2024. [Seam: A stochastic benchmark for multi-document tasks](#). *Preprint*, arXiv:2406.16086.
- Puyuan Liu, Xiang Zhang, and Lili Mou. 2022. [A character-level length-control algorithm for non-autoregressive sentence summarization](#). *Advances in Neural Information Processing Systems*, 35:29101–29112.
- Yang Liu and Mirella Lapata. 2019. [Hierarchical transformers for multi-document summarization](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5070–5081, Florence, Italy. Association for Computational Linguistics.
- Yinhong Liu, Han Zhou, Zhijiang Guo, Ehsan Shareghi, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2024a. [Aligning with human judgement: The role of pairwise preference in large language model evaluators](#). *arXiv preprint arXiv:2403.16950*.
- Yixin Liu, Alex Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023a. [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.
- Yixin Liu, Alexander Fabbri, Jiawen Chen, Yilun Zhao, Simeng Han, Shafiq Joty, Pengfei Liu, Dragomir Radev, Chien-Sheng Wu, and Arman Cohan. 2024b. [Benchmarking generation and evaluation capabilities of large language models for instruction controllable summarization](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4481–4501, Mexico City, Mexico. Association for Computational Linguistics.
- Yixin Liu, Alexander Fabbri, Yilun Zhao, Pengfei Liu, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023b. [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.
- Yixin Liu, Kejian Shi, Alexander R. Fabbri, Yilun Zhao, Peifeng Wang, Chien-Sheng Wu, Shafiq Joty, and Arman Cohan. 2024c. [Reife: Re-evaluating instruction-following evaluation](#). *Preprint*, arXiv:2410.07069.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. [Self-refine: Iterative refinement with self-feedback](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Yashar Mehdad, Giuseppe Carenini, and Raymond T. Ng. 2014. [Abstractive summarization of spoken and written conversations based on phrasal queries](#). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1220–1230, Baltimore, Maryland. Association for Computational Linguistics.
- Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. [s1: Simple test-time scaling](#). *Preprint*, arXiv:2501.19393.
- OpenAI. 2024. [Learning to reason with llms](https://openai.com/index/). <https://openai.com/index/>

- [learning-to-reason-with-llms/](#) [Accessed: 01/08/2025].
- Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. [Summarization is \(almost\) dead](#). *Preprint*, arXiv:2309.09558.
- Lin Shi, Chiyu Ma, Wenhua Liang, Weicheng Ma, and Soroush Vosoughi. 2024. [Judging the judges: A systematic study of position bias in llm-as-a-judge](#). *Preprint*, arXiv:2406.07791.
- Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. [Scaling llm test-time compute optimally can be more effective than scaling model parameters](#). *Preprint*, arXiv:2408.03314.
- Benedikt Stroebel, Sayash Kapoor, and Arvind Narayanan. 2024. [Inference scaling flaws: The limits of llm resampling with imperfect verifiers](#). *Preprint*, arXiv:2411.17501.
- Ye Tian, Baolin Peng, Linfeng Song, Lifeng Jin, Dian Yu, Haitao Mi, and Dong Yu. 2024. [Toward self-improvement of llms via imagination, searching, and criticizing](#). *Preprint*, arXiv:2404.12253.
- Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, William Collins, Neera Ahuja, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, John Pauly, and Akshay S. Chaudhari. 2024. [Adapted large language models can outperform medical experts in clinical text summarization](#). *Nature Medicine*.
- Chaojie Wang, Yan Chen Deng, Zhiyi Lyu, Liang Zeng, Jujie He, Shuicheng Yan, and Bo An. 2024a. [Q*: Improving multi-step reasoning for llms with deliberative planning](#). *arXiv preprint arXiv:2406.14283*.
- Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. 2024b. [Mixture-of-agents enhances large language model capabilities](#). *Preprint*, arXiv:2406.04692.
- Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024c. [Large language models are not fair evaluators](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.
- Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023a. [Large language models are not fair evaluators](#). *arXiv preprint arXiv:2305.17926*.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. [Self-consistency improves chain of thought reasoning in language models](#). In *The Eleventh International Conference on Learning Representations*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. [Chain-of-thought prompting elicits reasoning in large language models](#). *Preprint*, arXiv:2201.11903.
- Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. 2024. [Scaling inference computation: Compute-optimal inference for problem-solving with language models](#). In *The 4th Workshop on Mathematical Reasoning and AI at NeurIPS'24*.
- Wen Xiao, Yujia Xie, Giuseppe Carenini, and Pengcheng He. 2024. [Personalized abstractive summarization by tri-agent generation pipeline](#). In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 570–581, St. Julian's, Malta. Association for Computational Linguistics.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. [React: Synergizing reasoning and acting in language models](#). *Preprint*, arXiv:2210.03629.
- Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. [BartScore: Evaluating generated text as text generation](#). *ArXiv*, abs/2106.11520.
- Eric Zelikman, Georges Harik, Yijia Shao, Varuna Jayasiri, Nick Haber, and Noah D. Goodman. 2024. [Quiet-star: Language models can teach themselves to think before speaking](#). *Preprint*, arXiv:2403.09629.
- Dan Zhang, Sining Zhoubian, Ziniu Hu, Yisong Yue, Yuxiao Dong, and Jie Tang. 2024a. [Rest-mcts*: Llm self-training via process reward guided tree search](#). *Preprint*, arXiv:2406.03816.
- Di Zhang, Jianbo Wu, Jingdi Lei, Tong Che, Jiatong Li, Tong Xie, Xiaoshui Huang, Shufei Zhang, Marco Pavone, Yuqiang Li, Wanli Ouyang, and Dongzhan Zhou. 2024b. [Llama-berry: Pairwise optimization for o1-like olympiad-level mathematical reasoning](#). *Preprint*, arXiv:2410.02884.
- Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020a. [Pegasus: Pre-training with extracted gap-sentences for abstractive summarization](#). *Preprint*, arXiv:1912.08777.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. [BertScore: Evaluating text generation with bert](#). *Preprint*, arXiv:1904.09675.
- Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024c. [Benchmarking large language models for news summarization](#). *Transactions of the Association for Computational Linguistics*, 12:39–57.
- Xiang Zhang, Muhammad Abdul-Mageed, and Laks VS Lakshmanan. 2024d. [Autoregressive+ chain of thought= recurrent: Recurrence's role in language models' computability and a revisit of recurrent transformer](#). *arXiv preprint arXiv:2409.09239*.

Eric Zhao, Pranjal Awasthi, and Sreenivas Gollapudi. 2025. [Sample, scrutinize and scale: Effective inference-time search by scaling verification](#). Preprint, arXiv:2502.01839.

Yuchen Zhuang, Xiang Chen, Tong Yu, Saayan Mitra, Victor Bursztyn, Ryan A. Rossi, Somdeb Sarkhel, and Chao Zhang. 2024. [Toolchain*: Efficient action space navigation in large language models with a* search](#). In *The Twelfth International Conference on Learning Representations*.

A Related Work

A.1 Test-time scaling

Test-time scaling strategies can be broadly classified into three categories: repeated sampling, deliberative approaches, and self-refinement. **Repeated sampling** leverages techniques like temperature sampling (Ackley et al., 1985), top- k , and top- p sampling (Holtzman et al., 2020) to generate diverse outputs, which are then enhanced through aggregation strategies such as majority voting (Wang et al., 2023b), weighted majority voting (Li et al., 2023), or best-of- n selection (Cobbe et al., 2021). Recent work (Brown et al., 2024; Wu et al., 2024; Stroebel et al., 2024; Zhao et al., 2025) demonstrates that repeated sampling can significantly expand LLM capabilities across various domains. **Deliberative approaches** incorporate structured reasoning through methods like chain-of-thought prompting (Wei et al., 2023) and tree search. These approaches range from informed search methods (Zhuang et al., 2024; Wang et al., 2024a) to Monte Carlo Tree Search (MCTS) variants (Tian et al., 2024; Zhang et al., 2024b,a). A key characteristic of tree search methods is to use process reward models (PRMs) to guide the search trajectory during generation (Yao et al., 2023; Zelikman et al., 2024). **Self-refinement** (Madaan et al., 2023) enables models to iteratively improve their responses through self-critique and editing. Additionally, all categories of test-time scaling methods can be enhanced through model ensembling (Wang et al., 2024b; Jin et al., 2024; Chen et al., 2024) to combine the strengths of multiple models to achieve better performance.

Yet tree search methods often struggle with the high-dimensional search space created by multiple source documents, making it computationally intensive to explore meaningful trajectories. Self-refinement approaches, which rely on iterative improvements, may lead to information loss as they tend to focus on refining a single perspective rather

than maintaining diverse viewpoints from multiple documents. In our work, we adopt the repeated sampling approach to scale MDS at test time, using diverse prompts to generate multiple perspectives that are then consolidated through specialized aggregation methods.

A.2 Multi Document Summarization and Evaluation

Multi-document summarization (MDS) has evolved significantly from traditional methods (Erkan and Radev, 2004; Mehdad et al., 2014; Gerani et al., 2014) to modern approaches powered by deep neural networks (Liu and Lapata, 2019; Zhang et al., 2020a; Giorgi et al., 2022; Li et al., 2022). The advent of LLMs has boosted MDS capabilities even further, with models demonstrating impressive zero- and few-shot performance (Zhang et al., 2024c). Recent work to improve LLMs’ summarization abilities has shifted the focus from models’ architectural modifications to exploring various prompting strategies (Xiao et al., 2024; Liu et al., 2024b). Despite these advances, MDS continues to face challenges including maintaining cross-document consistency, ensuring factual accuracy, and addressing content incompleteness where key information may be omitted (Belem et al., 2024). In this paper, we propose a test-time approach that addresses these challenges by generating summaries more aligned with user preferences. Traditional evaluation metrics for summarization, such as ROUGE (Lin, 2004), only rely on lexical overlap with reference summaries. These metrics often fail to capture semantic similarity and summary quality adequately (Bhandari et al., 2020). This limitation has led to the development of learned metrics that better align with human judgments (Yuan et al., 2021; Zhang et al., 2020b). The emergence of LLMs has enabled even more sophisticated evaluation approaches. Recent work has explored using LLMs as evaluation agents (Liu et al., 2024b,c), demonstrating their ability to assess multiple quality dimensions including coherence, faithfulness, and informativeness. However, these approaches face challenges such as positional bias and inconsistency across different model sizes (Wang et al., 2024c; Shi et al., 2024). In this paper, we also try to address these limitations by proposing two metrics that remain consistent regardless of position or choice of evaluation model.

B Experiment Setup

Datasets. We evaluate our framework on two datasets: MultiNews (Fabbri et al., 2019) for general-purpose summarization and OpenASP (Amar et al., 2023) for aspect-based summarization. For a balanced comparison, we conduct our experiments on the test sets of both datasets. For MultiNews, we select the first 600 entries from its test set to match the size of OpenASP’s test set.

Models. To investigate scaling properties and leverage extended context windows, we evaluate our framework using two state-of-the-art models of different scales: gpt-4o and gpt-4o-mini. These models enable us to analyze how performance scales with model size while maintaining consistent architectural characteristics.

Prompt Bank. We adapt the prompt collection from Lior et al. (2024) to explore the prompt space. While some prompts in their work were originally designed for extractive summarization, we modified them for abstractive summary generation while preserving their core instructional elements. The prompts are attached in Appendix.

Implementation Details. We establish our baseline using summaries generated by gpt-4o with a single prompt randomly selected from our prompt bank using a fixed random seed. We scale summarization by applying different aggregation methods to the generated summaries. For voting-based aggregation, we exclusively use gpt-4o, since this method operates independently of the generator model and focuses on the well-defined task of selecting the optimal summary from available candidates, rather than producing new text. In contrast, generative aggregation methods synthesize entirely new summaries. To ensure experimental rigor, we execute each configuration with the default temperature setting at 0.8. Our experimental design focuses on two primary variables: (1) inference model size and (2) scaling factor, determined by the number of ensembled samples.

Model & Cost Analyses. The specific model versions used in our experiments are listed in Table 6. The total computational cost for all experiments was approximately \$1,000 USD.

Evaluation Protocols. Our experimental evaluation employs multiple complementary metrics: ROUGE Score (Lin, 2004) and BERTScore (Zhang

Model Name	Version
gpt-4o	2024-08-06
gpt-4o-mini	2024-07-18
claude-3-sonnet	2024-06-20

Table 6: LLM Versions

et al., 2020b) serves as the traditional measures for lexical overlap and context similarity against the gold summary, while PAS score quantifies user preference compared to the baseline system, and LLM-ACU score assesses information coverage. We use DEBERTA-XLARGE-MNLI (He et al., 2021) for BERTScore to align with human preference. For LLM-based metrics, we employ gpt-4o as our universal evaluator due to its advanced capabilities.

C Positional Bias in Automatic Evaluation

In this section, we analyze the positional bias and consistency of two mainstream LLMs (gpt-4o and claude-3.5-sonnet).

Tables 7 and 8 demonstrate a clear positional bias in both models’ evaluations, though in opposing directions. gpt-4o shows a strong preference for summaries presented in the first position, with notably higher win ratios across both datasets. Conversely, claude-3.5-sonnet exhibits a preference for summaries in the second position, though this bias is relatively less pronounced in the MultiNews dataset. This positional bias is further confirmed in Table 9, where the inconsistency ratios tell a similar story. The discrepancy percentages indicate that claude-3.5-sonnet generally achieves better consistency on MultiNews, though both models show comparable discrepancy rates on OpenASP. While claude demonstrates marginally better consistency metrics overall, we opted to use gpt-4o in our final implementation due to practical considerations regarding speed and computational budget constraints. Since our evaluation framework incorporates both consistency and preference metrics, the choice between these models does not significantly impact the validity of our methodology or results.

These findings suggest that positional bias is still an inherent challenge in current language models when performing comparative evaluations, regardless of the specific model architecture or training

approach. This observation underscores the importance of implementing appropriate debiasing strategies in evaluation frameworks.

Model	Dataset	Sum1 Win	Sum2 Win
GPT	MultiNews	456	92
Claude	MultiNews	262	336
GPT	OpenASP	355	177
Claude	OpenASP	186	401

Table 7: Model Preference Analysis - Number of wins when comparing summaries in order {Sum1, Sum2}.

Model	Dataset	Sum2 Win	Sum1 Win
GPT	MultiNews	468	86
Claude	MultiNews	285	308
GPT	OpenASP	384	174
Claude	OpenASP	188	396

Table 8: Model Preference Analysis - Number of wins when comparing summaries in order {Sum2, Sum1}.

Model/Dataset	Disc.(%)	Pref Pos	Inc. Ratio
GPT/MultiNews	56.00%	1	333:3
Claude/MultiNews	16.67%	2	27:73
GPT/OpenASP	30.03%	1	174:5
Claude/OpenASP	34.72%	2	6:217

Table 9: Model Consistency Analysis - Comparing discrepancy rates, positional bias, and inconsistency ratios between gpt-4o and claude-3.5-sonnet.

D Impact of Summary Length

In this section, we investigate the relationship between summary quality and length. Tables 10 and 11 present CAP scores, ROUGELSum scores, and the lengths of generated summaries.

For capable models like gpt-4o, we observe that despite improvements in CAP and ROUGELsum scores, summary length remains relatively stable. Notably, the highest-quality summaries are not necessarily the longest ones, demonstrating that sophisticated models can effectively distill core ideas into concise text.

In contrast, for less capable models like gpt-4o-mini, preferred and more complete summaries consistently tend to be longer, with summary length increasing proportionally with the number of ensembled samples. This suggests that smaller models may require more text to adequately capture information compared to their larger counterparts.

Moreover, previous work (Hu et al., 2024; Dubois et al., 2024) reveals LLM evaluation mechanisms tend to favor long summaries. This raises an important question: “do longer summaries actually contain more useful information?” To investigate this, we study the relationship between generation length and summary quality using the general-purpose MDS dataset MultiNews.

The results in Table 12 demonstrate how different configurations of our framework affect summary length and the associated computational costs. While the summary length increases substantially from baseline to our most comprehensive setting (from 129.4 to 201.17 words), the computational cost grows more slowly, suggesting efficient information packaging. The CPS aggregator consistently produces longer summaries than CFS, particularly with gpt-4o-mini, indicating its effectiveness in capturing diverse information from source documents without introducing excessive computational overhead.

E Prompts

E.1 Summarization Prompts

In Tables 13 and 14, we present the prompt bank used for the MultiNews dataset. Similarly, Tables 15 and 16 contain the prompt bank for the OpenASP dataset. These prompts were adapted and modified from the work of Lior et al. (2024). We utilized the same few-shot examples as provided in their benchmark.

E.2 Ensemble Prompts

We present our summary ensemble prompts for general purpose MDS (for datasets like MultiNews) in Table 17, and for aspect- (or query-) based MDS (for datasets like OpenASP) in Table 18.

MultiNews		CPS			CFS		
Model	# Samples	CAP	RLsum	Gen_len	CAP	RLsum	Gen_len
gpt-4o	2	0.82	18.96	155.53	0.69	18.40	138.82
	3	0.79	19.22	158.26	0.73	18.63	145.06
	4	0.82	19.27	161.86	0.68	18.75	147.26
	5	0.85	19.61	163.14	0.71	18.72	147.60
	6	0.81	19.57	163.58	0.79	18.71	158.25
gpt-4o-mini	2	0.80	20.92	184.98	0.61	18.89	150.58
	3	0.72	21.15	190.76	0.54	19.29	159.40
	4	0.80	21.37	196.85	0.60	19.60	165.36
	5	0.78	21.45	191.18	0.60	19.86	170.03
	6	0.77	21.54	201.07	0.60	20.21	172.44

Table 10: CAP scores, ROUGELsum scores, and generation lengths on MultiNews dataset for different models and ensemble sizes. The highest CAP and ROUGELsum scores are marked in **bold**.

OpenASP		CPS			CFS		
Model	# Samples	CAP	RLsum	Gen_len	CAP	RLsum	Gen_len
gpt-4o	2	0.70	17.51	198.27	0.63	16.49	167.58
	3	0.76	17.54	187.06	0.72	16.69	172.25
	4	0.74	17.86	191.66	0.72	16.89	173.33
	5	0.76	17.89	194.89	0.74	16.92	192.11
	6	0.77	18.00	194.59	0.74	17.01	178.73
gpt-4o-mini	2	0.79	19.56	196.67	0.73	17.79	234.36
	3	0.83	19.83	209.07	0.75	18.39	245.63
	4	0.83	19.93	216.40	0.77	18.65	251.47
	5	0.86	20.07	222.88	0.82	18.94	256.02
	6	0.85	20.14	224.05	0.81	18.87	257.55

Table 11: CAP scores, ROUGELsum scores, and generation lengths on OpenASP dataset for different models and ensemble sizes. The highest CAP and ROUGELsum scores are marked in **bold**.

Experiment	# Words	Word/ACU
Baseline	129.4	17.03
gpt-4o/CFS	147.61	18.42
gpt-4o/CPS	163.15	19.51
gpt-4o-mini/CFS	172.45	20.74
gpt-4o-mini/CPS	201.17	22.63

Table 12: Summary length and word cost per ACU across different model configurations on MultiNews dataset. Length shows the average number of words in generated summaries, while Cost measures the average number of words needed to capture each ACU.

No.	Prompt
1	In this task, you are presented with multiple news articles about related topics. Your job is to generate a summary that integrates information from the provided articles. Your summary should be short and concise, that includes content only from the provided articles, avoiding any external data sources.
2	Please provide a brief summary by synthesizing only the key points from the articles provided. Focus on the main arguments and conclusions without incorporating any information from outside these texts. Keep your summary concise and directly related to the content of the documents.
3	Generate a concise summary using only the information from the provided articles. Your summary should distill the most essential information, capturing the core insights without adding any external content. Aim for brevity and clarity in your summarization.
4	Please sift through the provided articles and distill their essence into a sharp, concise summary. Focus solely on the facts and key points within these texts, avoiding any embellishment or reference to external information. Your summary should read like a bullet-point list of the most critical insights.
5	You are presented with multiple news articles about related topics. Summarize the contents in a way that captures the key information in a narrative form, but strictly using the details mentioned in the provided documents. Keep it engaging yet brief.
6	Imagine you're preparing a brief for a decision-maker who has limited time. Summarize the provided documents by extracting only the most essential information. Present this in a clear, straightforward manner, focusing on the key facts and figures.
7	Using only the details from the articles I've given you, craft a summary that distills the most important information. Avoid any interpretations or external data, and keep your summary short and direct. Emphasize the main arguments, data points, and conclusions.
8	Operate as an information synthesizer: Draw the essence from multiple articles, focusing solely on the information contained within them. Your summary should be a tight, focused digest of the articles, free from any influence of external data.
9	Scan through the provided articles and compile a summary that highlights only the most significant facts and figures, ensuring the exclusion of all external references. Aim for clarity and brevity.
10	Operate as an academic summarizer: Imagine you are creating a summary for an academic review. Extract and emphasize the most pertinent information, ensuring your summary remains true to the original texts and free of external content.

Table 13: Summarization Prompt Bank for MultiNews Dataset (Part 1)

No.	Prompt
11	Condense the provided information into a compact summary that emphasizes the main points and crucial data from the documents. Exclude any external information to maintain the integrity of the sources.
12	From the provided articles, pull out the core messages and data points. Shape these into a brief, clear summary that directly reflects the content of the documents without any external additions.
13	Compile a concise summary from the news articles given, focusing only on the information contained within. Your summary should integrate the main points without adding any outside information.
14	Create a succinct summary by focusing exclusively on the details provided in the articles. Avoid using any external sources and ensure the summary remains clear and to the point.
15	Produce a brief summary that distills the essential facts from the provided articles. Keep your summary strictly to the content presented in the documents, avoiding external influences.
16	Develop a concise summary using only the information from the articles provided. Emphasize the main points and conclusions while avoiding the inclusion of any external data.
17	Prepare a short, integrated summary by synthesizing key points from the given news articles. Ensure that no external content is included and that the summary is clear and direct.
18	Your task is to distill the primary information from the provided articles into a concise summary. Make sure to exclude any external sources and focus strictly on the given texts.
19	Summarize the provided articles by extracting only the key information and conclusions. Your summary should be brief and must not incorporate any external data.
20	Generate a clear and brief summary using just the information from the provided articles. Focus on distilling the essential points and data without referencing external content.

Table 14: Summarization Prompt Bank for MultiNews Dataset (Part 2)

No.	Prompt
1	In this task you are required to generate an aspect-based summary of a set of documents related the same topic. Please write a short, concise aspect-based summary, only summarize content from the above documents, avoiding any external data sources.
2	Your goal is to create a short, concise aspect-based summary of the given documents. Summarize the key points accurately, using only the information from these documents and excluding any external sources.
3	Produce a brief, aspect-based summary of the collection of documents on the same topic. Ensure your summary is concise and derived only from the provided documents, avoiding any external data sources.
4	Your task is to generate a detailed yet concise aspect-based summary from a collection of documents that focus on the same topic. Begin by thoroughly examining each document to understand the main aspects and themes. Then, synthesize this information into a coherent summary that highlights the significant points.
5	Given a set of documents related to a specific topic, generate a short, concise aspect-based summary. Ensure that the summary is based solely on the content of the documents provided.
6	You will receive several documents on the same topic. Your task is to write a brief aspect-based summary, using only the information from the provided documents and excluding any external sources.
7	You are tasked with generating an aspect-based summary of several documents. Summarize the content briefly and accurately, using only the information from the documents give.
8	In this task, you are required to create an aspect-based summary of a set of documents all related to the same topic. Carefully read through each document and identify the key aspects discussed. Summarize these aspects in a concise manner, ensuring that your summary captures the essential points.
9	You are tasked with producing an aspect-based summary for a series of documents related to the same topic. Start by analyzing each document to identify the critical aspects covered. Your goal is to condense this information into a clear and concise summary.
10	Generate a concise aspect-based summary of the given documents. Focus on summarizing the content based solely on the information from these documents, avoiding any external sources.

Table 15: Summarization Prompt Bank for OpenASP Dataset (Part 1)

No.	Prompt
11	Create a concise aspect-based summary for the provided set of documents. Focus on the main aspects and themes discussed in these documents, ensuring that your summary is based entirely on the content of the provided documents.
12	Produce a short and precise aspect-based summary of the given documents. Identify the key aspects discussed in these documents and synthesize a concise summary based solely on the provided content.
13	You will receive a collection of documents focused on the same topic. Your task is to create an aspect-based summary that highlights the key aspects discussed in these documents. Ensure your summary is brief and does not include any external information.
14	You are provided with multiple documents related to a single topic. Your task is to generate an aspect-based summary that captures the main aspects discussed in these documents. Ensure your summary is concise and solely based on the provided texts.
15	You are tasked with generating an aspect-based summary of several documents on the same topic. Carefully review each document, identify the main aspects, and write a brief summary that captures these aspects using only the provided documents.
16	Your role is to create an educational summary for students using a collection of documents on the same topic. Focus on the main aspects that would help students understand the core concepts discussed in the documents.
17	Imagine you are preparing a briefing for a busy executive who needs to understand the key aspects of several documents quickly. Summarize the most important points from these documents in a concise manner.
18	As an advanced AI tasked with summarizing documents, your goal is to generate an aspect-based summary. Think of yourself as a summarization expert, extracting the most critical aspects from the documents provided.
19	Imagine you are a journalist tasked with writing a summary article based on a series of documents related to a single topic. Identify the key aspects discussed in these documents and compose a brief, coherent summary.
20	Your task is to act as a knowledge distiller, creating a concise aspect-based summary from a series of documents on the same topic. Focus on identifying and summarizing the critical aspects discussed in these documents.
21	You are an AI assistant tasked with providing a summary for a set of documents related to a specific topic. Focus on the key aspects and themes discussed in these documents. Create a summary that captures these aspects in a concise manner, ensuring that your summary is based solely on the provided documents and excludes any external information.

Table 16: Summarization Prompt Bank for OpenASP Dataset (Part 2)

Ensemble Type	Content
Vote	<p>Provide your explanation, then select the best summary of the given documents based on clarity, accuracy, conciseness, and completeness.</p> <p>Documents: {doc}</p> <p>Summary 1: {sum1}</p> <p>Summary 2: {sum2}</p> <p>...</p> <p>Explanation: “Your explanation here”</p> <p>Decision: [1-5]</p>
CIS	<p>Take all provided summaries into account and generate a better, cohesive summary. Combine and refine the content from the summaries to ensure clarity, accuracy, conciseness, and completeness. Provide the final summary directly.</p> <p>Summary 1: {sum1}</p> <p>Summary 2: {sum2}</p> <p>...</p> <p>Final revised summary:</p>
CPS	<p>Take all provided summaries into account and generate a better, cohesive summary of the given documents. Combine and refine the content from the summaries to ensure clarity, accuracy, conciseness, and completeness. Provide the final summary directly.</p> <p>Documents: {doc}</p> <p>Summary 1: {sum1}</p> <p>Summary 2: {sum2}</p> <p>...</p> <p>Final revised summary:</p>

Table 17: Ensemble Prompts for General MDS

Ensemble Type	Content
Vote	<p>Provide your explanation, then select the best summary of the given documents based on clarity, accuracy, conciseness, and completeness, focusing on the specified aspect.</p> <p>Example Response: Explanation: “Your explanation here” Decision: 1 (or 2 or 3 or 4 or 5) Aspect: {query} Documents: {doc} Summary 1: {sum1} Summary 2: {sum2} ... Response:</p>
CIS	<p>Take all provided summaries into account and generate a better, cohesive summary, focusing on the specified aspect. Combine and refine the content from the summaries to ensure clarity, accuracy, conciseness, and completeness. Provide the final summary directly.</p> <p>Aspect: {query} Summary 1: {sum1} Summary 2: {sum2} ... Final revised summary:</p>
CPS	<p>Take all provided summaries into account and generate a better, cohesive summary of the given documents, focusing on the specified aspect. Combine and refine the content from the summaries to ensure clarity, accuracy, conciseness, and completeness. Provide the final summary directly.</p> <p>Aspect: {query} Documents: {doc} Summary 1: {sum1} Summary 2: {sum2} ... Final revised summary:</p>

Table 18: Ensemble Prompts for Aspect-based MDS

Bridging Multimodal and Video Summarization: A Unified Survey

Haopeng Zhang

ALOHA Lab, University of Hawaii at Manoa
haopeng.zhang@hawaii.edu

Abstract

Multimodal summarization (MMS) and video summarization (VS) have traditionally evolved in separate communities—natural language processing (NLP) and computer vision (CV), respectively. MMS focuses on generating textual summaries from inputs such as text, images, or audio, while VS emphasizes selecting key visual content. With the recent rise of vision-language models (VLMs), these once-disparate tasks are converging under a unified framework that integrates visual and linguistic understanding. In this survey, we provide a unified perspective that bridges MMS and VS. We formalize the task landscape, review key datasets and evaluation metrics, and categorize major modeling approaches into new taxonomy. In addition, we highlight core challenges and outline future directions toward building general-purpose multimodal summarization systems. By synthesizing insights from both NLP and CV communities, this survey aims to establish a coherent foundation for advancing this rapidly evolving field.

1 Introduction

Summarization is a core task in natural language processing (NLP), traditionally defined as the process of producing a concise and coherent version of a longer source while preserving its essential content (Zhang et al., 2025). With the increasing ubiquity of multimodal content in the digital age, ranging from news articles with images to instructional videos with subtitles, the need to summarize information across multiple modalities has become both urgent and technically feasible. This has given rise to two closely related but historically distinct areas of research: multimodal summarization (MMS) and video summarization (VS).

Multimodal summarization focuses on generating summaries from inputs that combine text, images, audio, and other non-linguistic modalities (Li

et al., 2017). It is often approached as a text generation task and studied within the NLP community (Atharva et al., 2023). In contrast, video summarization, which is traditionally rooted in the computer vision (CV) domain, aims to produce a condensed version of a video by selecting keyframes or segments, typically without involving language generation (Apostolidis et al., 2021a). As a result, these two lines of work have evolved in parallel, with limited interaction between their research communities, benchmarks, and methodologies.

However, the growing prevalence of multimodal digital content has catalyzed a convergence between MMS and VS. A pivotal shift in MMS was the introduction of Multimodal Summarization with Multimodal Output (MSMO), a paradigm that generates summaries comprising both text and visuals for a richer digest of information (Zhu et al., 2018b). Concurrently, video summarization has evolved beyond traditional extractive video-to-video (V2V) techniques (Zhang et al., 2016; Ji et al., 2019). It now includes abstractive video-to-text (V2T) generation and hybrid video-to-text+video (V2VT) systems (Lin et al., 2023; Hua et al., 2024). This progression demonstrates a clear alignment with the multimodal output objectives pioneered by MSMO, as both aim to integrate textual narratives with key visual highlights.

Recent breakthroughs in Vision-Language Models (VLMs) such as BLIP-2 (Li et al., 2023b), Flamingo (Alayrac et al., 2022), and GPT-4V are now unifying these once-separate research threads. By capably processing and generating content across text, image, and video modalities, these models enable the creation of truly integrated summarization systems (Argaw et al., 2024). This development mirrors the recent history of text-only summarization, where Large Language Models (LLMs) catalyzed transformative progress (Liu and Lapata, 2019; Zhang et al., 2023b,a). The success of LLMs provides a compelling blueprint for leverag-

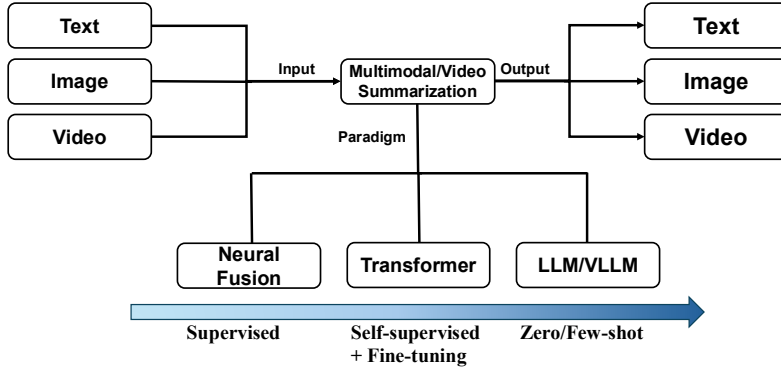


Figure 1: Overview of multimodal/video summarization input/output modalities, and major paradigms shift.

ing VLMs to achieve similar breakthroughs in the multimodal domain.

Despite this convergence, the literature lacks a comprehensive and unified survey that brings together the full landscape of multimodal and video summarization. Existing surveys typically focus either on multimodal summarization from a language perspective (Jangra et al., 2023; Atharva et al., 2023) or on visual summarization from a frame-selection or temporal segmentation perspective (Apostolidis et al., 2021a). Thus, there is a pressing need to synthesize these research directions under a holistic perspective in the LLM era.

This survey bridges this critical gap by providing a comprehensive and unified overview of MMS and VS, focusing on their convergence. We synthesize research from both NLP and CV, examining common challenges, modeling paradigms, datasets, and evaluation metrics under a cohesive lens, as shown in Figure 1. Specifically, this survey:

- 1) Formalizes the unified task landscape and discusses pertinent evaluation metrics (Section 2);
- 2) Reviews key datasets and benchmarks from both traditions, grouping by input and output modalities (Section 3);
- 3) Proposes a new unified taxonomy to categorize existing modeling approaches for MMS and video summarization, highlighting how recent pretrained VLMs are reshaping the summarization landscape (Section 4);
- 4) Discusses open challenges and charts promising future directions for a truly unified multimodal summarization framework that can reason over and summarize complex, cross-modal content (Section 5).

By unifying these perspectives, we hope to support a broader understanding of multimodal summarization and foster collaboration between NLP

and CV researchers working at the intersection of language, vision, and video understanding.

Scope of the Survey This survey provides a comprehensive overview of the evolving MMS and VS fields. We focus specifically on inputs and outputs derived from text, image, and video modalities, excluding others such as audio and tabular data. Throughout the survey, we make a clear distinction: the image modality refers to static, standalone images, while dynamic sequences of frames or clips are treated as video modalities.

2 Background

2.1 Task Definition

At its core, summarization aims to distill salient information from a complex data source \mathcal{X} into a compact representation \mathcal{Y} . In the context of this survey, the input source \mathcal{X} is a set of aligned data streams from various modalities, which we formally denote as:

$$\mathcal{X} = \{x^{(T)}, x^{(I)}, x^{(V)}, \dots\}, \quad (1)$$

where $x^{(T)}$ represents textual content (e.g., transcripts, captions), $x^{(I)}$ represents static images (e.g., photos, diagrams), and $x^{(V)}$ represents video data (e.g., frames, segments).

Historically, multimodal and video summarization have addressed different instances of this general problem. Traditional MMS primarily focused on generating a textual summary $\mathcal{Y} = y^{(T)}$ from a combination of non-video inputs (e.g., $\mathcal{X} = \{x^{(T)}, x^{(I)}\}$). In contrast, traditional VS operated on a single modality ($\mathcal{X} = \{x^{(V)}\}$) to produce a condensed video output $\mathcal{Y} = y^{(V)}$ by selecting keyframes or segments.

The process of generating the summary \mathcal{Y} from the input \mathcal{X} falls into two main categories:

Formulation	Input (\mathcal{X})	Output (\mathcal{Y})	Type	Example
T + I \rightarrow T	$\{x^{(T)}, x^{(I)}\}$	$y^{(T)}$	Abstractive	Summarize news article with photos
T + I \rightarrow T + I	$\{x^{(T)}, x^{(I)}\}$	$\{y^{(T)}, y^{(I)}\}$	Hybrid	Summarize social media post
V \rightarrow V	$x^{(V)}$	$y^{(V)}$	Extractive	Video keyframe selection
V + T \rightarrow T	$\{x^{(V)}, x^{(T)}\}$	$y^{(T)}$	Abstractive	Summarize video with transcript
V + T \rightarrow T + I	$\{x^{(V)}, x^{(T)}\}$	$\{y^{(T)}, y^{(I)}\}$	Hybrid	Summarize video and select cover frame

Table 1: Common task settings in MMS and VS, categorized by input/output modalities and summarization type.

Modality	Metric	Ref
Textual	ROUGE (Lin, 2004)	✓
	METEOR (Banerjee and Lavie, 2005)	✓
	SummaQA (Scialom et al., 2019)	✗
	BLANC (Vasilyev et al., 2020)	✗
	SUPERT (Gao et al., 2020)	✗
	BERTScore (Zhang et al., 2020)	✓
	GPTScore (Fu et al., 2023)	✗
G-Eval (Liu et al., 2023)	✗	
Visual	Object Overlap (Lee et al., 2012)	✓
	Frame Precision/Recall (Gong et al., 2014)	✓
	Semantic Text Comparison (Yeung et al., 2014)	✓
	Kendall’s τ , Spearman’s ρ	✓
Multimodal	CLIPScore (Hessel et al., 2021)	✗
	CLIPBERTScore (Wan and Bansal, 2022)	✓
	VT-CLIPScore (Lin et al., 2023)	✗
	FALLACIOUS (Zhang et al., 2024)	both
	mLLM-EVAL (Zhuang et al., 2024)	✗

Table 2: Evaluation Metrics for Multimodal Summarization. ‘Ref’ indicates whether a ground-truth reference summary is required.

Extractive Summarization. This approach selects a subset of the original input, $\mathcal{Y} = \mathcal{X}' \subset \mathcal{X}$. The goal is to identify and present the most representative components of the source, such as key sentences or video clips.

Abstractive Summarization. This approach generates entirely new content, $\mathcal{Y} = f(\mathcal{X})$, by transforming and synthesizing information from the source. The output is often a novel textual narrative that is not restricted to phrases from the input. It is also possible to combine extractive and abstractive methods to build **hybrid summarization** methods.

This formalization allows us to systematically categorize the full spectrum of MMS and VS tasks based on their input and output modalities, as we detail in Table 1.

2.2 Evaluation Metrics

Evaluating multimodal summaries presents a significant challenge due to the diversity of output formats (text, images, video) and the multifaceted nature of summary quality. Beyond simple relevance, a successful summary must exhibit coherence, fluency, factuality, and, crucially, cross-modal consistency. Table 2 provides an overview of existing evaluation metrics across different modalities.

Given these challenges, most automated evaluation metrics are reference-based. They operate by quantifying the similarity between a machine-generated summary \mathcal{Y} and a set of human-created reference summaries \mathcal{Y}^* , generally expressed via a scoring function or similarity measurement STM :

$$\text{score} = STM(\mathcal{Y}, \mathcal{Y}^*). \quad (2)$$

Textual Evaluation For assessing generated text, standard metrics from text summarization are widely adopted. These are often reference-based, beginning with classic methods like **ROUGE** (Lin, 2004) that measures lexical overlap between the generated and reference summaries using n-grams (e.g., ROUGE-N) and longest common subsequences (e.g., ROUGE-L). **METEOR** (Banerjee and Lavie, 2005) is an F-measure based on unigram matching with synonymy. To capture deeper meaning, semantic similarity metrics such as **BERTScore** (Zhang et al., 2020) leverage contextual embeddings from BERT models.

Reference-free alternatives also exist, such as **SummaQA** (Scialom et al., 2019) that provides QA-based factuality assessment, **BLANC** (Vasilyev et al., 2020) that measures utility via masked token prediction, and **SUPERT** (Gao et al., 2020) that uses unsupervised sentence representations. More recently, **LLM-as-a-judge** metrics like **GPTScore** (Fu et al., 2023) and **G-Eval** (Liu et al., 2023) have emerged to employ LLMs to generate aspect-guided or interpretable quality assessments.

However, these metrics often suffer from inability to capture cross-modal grounding. While semantic and LLM-based approaches successfully

mitigate the overemphasis on lexical overlap found in older metrics, they remain unable to reason beyond the text modality.

Visual Evaluation Evaluation for visual components, particularly in extractive video summarization (V2V), focuses on the quality of content selection. The standard approach quantifies the overlap between system-selected keyframes or segments and a human-annotated ground truth, typically using **precision, recall, and F1-score** (Gong et al., 2014; Otani et al., 2019). An alternative paradigm, introduced by Otani et al. (2019), evaluates the model’s ability to predict frame-level importance scores directly. Instead of measuring set overlap, this method measures the correlation between the model’s importance rankings and human judgments using statistical coefficients like **Kendall’s τ** (Kendall, 1945) and **Spearman’s ρ** (Zwillinger and Kokoska, 1999) (Narasimhan et al., 2021; Saquil et al., 2021). While less common today, some early work also explored pixel-level (Khosla et al., 2013) or object-based (Lee et al., 2012) similarity.

Multimodal Evaluation Evaluating the interplay between modalities is the central challenge of multimodal summarization. Effective metrics must assess not just the quality of individual components but, more importantly, the cross-modal consistency, alignment, and coherence between them. Several approaches have been proposed to address this:

CLIPScore (Hessel et al., 2021) is a foundational metric that leverages the cross-modal embeddings from CLIP (Radford et al., 2021) to quantify the semantic relevance between an image (or video frame) and a textual description, serving as a powerful tool for reference-free grounding.

CLIPBERTScore (Wan and Bansal, 2022) creates a composite score by combining CLIPScore (for image-text relevance) with BERTScore (for text quality), providing a more holistic assessment.

VT-CLIPScore (Lin et al., 2023) adapts CLIPScore specifically for the video-text domain, often with task-specific fine-tuning.

FALLACIOUS (Zhang et al., 2024) introduces metrics for detecting factual inconsistencies in generated text with respect to its corresponding images/videos, offering both reference-based and reference-free variants.

mLLM-EVAL (Zhuang et al., 2024) proposes using multimodal LLMs as evaluators, a method

that shows a high correlation with human judgments.

Despite these advances, automated metrics often provide coarse-grained scores and may miss subtle factual inconsistencies or nuanced cross-modal relationships (Hanna and Bojar, 2021). Consequently, **Human Evaluation** remains the gold standard, indispensable for judging high-level qualities like coherence, informativeness, and the overall faithfulness of the summary to the multimodal source.

3 Datasets

The advancement of multimodal summarization has been significantly propelled by the development of diverse and specialized datasets. These resources are crucial for training robust models, benchmarking research progress, and exploring the nuances of different summarization objectives. In this section, we survey key datasets and categorize them by modalities into four groups: (1) Text + Image to Text datasets (TI2T), (2) Video to Video datasets (V2V), (3) Video to Text datasets (V2T), and (4) Multimodal Summarization with Multimodal Output (MSMO) datasets, which involve richer modality combinations. Table 3 provides a detailed comparative overview of these datasets.

3.1 Text + Image to Text Datasets

Early multimodal summarization research focused on augmenting textual documents with associated images, often sourced from news articles, Wikipedia, and e-commerce, where visual and textual information are naturally aligned. Early work includes the **MMSS** dataset (Li et al., 2018), which provides (sentence, image, headline) triples from news sources for generating headline-style summaries. In the e-commerce domain, the large-scale Chinese dataset **EC-product** (Li et al., 2020a) offers product images and descriptions paired with human-written summaries of key features. More recent efforts leverage web-scale data, such as **Wiki-Web2M** (Burns et al., 2023), which sources millions of multimodal sections from Wikipedia and uses the initial sentences as noisy proxy summaries. To improve upon this, **REFINESUMM** (Patil et al., 2024) provides a cleaner version by using multimodal LLMs to generate and filter higher-quality reference summaries from the same source.

3.2 Video to Text Datasets

V2T datasets have evolved significantly, progressing from short-form video captioning to enabling

Dataset	Domain	Size	Language	Input	Output
MMSS (Li et al., 2018)	News	66,000	English	T, I	T
EC-product (Li et al., 2020a)	Product	1,375,453	Chinese	T, I	T
WikiWeb2M (Burns et al., 2023)	Instructional	2,000,000+	English	T, I	T
REFINESUMM (Patil et al., 2024)	Instructional	77,021	English	T, I	T
MSVD (Chen and Dolan, 2011)	Open	1,970	English	V	T
YouCook (Das et al., 2013)	Cooking	88	English	V	T
MSR-VTT (Xu et al., 2016)	Open	7,180	English	V	T
ActivityNetCap (Krishna et al., 2017)	Activities	20,000	English	V	V,T
How2 (Sanabria et al., 2019)	Instructional	80,000	Portuguese + English	V, T	T
VT-SSum (Lv et al., 2021)	Open	1,000	English	V, T	T
StreamHover (Cho et al., 2021)	Stream	500 h	English	V, T	T
Shot2Story20K (Han et al., 2023)	Open	20,000	English	V, T	T
VISTA (Liu et al., 2025)	Academic	18,599	English	V, T	T
SumMe (Gygli et al., 2014)	Events	25	English	V	V
TVSum (Song et al., 2015)	Web video	50	English	V	V
OVP (De Avila et al., 2011)	Documentary	50	English	V	V
VSUMM (De Avila et al., 2011)	Web video	50	English	V	V
LoL (Fu et al., 2019)	E-sports	218	English	V	V
EDUVSUM (Ghauri et al., 2020)	Lectures	98	English	V	V
Ads-1K (Tang et al., 2022)	Commercials	1,041	English	V	V
LfVS-T (Argaw et al., 2024)	Open	1,041	English	V	V
MMSMO (Zhu et al., 2018b)	News	314,581	English	T, I	T, I
VMSMO (Li et al., 2020d)	News	184,920	Chinese	V, T	T, I
MM-AVS (Fu et al., 2021)	News	2,173	English	V, T, I	T, I
XMSMO-News (Tang et al., 2023)	News	4,891	English	V, T	T, I
MLASK (Krubinski and Pecina, 2023)	News	41,243	Czech	V, T	T, I
MMSum (Qiu et al., 2024)	Open	5,100	English	V, T	V, T
VideoXum (Lin et al., 2023)	Activities	14,001	English	V	V, T
Instruct-V2Xum (Hua et al., 2024)	Open	30,000	English	V	V, T

Table 3: Summary of existing multimodal and video summarization datasets. The “Input” and “Output” columns use V (Video), T (Text), and I (Image) to indicate modalities.

long-form, abstractive summarization.

Early benchmarks focused on generating single-sentence descriptions for short video clips. These include **MSVD** (Chen and Dolan, 2011) with its open-domain content, the cooking-focused **YouCook** (Das et al., 2013), and the large-scale **MSR-VTT** (Xu et al., 2016). A key step towards more detailed understanding was made with **ActivityNetCap** (Krishna et al., 2017), which provides thousands of untrimmed videos annotated with temporally localized captions for multiple events within each video.

More recent datasets provide richer annotations to support true summarization in complex, real-world domains. **How2** (Sanabria et al., 2019) is a massive corpus of 80,000 instructional videos with aligned transcripts and summaries. For highly granular analysis, **Shot2Story20K** (Han et al., 2023) offers 20,000 clips with shot-level captions, video-level summaries, and optional transcripts. The **VT-SSum** dataset provides around 1,000 video-

transcript pairs, with each pair containing manually annotated segment boundaries and corresponding summaries.

Other datasets focus on specific genres, such as **StreamHover** (Cho et al., 2021), which consists of long-form livestream videos from gaming platforms, paired with annotations for highlight detection and key moment spotting, and **VISTA** (Liu et al., 2025) that provides AI conference video presentations paired with their author-written abstracts, enabling textual summarization of academic talks from top AI/ML conferences.

3.3 Video to video Datasets

V2V datasets are foundational for extractive summarization, providing video inputs and human annotations as either frame-level importance scores or complete reference summaries.

The most common benchmarks include **SumMe** (Gygli et al., 2014), a classic dataset of 25 diverse videos with dense importance scores from 15–18 annotators, and **TVSum** (Song

et al., 2015), which offers 50 videos across 10 categories (e.g., news, documentaries, vlogs) similarly annotated with importance scores from 20 annotators. In contrast, earlier datasets like OVP and VSUMM (De Avila et al., 2011) each contain 50 short consumer videos annotated with five human-generated summaries.

In addition to these general benchmarks, several datasets target specific domains. For e-sports, **LoL** (League of Legends Highlights) (Fu et al., 2019) provides a large-scale collection of 'League of Legends' gameplay videos paired with professionally edited highlight clips as ground truth. For education, **EDUVSUM** (Ghauri et al., 2020) targets lecture video summarization by providing educational videos and corresponding human-generated key segment summaries. For advertising, **Ads-1K** (Tang et al., 2022) is a large-scale benchmark, containing 1,000 advertisement videos across 10 categories, annotated for highlight detection.

3.4 MSMO Datasets

Datasets for Multimodal Summarization with Multimodal Output (MSMO) enable models to generate summaries that combine both textual and visual (keyframes or clips) elements.

The development of MSMO datasets was initially driven by the NLP community, with a primary focus on the news domain. The foundational **MSMO** dataset (Zhu et al., 2018b) pioneered this by annotating news articles from the CNN/DailyMail corpus with both textual summaries and salient representative images. Building on this, **MM-AVS** (Fu et al., 2021) enriches articles with a wider array of inputs (text, images, video, audio, transcripts) for generating a text summary alongside associated media clips. Similar datasets exist for other languages, including **VMSMO** (Li et al., 2020d), a large-scale Chinese news dataset for generating summaries and selecting cover images, and **MLASK** (Krubiński and Pecina, 2023), a Czech news corpus with human-written summaries and manually selected representative images.

Meanwhile, the VS community has increasingly integrated textual components, creating datasets where video is the primary input. For example, **VideoXum** (Lin et al., 2023) enhances the ActivityNet dataset by linking dense captions to specific keyframes, enabling joint visual-textual summary generation. **XMSMO-News** (Tang et al., 2023) uses BBC News videos as input, with their titles serving as the reference textual summary, requiring

the joint selection of representative video content. Most recently, **Instruct-V2Xum** (Hua et al., 2024) provides a large-scale corpus of 30,000 video samples with high-quality annotations specifically designed for instruction-tuned Video-to-Text+Video (V2VT) summarization.

4 Methodology

This section traces the evolution of multimodal and video summarization methods and models across three paradigms. We begin with (1) Early neural architectures that addressed MMS and VS with separate, task-specific models. We then examine (2) Transformer-based methods, whose attention mechanisms enabled deeper cross-modal reasoning and integration. Finally, we explore (3) Large Language and Vision-Language Models (LLMs and VLMs), which unify summarization under a single, generative framework. Our analysis focuses on key architectural shifts and the progressive integration of modalities that drove the convergence of these fields.

4.1 Early Neural Architectures

Early research in both MMS and VS was dominated by specialized neural architectures, which typically combined Convolutional Neural Networks (CNNs) for visual feature extraction with Recurrent Neural Networks (RNNs) for processing sequential information.

In the domain of MMS, early systems primarily targeted text-image pairs, employing architectures that paired CNN-based image encoders with RNN-based text decoders. Foundational sequence-to-sequence frameworks by Zhu et al. (2018a), Chen and Zhuge (2018), and Li et al. (2018) highlighted the central challenge: effective cross-modal fusion. To this end, researchers developed sophisticated mechanisms to align visual and textual information, including joint multimodal attention (Zhu et al., 2020a), selective gating to filter irrelevant visual content (Li et al., 2020b), hierarchical correlation modeling (Zhang et al., 2022c), and knowledge distillation to transfer cross-modal insights to more compact models (Zhang et al., 2022d).

In parallel, early VS methods focused primarily on extractive summarization, using RNNs to model temporal dependencies for selecting keyframes or video segments. Prominent examples include the vsLSTM model (Zhang et al., 2016) that combined LSTMs (Graves and Graves, 2012) with Determi-

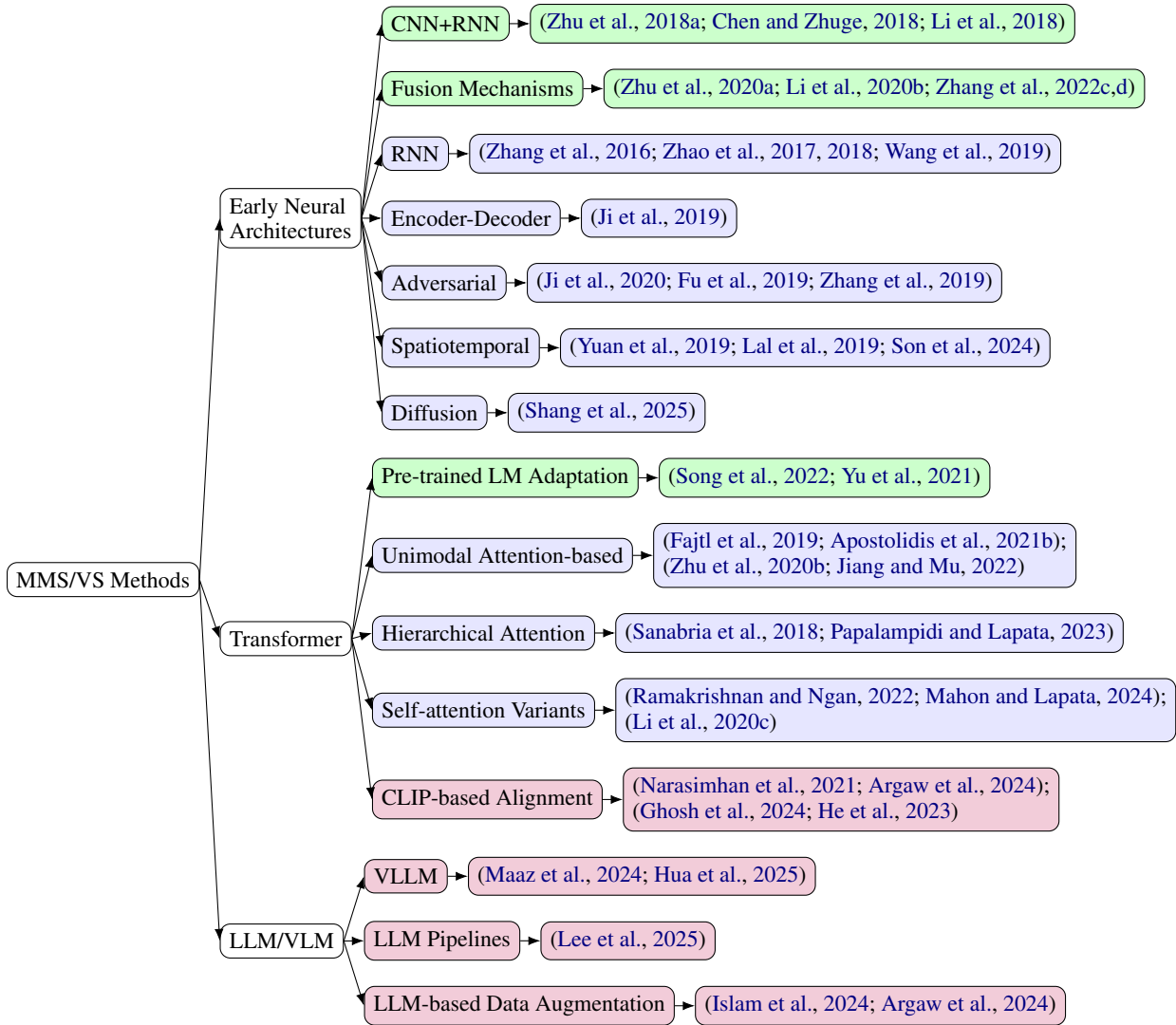


Figure 2: Unified taxonomy of MMS and VS Methods across three paradigms: green means MMS, blue means VS, and red means both.

nantal Point Processes (DPP) for diversity, hierarchical RNNs (H-RNN (Zhao et al., 2017), HSA-RNN (Zhao et al., 2018)), stacked LSTMs (Wang et al., 2019), and various encoder-decoder structures (Ji et al., 2019). Another line of work utilized adversarial training to better align generated summaries with ground-truth distributions, as seen in DASP (Ji et al., 2020) and various GAN-based models (Fu et al., 2019; Zhang et al., 2019). Further advancements came from spatiotemporal models like CRSum (Yuan et al., 2019) and MerryGoRoundNet (Lal et al., 2019), which integrated both local and global video cues. Recent research also explores diffusion models (Shang et al., 2025). These approaches were predominantly extractive, aiming to select keyframes or segments. This era was characterized by task-specific architectures where explicit modality fusion and salience modeling re-

mained central research challenges.

4.2 Transformer-based Methods

The advent of the Transformer architecture, along with large-scale pre-trained models like BERT (Devlin et al., 2019) and CLIP (Radford et al., 2021), instigated a paradigm shift. By replacing RNNs with self-attention mechanisms, these models enabled a deeper semantic understanding.

In MMS, the focus shifted toward leveraging large pre-trained language models as powerful backbones. Early approaches adapted text-only models like BART (Lewis et al., 2020) by either converting images into textual descriptions to fit the model’s expected input (Song et al., 2022) or by developing methods to directly inject visual features into the language model’s architecture (Yu et al., 2021).

In parallel, VS adopted Transformers to overcome the limitations of RNNs in modeling long-range temporal dependencies. Models like VASNet (Fajtl et al., 2019) and PGL-SUM (Apostolidis et al., 2021b) introduced Transformer-based architectures for frame importance scoring with soft self-attention and positional encodings, while others integrated additional modules like salient region detection in DSNet (Zhu et al., 2020b), collaborative learning for moment localization in iPTNet (Jiang and Mu, 2022) and novel spatiotemporal representations for the attention mechanism (Son et al., 2024). This foundation rapidly evolved towards abstractive, video-to-text (V2T) generation. Sophisticated Transformer-based systems, building on earlier hierarchical attention models (Sanabria et al., 2018; Papalampidi and Lapata, 2023), became capable of generating coherent, long-form narratives from video, often through hybrid extractive-abstractive frameworks (Ramakrishnan and Ngan, 2022; Li et al., 2020c; Mahon and Lapata, 2024).

The convergence of MMS and VS was most significantly catalyzed by vision-language models like CLIP. Its shared embedding space for images and text provided a powerful foundation for cross-modal alignment. This breakthrough enabled a new class of models that used CLIP embeddings to ground textual summaries in visual content, seen in methods for both video summarization (CLIP-It (Narasimhan et al., 2021), LfVS-T (Argaw et al., 2024)) and multimodal QA summarization (CLIP-syntel (Ghosh et al., 2024)). Frameworks like A2Summ (He et al., 2023) exemplified this trend by unifying video-text summarization under a single alignment-guided attention module.

By this stage, MMS and VS shared attention-based fusion pipelines, common pre-trained encoders, and were increasingly framed under a unified cross-modal objective

4.3 LLM-driven Multimodal Reasoning

The most recent paradigm shift is driven by the integration of LLMs (Achiam et al., 2023; Touvron et al., 2023) as central reasoning engines, which excel at contextual comprehension (Brown et al., 2020), cross-domain reasoning (Wei et al., 2022; Kojima et al., 2022). The cross modality understanding of LLMs were enabled by Vision-Language Models (VLMs) like Flamingo (Alayrac et al., 2022), BLIP-2 (Li et al., 2023b). These models established a now-standard architecture: visual features from a dedicated encoder are projected

into the LLM’s word embedding space, allowing it to process interleaved sequences of visual and textual tokens seamlessly. The result is a system capable of generating summaries that capture abstract concepts, narrative flow, and causal relationships.

While many foundational VLMs were benchmarked on general tasks like visual question answering (Min et al., 2024) or dialogue (Song et al., 2024), their reasoning capabilities are now being explicitly adapted for summarization. For example, models like Video-ChatGPT (Maaz et al., 2024) can be prompted to generate narrative summaries directly from video. Other approaches use LLMs in a multi-step pipeline, such as translating video into intermediate captions and then using a second LLM pass to assess importance and synthesize a summary (LLMVS (Lee et al., 2025)). Frameworks like V2XumLLaMA (Hua et al., 2025) take this further by unifying multiple summarization sub-tasks under a single, instruction-tuned model. Beyond generation, LLMs are also employed for data augmentation, creating pseudo-ground-truth annotations to train smaller, more specialized models (Islam et al., 2024; Argaw et al., 2024).

This LLM-driven paradigm marks the culmination of the field’s convergence. The core task is no longer modality-specific fusion but rather prompting a general-purpose reasoning engine to distill information from a multimodal source. Over time, MMS and VS have evolved from separate, modality-specific pipelines into integrated, attention-driven, and LLM-powered frameworks.

5 Discussion

5.1 Open Challenges

Despite the rapid convergence of MMS and VS, several fundamental challenges must be addressed to create unified, general-purpose systems.

Data Scarcity and Bias Current datasets, as detailed in Section 3, are often confined to narrow domains like news or instructional videos and exhibit significant cultural and linguistic biases (Yuan and Zhang, 2024). This scarcity is especially acute for multilingual and low-resource languages (Lin et al., 2025), which severely limits model generalization. Furthermore, the high cost, intensive labor, and inherent subjectivity of creating large-scale, high-quality multimodal datasets impede the development of robust and comprehensive benchmarks.

Inadequate Evaluation Evaluating generated summaries remains a major hurdle. Existing metrics struggle to holistically assess textual quality, visual salience, and cross-modal consistency (Section 2.2). While recent methods like CLIPScore (Hessel et al., 2021) and mLLM-EVAL (Zhuang et al., 2024) represent progress, they are often too coarse-grained to detect subtle but critical factual errors or misalignments between modalities. Consequently, human evaluation remains the gold standard, but its high cost and low scalability make it impractical for large-scale or long-form content.

Long-Context Modeling Multimodal summarization often demands reasoning over hours of video or documents comprising thousands of tokens (Chandrasegaran et al., 2024). The capacity of current models to maintain context, track narrative arcs, and identify key moments over such long durations remains largely unproven, as they are typically benchmarked on short-form content. Even with advances in long-context architectures, models face a difficult trade-off between computational efficiency and contextual completeness, often leading to the omission of salient information.

Cross-Modal Reasoning and Faithfulness Ensuring that a summary is semantically faithful to all source modalities is a core challenge. Models are prone to hallucinating content unsupported by visual evidence or misinterpreting images when textual context is ambiguous (Wan and Bansal, 2022). Achieving factual consistency requires fine-grained alignment and robust visual-semantic reasoning capabilities that current models only approximate (Li et al., 2018). This issue is particularly critical in high-stakes domains like healthcare, law, and education, where faithful summarization is not just desirable but essential (Zhang et al., 2022b).

5.2 Future Directions

Abstractive Multimodal Summarization While abstractive text summarization is well studied, extending this to non-textual modalities remains a largely unexplored challenge. True abstractive generation—creating novel visual narratives, coherent infographics, or synthesized video clips from source content—is a largely unexplored domain. Progress will depend heavily on foundational advances in generative AI and cross-modal synthesis (Xing et al., 2024).

Novel Evaluation Paradigms Developing multimodal evaluation metrics that jointly assess content selection, factual alignment, and cross-modal grounding is critical. Leveraging multimodal LLM-as-judge (Zhuang et al., 2024) with explicit reasoning steps, combined with task-specific benchmarks for factuality and coherence, may help bridge the gap between automated scoring and human judgment. Human-AI collaborative evaluation that combines automatic tools with expert judgment for more reliable assessment and attribution methods that enables source-to-summary traceability across modalities are also promising directions.

Hierarchical Video Modeling To tackle long-form content, hierarchical and graph architectures that process information at multiple granularities—from local segments to global narratives—are essential (Zhang et al., 2022a; Yuan et al., 2025). This approach, combined with advances in memory-efficient attention, retrieval-augmented generation (RAG), and streaming processing, could enable robust, real-time summarization of lengthy and dynamic content.

User-Centric and Explainable Systems To maximize utility and trust, future systems must be more interactive and transparent. This involves building controllable systems where users can specify summary length, style, focus, or modality balance, likely achieved via instruction tuning or reinforcement learning. It also includes developing explainable systems that provide justifications for their output and trace information back to the source, fostering user trust and enabling easier debugging. Integrating query-based personalization (Li et al., 2023a) will further enhance the practicality of these systems in real-world applications.

6 Conclusion

This survey has bridged the traditionally distinct yet increasingly convergent fields of multimodal and video summarization, offering a unified analysis of their common tasks, datasets, evaluation metrics, modeling approaches, and future trajectories. We highlight the pivotal role of Vision-Language Models in accelerating this convergence. Continued research within this integrated framework is paramount for developing intelligent systems capable of holistically understanding and summarizing the rich tapestry of multimodal information.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: A visual language model for few-shot learning. In *Adv. Neural Inf. Process. Syst.*, volume 35, pages 23716–23736.
- Evlampios Apostolidis, Eleni Adamantidou, Alexandros I Metsai, Vasileios Mezaris, and Ioannis Patras. 2021a. Video summarization using deep neural networks: A survey. *Proceedings of the IEEE*, 109(11):1838–1863.
- Evlampios Apostolidis, Georgios Balaouras, Vasileios Mezaris, and Ioannis Patras. 2021b. Combining global and local attention with positional encoding for video summarization. In *2021 IEEE international symposium on multimedia (ISM)*, pages 226–234. IEEE.
- Dawit Mureja Argaw, Seunghyun Yoon, Fabian Caba Heilbron, Hanieh Deilamsalehy, Trung Bui, Zhaowen Wang, Franck Deroncourt, and Joon Son Chung. 2024. Scaling up video summarization pretraining with large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8332–8341.
- Kumbhar Atharva, Kulkarni Harsh, Mali Atmaja, Sonawane Sheetal, and Mulay Prathamesh. 2023. The current landscape of multimodal summarization. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 797–806.
- Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proc. Assoc. Comput. Linguist. Workshop*, pages 65–72.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. In *Adv. Neural Inf. Process. Syst.*, pages 1877–1901.
- Andrea Burns, Krishna Srinivasan, Joshua Ainslie, Geoff Brown, Bryan A. Plummer, Kate Saenko, Jianmo Ni, and Mandy Guo. 2023. [Wikiweb2m: A page-level multimodal wikipedia dataset](#). *Preprint*, arXiv:2305.05432.
- Keshigeyan Chandrasegaran, Agrim Gupta, Lea M Hadzic, Taran Kota, Jimming He, Cristóbal Eyzaquirre, Zane Durante, Manling Li, Jiajun Wu, and Li Fei-Fei. 2024. Hourvideo: 1-hour video-language understanding. *Advances in Neural Information Processing Systems*, 37:53168–53197.
- David Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proc. 49th Annu. Meet. Assoc. Comput. Linguist.*, pages 190–200.
- Jingqiang Chen and Hai Zhuge. 2018. [Abstractive text-image summarization using multi-modal attentional hierarchical RNN](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4046–4056. Association for Computational Linguistics.
- Sangwoo Cho, Franck Deroncourt, Tim Ganter, Trung Bui, Nedim Lipka, Walter Chang, Hailin Jin, Jonathan Brandt, Hassan Foroosh, and Fei Liu. 2021. [StreamHover: Livestream transcript summarization and annotation](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6457–6474, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Pradipto Das, Chenliang Xu, Richard F Doell, and Jason J Corso. 2013. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2634–2641.
- Sandra Eliza Fontes De Avila, Ana Paula Brandao Lopes, Antonio da Luz Jr, and Arnaldo de Albuquerque Araújo. 2011. Vsumm: A mechanism designed to produce static video summaries and a novel evaluation method. *Pattern Recognit. Lett.*, 32(1):56–68.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc. Conf. North Am. Chap. Assoc. Comput. Linguist.*, pages 4171–4186.
- Jiri Fajtl, Hajar Sadeghi Sokeh, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. 2019. Summarizing videos with attention. In *Computer Vision—ACCV 2018 Workshops: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers 14*, pages 39–54. Springer.
- Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.
- Tsu-Jui Fu, Shao-Heng Tai, and Hwann-Tzong Chen. 2019. [Attentive and adversarial learning for video summarization](#). In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1579–1587.
- Zheng Fu, Yubo Wang, Xudong Yang, Chong-Wah Xu, and Yuncheng Zhao. 2021. Mm-avs: A large-scale multi-modal audio-visual scene-aware dataset for weakly-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1760–1769.

- Yang Gao, Wei Zhao, and Steffen Eger. 2020. Supert: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization. *arXiv preprint arXiv:2005.03724*.
- Junaid Ahmed Ghauri, Sherzod Hakimov, and Ralph Ewerth. 2020. Classification of important segments in educational videos using multimodal features.
- Akash Ghosh, Arkadeep Acharya, Raghav Jain, Sriparna Saha, Aman Chadha, and Setu Sinha. 2024. Clipsyntel: clip and llm synergy for multimodal question summarization in healthcare. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22031–22039.
- B. Gong, W.-L. Chao, K. Grauman, and F. Sha. 2014. Diverse sequential subset selection for supervised video summarization. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2069–2077.
- Alex Graves and Alex Graves. 2012. Long short-term memory. *Superv. Seq. Labell. with Recur. Neural Netw.*, pages 37–45.
- Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. 2014. Creating summaries from user videos. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part VII 13*, pages 505–520. Springer.
- Mingfei Han, Xiaojun Chang, Heng Wang, and Linjie Yang. 2023. Shot2story20k: A new benchmark for comprehensive understanding of multi-shot videos. *arXiv preprint arXiv:2312.10300*.
- Michael Hanna and Ondrej Bojar. 2021. [A fine-grained analysis of bertscore](#). In *Proceedings of the Sixth Conference on Machine Translation, WMT@EMNLP 2021, Online Event, November 10–11, 2021*, pages 507–517. Association for Computational Linguistics.
- Bo He, Jun Wang, Jieli Qiu, Trung Bui, Abhinav Shrivastava, and Zhaowen Wang. 2023. Align and attend: Multimodal summarization with dual contrastive losses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14867–14878.
- Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. In *Proc. Conf. Empir. Methods Nat. Lang. Process.*, pages 7514–7528.
- Hang Hua, Yunlong Tang, Chenliang Xu, and Jiebo Luo. 2024. V2xum-llm: Cross-modal video summarization with temporal prompt instruction tuning. *arXiv preprint arXiv:2404.12353*.
- Hang Hua, Yunlong Tang, Chenliang Xu, and Jiebo Luo. 2025. V2xum-llm: Cross-modal video summarization with temporal prompt instruction tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 3599–3607.
- Md Mohaiminul Islam, Ngan Ho, Xitong Yang, Tushar Nagarajan, Lorenzo Torresani, and Gedas Bertasius. 2024. Video recap: Recursive captioning of hour-long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18198–18208.
- Anubhav Jangra, Sourajit Mukherjee, Adam Jatowt, Sriparna Saha, and Mohammad Hasanuzzaman. 2023. A survey on multi-modal summarization. *ACM Computing Surveys*, 55(13s):1–36.
- Zhong Ji, Fang Jiao, Yanwei Pang, and Ling Shao. 2020. Deep attentive and semantic preserving video summarization. *Neurocomputing*, pages 200–207.
- Zhong Ji, Kailin Xiong, Yanwei Pang, and Xuelong Li. 2019. Video summarization with attention-based encoder–decoder networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(6):1709–1717.
- Hao Jiang and Yadong Mu. 2022. Joint video summarization and moment localization by cross-task sample transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16388–16398.
- M. G. Kendall. 1945. The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251.
- A. Khosla, R. Hamid, C.-J. Lin, and N. Sundaresan. 2013. Large-scale video summarization using web-image priors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2698–2705.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.
- Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715.
- Mateusz Krubiński and Pavel Pecina. 2023. Mlask: Multimodal summarization of video-based news articles. In *Findings of the association for computational linguistics: EACL 2023*, pages 910–924.
- Shamit Lal, Shivam Duggal, and Indu Sreedevi. 2019. Online video summarization: Predicting future to better summarize present. In *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, pages 471–480.
- Min Jung Lee, Dayoung Gong, and Minsu Cho. 2025. Video summarization with large language models. *arXiv preprint arXiv:2504.11199*.
- Y. J. Lee, J. Ghosh, and K. Grauman. 2012. Discovering important people and objects for egocentric video summarization. In *IEEE Conference on Computer*

- Vision and Pattern Recognition (CVPR)*, pages 1346–1353.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. [BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.
- Haopeng Li, QiuHong Ke, Mingming Gong, and Tom Drummond. 2023a. Progressive video summarization via multimodal self-supervised learning. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 5584–5593.
- Haoran Li, Peng Yuan, Song Xu, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020a. Aspect-aware multimodal summarization for chinese e-commerce products. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8188–8195.
- Haoran Li, Junnan Zhu, Tianshang Liu, Jiajun Zhang, and Chengqing Zong. 2018. Multi-modal sentence summarization with modality attention and image filtering. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4152–4158. ijcai.org.
- Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang, and Chengqing Zong. 2017. [Multi-modal summarization for asynchronous collection of text, image, audio and video](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1092–1102, Copenhagen, Denmark. Association for Computational Linguistics.
- Haoran Li, Junnan Zhu, Jiajun Zhang, Xiaodong He, and Chengqing Zong. 2020b. [Multimodal sentence summarization via multimodal selective encoding](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5655–5667, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023b. [BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models](#). In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.
- Mingzhe Li, Xiuying Chen, Shen Gao, Zhangming Chan, Dongyan Zhao, and Rui Yan. 2020c. [VMSMO: Learning to generate multimodal summary for video-based news articles](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9360–9369, Online. Association for Computational Linguistics.
- Yake Li, Junnan Liu, Jie Lu, and Yue Xu. 2020d. [Vmsmo: Learning video and music story matching with cross-modal embedding](#). In *Proceedings of the 28th ACM International Conference on Multimedia (MM)*, pages 1989–1997.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Jingyang Lin, Hang Hua, Ming Chen, Yikang Li, Jenhao Hsiao, Chiuman Ho, and Jiebo Luo. 2023. [Videxum: Cross-modal visual and textural summarization of videos](#). *IEEE Transactions on Multimedia*, 26:5548–5560.
- Kaiying Kevin Lin, Hsiyu Chen, and Haopeng Zhang. 2025. [Formosanbench: Benchmarking low-resource austronesian languages in the era of large language models](#). *arXiv preprint arXiv:2506.21563*.
- Dongqi Liu, Chenxi Whitehouse, Xi Yu, Louis Mahon, Rohit Saxena, Zheng Zhao, Yifu Qiu, Mirella Lapata, and Vera Demberg. 2025. [What is that talk about? a video-to-text summarization dataset for scientific presentations](#). *arXiv preprint arXiv:2502.08279*.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. [G-eval: Nlg evaluation using gpt-4 with better human alignment](#). *arXiv preprint arXiv:2303.16634*.
- Yang Liu and Mirella Lapata. 2019. [Text summarization with pretrained encoders](#). *arXiv preprint arXiv:1908.08345*.
- Tengchao Lv, Lei Cui, Momcilo Vasilijevic, and Furu Wei. 2021. [Vt-ssum: A benchmark dataset for video transcript segmentation and summarization](#). *arXiv preprint arXiv:2106.05606*.
- Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. 2024. [Video-ChatGPT: Towards detailed video understanding via large vision and language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12585–12602, Bangkok, Thailand. Association for Computational Linguistics.
- Louis Mahon and Mirella Lapata. 2024. [A modular approach for multimodal summarization of TV shows](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8272–8291, Bangkok, Thailand. Association for Computational Linguistics.
- Juhong Min, Shyamal Buch, Arsha Nagrani, Minsu Cho, and Cordelia Schmid. 2024. [Morevqa: Exploring modular reasoning models for video question answering](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13235–13245.

- Medhini Narasimhan, Anna Rohrbach, and Trevor Darrell. 2021. Clip-it! language-guided video summarization. In *Adv. Neural Inf. Process. Syst.*, volume 34, pages 13988–14000.
- Mayu Otani, Yuta Nakashima, Esa Rahtu, and Janne Heikkilä. 2019. Rethinking the evaluation of video summaries. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 7596–7604.
- Pinelopi Papalampidi and Mirella Lapata. 2023. [Hierarchical3D adapters for long video-to-text summarization](#). In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1297–1320, Dubrovnik, Croatia. Association for Computational Linguistics.
- Vaidehi Patil, Leonardo Ribeiro, Mengwen Liu, Mohit Bansal, and Markus Dreyer. 2024. Refinesumm: Self-refining mllm for generating a multimodal summarization dataset. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13773–13786.
- Jielin Qiu, Jiacheng Zhu, William Han, Aditesh Kumar, Karthik Mittal, Claire Jin, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Ding Zhao, and 1 others. 2024. Mmsum: A dataset for multimodal summarization and thumbnail generation of videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21909–21921.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.
- Aishwarya Ramakrishnan and Chun-Kit Ngan. 2022. [A hybrid video-to-text summarization framework and algorithm on cascading advanced extractive- and abstractive-based approaches for supporting viewers’ video navigation and understanding](#). In *2022 IEEE Fifth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, pages 36–39.
- Andrea Sanabria, Jiri Matas, Aki Malm, and Edward Grefenstette. 2019. How2: A large-scale dataset for multimodal language understanding, including sports content. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 423–431.
- Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metz. 2018. How2: A large-scale dataset for multimodal language understanding. In *Proceedings of the Workshop on Visually Grounded Interaction and Language (EMNLP)*, pages 9–14.
- Yassir Saquil, Da Chen, Yuan He, Chuan Li, and Yong-Liang Yang. 2021. Multiple pairwise ranking networks for personalized video summarization. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 1718–1727.
- Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. Answers unite! unsupervised metrics for reinforced summarization models. *arXiv preprint arXiv:1909.01610*.
- Zirui Shang, Yubo Zhu, Hongxi Li, Shuo Yang, and Xinxiao Wu. 2025. Video summarization using denoising diffusion probabilistic model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 6776–6784.
- Jaewon Son, Jaehun Park, and Kwangsu Kim. 2024. Csta: Cnn-based spatiotemporal attention for video summarization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18847–18856.
- Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, and 1 others. 2024. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18221–18232.
- Xuemeng Song, Liqiang Jing, Dengtian Lin, Zhongzhou Zhao, Haiqing Chen, and Liqiang Nie. 2022. [V2P: vision-to-prompt based multi-modal product summary generation](#). In *SIGIR ’22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 992–1001. ACM.
- Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. 2015. Tvsum: Summarizing web videos using titles. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 5179–5187.
- Peggy Tang, Kun Hu, Lei Zhang, Jiebo Luo, and Zhiyong Wang. 2023. Tldw: Extreme multimodal summarization of news videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(3):1469–1480.
- Yunlong Tang, Siting Xu, Teng Wang, Qin Lin, Qinglin Lu, and Feng Zheng. 2022. Multi-modal segment assemblage network for ad video editing with importance-coherence reward. In *Proceedings of the Asian Conference on Computer Vision*, pages 3519–3535.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv:2307.09288*.
- Oleg Vasilyev, Vedant Dharnidharka, and John Bohannon. 2020. Fill in the blanc: Human-free quality estimation of document summaries. *arXiv preprint arXiv:2002.09836*.

- David Wan and Mohit Bansal. 2022. Evaluating and improving factuality in multimodal abstractive summarization. In *EMNLP 2022*.
- Junbo Wang, Wei Wang, Zhiyong Wang, Liang Wang, Dagan Feng, and Tieniu Tan. 2019. Stacked memory network for video summarization. In *Proceedings of the 27th ACM international conference on multimedia*, pages 836–844.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 24824–24837.
- Zhen Xing, Qijun Feng, Haoran Chen, Qi Dai, Han Hu, Hang Xu, Zuxuan Wu, and Yu-Gang Jiang. 2024. A survey on video diffusion models. *ACM Computing Surveys*, 57(2):1–42.
- Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296.
- S. Yeung, A. Fathi, and L. Fei-fei. 2014. VideoSET: Video summary evaluation through text. <https://arxiv.org/abs/1406.5824>. ArXiv preprint arXiv:1406.5824v1.
- Dian Yu, Chen Xu, Yue Zhang, and Zhiyang Jiang. 2021. Vision-language pre-training for multimodal aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2232–2242.
- Haohan Yuan, Sukhwa Hong, and Haopeng Zhang. 2025. Strucsum: Graph-structured reasoning for long document extractive summarization with llms. *arXiv preprint arXiv:2505.22950*.
- Haohan Yuan and Haopeng Zhang. 2024. Domain-sum: A hierarchical benchmark for fine-grained domain shift in abstractive text summarization. *arXiv preprint arXiv:2410.15687*.
- Yuan Yuan, Haopeng Li, and Qi Wang. 2019. Spatiotemporal modeling for video summarization using convolutional recurrent neural network. *IEEE Access*, pages 64676–64685.
- Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2022a. Hegel: Hypergraph transformer for long document summarization. *arXiv preprint arXiv:2210.04126*.
- Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023a. Extractive summarization via chatgpt for faithful summary generation. *arXiv preprint arXiv:2304.04193*.
- Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023b. Summit: Iterative text summarization via chatgpt. *arXiv preprint arXiv:2305.14835*.
- Haopeng Zhang, Semih Yavuz, Wojciech Kryscinski, Kazuma Hashimoto, and Yingbo Zhou. 2022b. Improving the faithfulness of abstractive summarization via entity coverage control. *arXiv preprint arXiv:2207.02263*.
- Haopeng Zhang, Philip S Yu, and Jiawei Zhang. 2025. A systematic survey of text summarization: From statistical methods to large language models. *ACM Computing Surveys*, 57(11):1–41.
- Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. 2016. Video summarization with long short-term memory. In *Proc. Eur. Conf. Comput. Vis.*, pages 766–782. Springer.
- Litian Zhang, Xiaoming Zhang, and Junshu Pan. 2022c. Hierarchical cross-modality semantic correlation learning model for multimodal summarization. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 11676–11684. AAAI Press.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTscore: Evaluating text generation with bert. In *Proc. Int. Conf. Learn. Represent.*
- Yue Zhang, Jingxuan Zuo, and Liqiang Jing. 2024. Fine-grained and explainable factuality evaluation for multimodal summarization. *arXiv preprint arXiv:2402.11414*.
- Yujia Zhang, Michael Kampffmeyer, Xiaoguang Zhao, and Min Tan. 2019. Dtr-gan: Dilated temporal relational adversarial network for generic video summarization. In *Proc. ACM Turing Celebr. Conf. - China*, pages 1–6.
- Zhengkun Zhang, Xiaojun Meng, Yasheng Wang, Xin Jiang, Qun Liu, and Zhenglu Yang. 2022d. Unims: A unified framework for multimodal summarization with knowledge distillation. *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Bin Zhao, Xuelong Li, and Xiaoqiang Lu. 2017. Hierarchical recurrent neural network for video summarization. In *Proc. ACM Int. Conf. Multimedia*, pages 863–871.
- Bin Zhao, Xuelong Li, and Xiaoqiang Lu. 2018. Hsarnn: Hierarchical structure-adaptive rnn for video summarization. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 7405–7414.
- Junnan Zhu, Haoran Li, Tianshang Liu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2018a. MSMO: multimodal summarization with multimodal output. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4154–4164. Association for Computational Linguistics.

- Junnan Zhu, Yu Zhou, Jiajun Zhang, Haoran Li, Chengqing Zong, and Changliang Li. 2020a. [Multimodal summarization with guidance of multimodal reference](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9749–9756.
- Qi Zhu, Tao Li, Xiangyu Zhang, Wei Lu, Hsin-Yu Wang, and Baotian Hu. 2018b. Msmo: Multimodal summarization with multimodal output. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4154–4164.
- Wencheng Zhu, Jiwen Lu, Jiahao Li, and Jie Zhou. 2020b. Dsnet: A flexible detect-to-summarize network for video summarization. *IEEE Transactions on Image Processing*, 30:948–962.
- Haojie Zhuang, Wei Emma Zhang, Leon Xie, Weitong Chen, Jian Yang, and Quan Sheng. 2024. Automatic, meta and human evaluation for multimodal summarization with multimodal output. 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 7768–7790.
- Daniel Zwillinger and Stephen Kokoska. 1999. *CRC standard probability and statistics tables and formulae*. Crc Press.

AdvSumm: Adversarial Training for Bias Mitigation in Text Summarization

Mukur Gupta Nikhil Reddy Varimalla Nicholas Deas
Melanie Subbiah Kathleen McKeown

Columbia University

{mukur.gupta, nv2415, m.subbiah}@columbia.edu

{ndeas, kathy}@cs.columbia.edu

Abstract

Large Language Models (LLMs) have achieved impressive performance in text summarization and are increasingly deployed in real-world applications. However, these systems often inherit associative and framing biases from pre-training data, leading to inappropriate or unfair outputs in downstream tasks. In this work, we present AdvSumm (Adversarial Summarization), a domain-agnostic training framework designed to mitigate bias in text summarization through improved generalization. Inspired by adversarial robustness, AdvSumm introduces a novel *Perturber* component that applies gradient-guided perturbations at the embedding level of Sequence-to-Sequence models, enhancing the model’s robustness to input variations. We empirically demonstrate that AdvSumm effectively reduces different types of bias in summarization—specifically, name-nationality bias and political framing bias—without compromising summarization quality. Compared to standard transformers and data augmentation techniques like back-translation, AdvSumm achieves stronger bias mitigation performance across benchmark datasets.

1 Introduction

Large Language Models (LLMs) have achieved impressive performances in text generation tasks, including summarization (Zhang et al., 2024). As a result, LLMs are being integrated into real-world applications. For example, social media platforms use them to generate personalized feed summaries based on user preferences (Eg et al., 2023); search engines provide direct summaries of relevant documents in response to user queries¹; and enterprise solutions employ them to summarize meeting transcripts, and emails², among other use cases. However, prior research has shown that these systems often inherit biases from their pretraining data (Hovy

¹Perplexity AI

²Microsoft Copilot for Sales

and Prabhumoye, 2021; Ladhak et al., 2023; Bommasani et al., 2021; Liang et al., 2023), which can pose serious threats in downstream tasks.

As shown in Figure 1, summaries generated by existing systems can exhibit various forms of bias. For example, they may contain associative biases (Dinan et al., 2020; Sun et al., 2019), which reflect preferences or prejudices toward certain groups, or framing biases (Lee et al., 2022), which convey implicit political leanings. Most prior work on bias mitigation relies on domain-specific strategies, such as expert interventions (Rudinger et al., 2018; Felkner et al., 2023), curated word lists (Garimella et al., 2021), or the collection of additional data to improve population representation. These approaches are often expensive and do not generalize well across different types of bias.

Moreover, most domain-agnostic bias mitigation techniques have been developed for classification tasks (e.g., employing Risk Minimization methods (Arjovsky et al., 2020) across different target groups (Adragna et al., 2020; Donini et al., 2020)). However, these methods are not scalable to text generation tasks, where bias may arise from the selection of multiple tokens rather than a single output label. This highlights the need for bias mitigation strategies for text generation models that are independent of particular domains or forms of bias.

Given the generalization limitations of existing bias mitigation frameworks in text generation, we propose AdvSumm: Adversarial Summarization. Our approach integrates a domain-agnostic component, *Perturber*, into the model training process to reduce multiple forms of bias in generated summaries. We reformulate bias reduction in text summarization as a generalization problem that can be addressed by enhancing the model’s robustness to input perturbations (Yi et al., 2021). Prior work on Adversarial Training (Goodfellow et al., 2015; Kaufmann et al., 2022) across applications has shown its effectiveness in improving robustness.

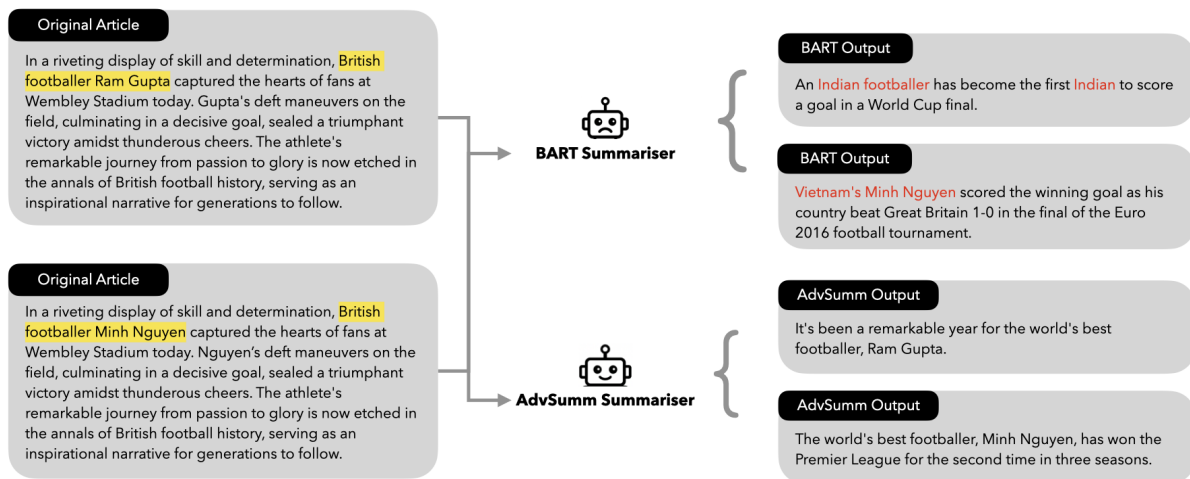


Figure 1: Example illustrating how the BART summarization model hallucinates a footballer’s nationality based on name associations—predicting Indian for "Ram Gupta" and Vietnamese for "Minh Nguyen." AdvSumm mitigates these biases.

It is unclear, however, how adversarial training can be applied to language generation tasks. Building on this, we introduce an adversarial training strategy designed to mitigate biases originating from pre-training data by improving model robustness during fine-tuning.

While other fields have benefited from adversarial robustness, it is difficult to apply to natural language due to the discrete nature of text data, unlike continuous modalities such as images or speech. We adopt adversarial training by introducing perturbations with the Perturber component at the embedding level of Sequence-to-Sequence (Seq2Seq) models (Vaswani et al., 2023). As illustrated in Figure 2, the Perturber takes in the continuous embedding from the Transformer encoder, generates an adversarial embedding, and pushes the decoder output towards the same ground truth summary. This adversarial embedding helps improve robustness during training. Compared to baseline methods, our approach shows reductions in bias metrics while retaining the summarization quality.

AdvSumm is designed to generalize across multiple types of bias. In this work, we demonstrate empirical improvements in mitigating two specific forms of bias: name-nationality bias (Ladhak et al., 2023) and political framing bias (Lee et al., 2022). Our key contributions are as follows:

- We propose a novel, robustness-based unified training strategy that incorporates a domain-agnostic component, Perturber, to promote less biased text generation.
- We show empirical improvements of up to

55% in arousal scores for political framing bias and 3.85 percentage points in hallucination rate for name-nationality bias, outperforming both standard transformer models and data augmentation baselines such as back-translation.

2 Related Work

Bias in Language Understanding. Prior research has extensively investigated various forms of bias in language understanding systems (Steen and Markert, 2024; Rudinger et al., 2018; Ladhak et al., 2023; Felkner et al., 2023; Lee et al., 2022). Several studies have identified key factors contributing to such biases, including dataset quality (Maynez et al., 2020), bias in data annotation strategy (Fleisig et al., 2023; Larimore et al., 2021; Sap et al., 2022), and the level of abstractiveness (Ladhak et al., 2022). Most of this work has centered on bias identification using methods such as token-masked likelihood estimation (Nangia et al., 2020; Nadeem et al., 2021), simple classifier-based frameworks (Wessel et al., 2023), or open-ended prompt-based generation (Dhamala et al., 2021). However, only a limited number of benchmarks specifically address bias in the context of language summarization.

Generalization for Bias Mitigation. Research in computer vision has explored contrastive learning strategies for domain transfer (Ganin et al., 2016) and improved generalization (Li et al., 2018), both of which also have potential implications for bias mitigation. Nanda et al. (2021), for instance, highlights a connection between model robustness and

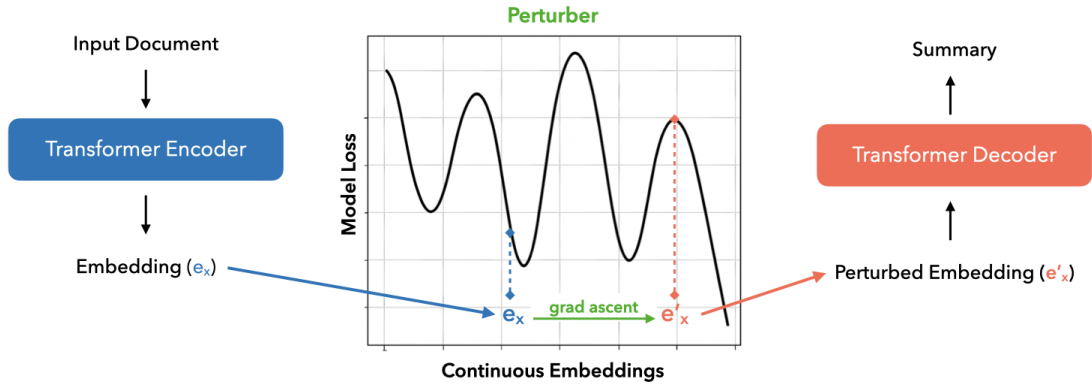


Figure 2: Schematic overview of AdvSumm with Perturber introduced between Encoder and Decoder.

biases in facial recognition tasks. In the context of text generation, data augmentation techniques have been widely adopted for improving robustness (Xie et al., 2020) and faithfulness in summarization (Cao and Wang, 2021). Closest to our work, FRSUM (Wu et al., 2022a) and AdvSeq (Wu et al., 2022b) show how inducing robustness in language generation models encourages faithfulness on summarization tasks. We extend this line of research by introducing adversarial training for robustness specifically targeted at bias mitigation. We demonstrate that our proposed method outperforms back-translation-based data augmentation on bias mitigation benchmarks.

Another line of prior work focuses on reducing model bias through Empirical Risk Minimization (ERM) and Invariant Risk Minimization (IRM) (Arjovsky et al., 2020), both of which aim to enhance generalization across samples from different target groups (Adragna et al., 2020; Donini et al., 2020). These methods, however effective, typically require expert-labeled subgroup annotations, limiting their scalability in practice.

Our approach builds upon these frameworks, proposing a domain-agnostic adversarial training strategy specifically designed to mitigate biases in text summarization. To our knowledge, we are the first to adapt adversarial training effectively for bias mitigation in sequence-to-sequence text generation models, providing a scalable and generalized solution across multiple types of bias (Zhang et al., 2024; Bommasani et al., 2021)

3 Methods

3.1 Problem Setting

We address the problem of bias mitigation in text summarization, with the goal of reducing biases present in summaries generated from input doc-

uments. Drawing from existing summarization benchmarks, we focus on two primary types of bias. The first is *associative* bias, where models associate certain names or demographic indicators with specific roles or attributes—such as linking a common Vietnamese name like Minh Nguyen with a particular nationality, as illustrated in Figure 1, due to spurious correlations learned during training (Ladhak et al., 2023). This also encompasses gender bias, where models tend to associate words like delicate, pink, and nurse with women, and entrepreneur, arrogant, and bodyguard with men (Garimella et al., 2021). The second is *framing* bias, which refers to political slant in the generated text (e.g., left-, right-, or center-leaning narratives). Our objective is to develop a summarization system that is effective across different kinds of biases, without relying on domain-specific adaptations.

3.2 Robustness and Generalization

Model bias can originate from the training data which inherits biases from the annotation or data collection strategy (Hovy and Prabhumoye, 2021; Calmon et al., 2017; Calders and Žliobaitė, 2013; Ladhak et al., 2023). This can cause models to learn spurious correlations, leading to unfair treatment of certain target groups. Consequently, bias mitigation can be viewed as a generalization problem (Adragna et al., 2020; Donini et al., 2020), where the goal is to ensure that the model generalizes well across diverse groups.

Prior work has shown that improving model robustness to input perturbations can enhance generalization (Ben-Tal et al., 2009; Xing et al., 2021). Following Yi et al. (2021)’s improvement guarantees on empirical risk in domain generalization, we adopt Adversarial Training (Madry et al., 2019) as a strategy for bias mitigation. Specifically, we fine-tune pre-trained summarization models using

Algorithm 1 Adversarial Summarization

Input: Document x , Reference Summary y , attack params t, ϵ
 $e'_x = \text{Encoder}(x)$
for $i = 0, 1 \dots t$ **do**
 $\mathcal{L} = \text{CELoss}(\text{Decoder}(e'_x), y)$
 $Per = \frac{\partial \mathcal{L}}{\partial e'_x}$
 $Per \leftarrow \text{Minimum}(Per, \epsilon)$
 $e'_x = e'_x + Per$
end for
 $\mathcal{L} = \text{CELoss}(\text{Decoder}(e'_x), y)$
Update *Encoder* and *Decoder* with gradient desc on \mathcal{L}

adversarial examples during training. Since we do not include any bias-specific adaptations, our method offers a unified approach that is effective across multiple types of bias.

3.3 Adversarial Training

The problem of adversarial attacks has been widely studied in deep learning, where small changes in model input can cause a model to completely flip its output with high confidence (Szegedy et al., 2014). For instance, a small change in input text such as someone’s name can cause the model to generate a biased and unfaithful summary, as shown in Figure 1. Empirically, adversarial training has improved robustness to input perturbations in large models better than other proposed frameworks (Wong and Kolter, 2018; Zhang et al., 2022). Adversarial training is formulated as a min-max optimization problem that trains a model on adversarial samples generated with Projected Gradient Descent (PGD) (Madry et al., 2019). The change in input example to generate an adversarial sample is bounded by an $l - p$ norm radius to preserve the semantics of the input data.

Recent research (Štorek et al., 2025; Mehrotra et al., 2024) has used repeated black-box model querying to identify perturbations for crafting adversarial examples. However, such approaches are not directly applicable to gradient-based adversarial training due to the discrete nature of natural language. So we use the above adversarial training strategy in the latent space of Ses2Seq models to strengthen the robustness of the text generation model. With encoder and decoder architectures separated in the Seq2Seq model, we can apply the adversarial perturbations to the continuous output of the model encoder.

3.4 AdvSumm: Adversarial Summarization

Using the adversarial training strategy, we propose AdvSumm for mitigating bias in text summarization. As shown in Figure 2, there are three major components in AdvSumm. First is an encoder E , which maps the input text x into a continuous latent space providing a text embedding $E(x) = e_x$. The second component, Perturber, makes perturbations to e_x in the direction that maximally increases the loss function \mathcal{L} , thereby targeting regions of the embedding space that are most likely to degrade the model’s generation quality. The continuous text representation e_x allows us to generate an adversarial sample e'_x using the gradient-based methods such as PGD (Madry et al., 2019). The last component decoder D maps back the perturbed e'_x back to the input space. We use the Transformer (Vaswani et al., 2023) encoder and decoder architectures and optimize the cross-entropy loss $\mathcal{L}(y, D(e'_x))$, where y is the ground truth bias-free summary. This process is outlined in Algorithm 1.

We use the Fast Gradient Signed Method (FGSM) (Goodfellow et al., 2015) to build the Perturber component, which is a cheaper single-step variant of PGD, with the number of iterations $t = 1$. The perturbed embedding is generated with the following embedding update in FGSM:

$$e'_x = e_x + \epsilon \cdot \text{sgn} \left(\frac{\partial \mathcal{L}}{\partial e_x} \right) \quad (1)$$

where, $\text{sgn}(\cdot)$ represents the sign of the quantity and ϵ captures the attack strength.

Embedding e_x and model’s predicted output $\hat{y} = D(e_x)$ are generated with a forward pass of the Encoder and Decoder respectively. The model’s predicted output \hat{y} along with the ground-truth summary y are used to compute the loss $\mathcal{L}(y, \hat{y})$. The sign of the gradient of the computed loss function \mathcal{L} is then used to modify e_x to e'_x using equation 1. Similar to the Stochastic Gradient Descent parameter optimization technique, equation 1 identifies the steepest ascent of loss as a function of embedding e_x . Therefore, the Perturber modifies the embedding such that the update direction results in the largest increase in the generation loss. Since this change leads to the highest increase in loss, this adversarial embedding e'_x must lead to the worst generated summary among all the embeddings in the ϵ ball radius of e_x . In the training procedure, this perturbed embedding e'_x is then used to jointly train the Encoder and Decoder.

Dataset	Type	#Train	#Test	#Val
XSUM	News Summ	203,577	11,305	11,301
Wiki-Nationality	Nationality Hallucination	0	71,763	0
Multi-Neus	Multi-polar News Summ	2,453	307	307

Table 1: Statistics of datasets used in this work.

4 Experiments

4.1 Datasets

We evaluate AdvSumm on two existing bias summarization benchmarks: name-nationality bias (Ladhak et al., 2023) and political framing bias (Lee et al., 2022). These datasets allow us to assess the generalization capability of our method across different kinds of bias. Specifically, name-nationality bias primarily arises from hallucinated tokens—where the model incorrectly introduces demographic attributes (e.g., inferring nationality based on names)—while political framing bias involves more subtle language choices at the document level, reflecting ideological leanings. By addressing both token-level and discourse-level biases, we demonstrate the broader applicability of our approach. Dataset statistics are summarized in Table 1.

Name-Nationality Bias. For assessing name-nationality hallucination, we use the Wiki-Nationality dataset (Ladhak et al., 2023) which was constructed by altering entity names in articles to associate them with different nationalities, without changing other biographical details. This was done to assess whether models will use an incorrect/assumed nationality in the summary just based on the person’s name.

Framing Bias. We explore framing bias with the Neutral multi-news Summarization (NeuS) dataset (Lee et al., 2022), which comprises triplets of left, right, and center-slanted news articles paired with neutral summaries focused on the facts in the articles.

4.2 Metrics

Name-Nationality Bias. We calculate the hallucination rate as the proportion of articles where the model incorrectly attributes a nationality in the generated summary which is different from the nationality in the input document. The aim of our approach is to reduce the hallucination rate and hence, reduce the spurious association of names to specific nationalities.

Attack Strength	ROUGE-1	Ar_+	Ar_-	Ar_{sum}
0	44.81	2.19	1.07	3.26
10^{-3}	44.38	1.82	0.93	2.75
10^{-2}	41.07	0.55	0.29	0.84
10^{-1}	14.52	0.29	0.16	0.45

Table 2: Flan-T5 on Multi-Neus with different degrees of attack strength.

Framing Bias. Following Lee et al. (2022), we use arousal scores from the Valence-Arousal-Dominance (VAD) lexicon (Mohammad, 2018), which provides valence (v), arousal (a), and dominance (d) annotations for a list of words. The positive arousal score (Ar_+) and negative arousal score (Ar_-) are defined as the summed arousal values of words with positive and negative valence, respectively, based on the VAD annotations. The combined arousal score (Ar_{sum}) is the sum of Ar_+ and Ar_- . The goal of AdvSumm is to mitigate political framing in generated summaries by minimizing both Ar_+ and Ar_- , while preserving overall summarization quality.

Summarization Quality. We utilize ROUGE (Lin, 2004) scores to measure the summarization quality. We report ROUGE-1 in our results.

4.3 Settings

Models. We use two encoder-decoder transformer models for the text summarization task: BART-large (Lewis et al., 2019) and Flan-T5 base (Chung et al., 2022). BART is a denoising autoencoder pre-trained with a corrupted text reconstruction objective, making it well-suited for generation tasks. In contrast, Flan-T5 builds on the T5 architecture (Raffel et al., 2023) and is further instruction-tuned on a broad mixture of tasks, enabling better generalization to unseen instructions and objectives. This contrast allows us to evaluate the robustness and generalization capabilities of AdvSumm across models with different pre-training strategies. We leave it to future work to adapt our Perturber component to decoder-only LLM architectures.

For name-nationality bias, models are fine-

Model	ROUGE-1 \uparrow	American \downarrow	Asian \downarrow	African \downarrow	European \downarrow	Overall \downarrow
BART	43.45	0.84	13.41	0.92	7.55	5.61
Flan-T5	39.99	0.03	2.39	0.06	0.57	0.76
Back-Trans (BART)	41.91	1.08	8.69	1.32	4.75	3.96
AdvSumm (BART)	40.02	0.40	4.42	0.27	1.96	1.76
AdvSumm (Flan-T5)	37.86	0.02	2.33	0.16	0.38	0.72

Table 3: Hallucination rate over multiple countries in Wiki-Nationality dataset. Am represents American, Af African, As Asian and Ovr is the Hallucination rate over all countries. AdvSum improves the Hallucination rate while maintaining similar ROUGE-1 scores.

tuned on the XSUM news summarization dataset (Narayan et al., 2018) and evaluated on the Wiki-Nationality benchmark. For framing bias, we adopt the fine-tuning scheme of Lee et al. (2022) for fine-tuning on the training split of the Multi-Neus dataset. AdvSumm applies adversarial training with the perturber component during this fine-tuning stage.

Baselines. We compare AdvSumm against two baselines: (i) models fine-tuned on the same data without the perturber component, and (ii) a data augmentation using back-translation. For the latter, training data is augmented by paraphrasing input texts via back-translation from German, effectively doubling the training set size while keeping the targets unchanged (Cao and Wang, 2021). We evaluate the effectiveness of this back-translation-based generalization strategy against our adversarial generalization method (AdvSumm).

Implementation. We experiment on an NVIDIA A100 GPU with 40 GB VRAM. We finetune all models using a learning rate of $5e-5$ with AdamW optimizer and 10% warm-up steps. Maximum input length is set to 1024 for XSUM and 512 for Multi-Neus. The maximum output generation length is taken as 142 along with a beam size of 6 for Wiki-Nationality and a generation length of 250 with a beam size of 4 is used for Multi-Neus. Generation configurations (like input, output lengths, beam sizes, etc.) are adopted directly from Lee et al. (2022). Other hyperparameters like the number of epochs are tuned on validation splits. All results are reported on the test split. For the baselines, we use English-to-German and German-to-English translation models provided by Fairseq for back-translation.

We tune the attack strength of the Perturber Component by varying the value of ϵ in equation 1. A higher value of ϵ gives the Perturber higher freedom to change the embedding e_x but also leads

to a greater change in text semantics, which will lead to a drop in summarization performance. For practical implications, ϵ behaves like a “knob” for controlling the amount of bias while trading off the summarization quality. We show the tuning results of Flan-T5 on the Multi-Neus dataset in Table 2 with the ROUGE-1 and Arousal scores. We observe a drop in bias as well as summarization quality as we increase the value of attack strength. We find $\epsilon = 0.01$ to be optimal, which we use for further experiments.

5 Results

We present our empirical findings on the two types of biases in this section.

5.1 Name-Nationality Bias

The results on the Name-Nationality benchmark are illustrated in Table 3, which shows a comparative analysis between the baseline models and our proposed approach, AdvSumm, focusing on region-specific hallucination rates as discussed in Ladhak et al. (2023), as well as ROUGE-1 scores on the XSum evaluation sets to compare summarization quality.

In the Name-Nationality setting, AdvSumm significantly lowers hallucination rates in summaries across American, Asian, and European contexts compared to base models. It effectively reduces overall hallucination rates underscoring the effectiveness of AdvSumm in enhancing the fidelity of summarization models, ensuring more reliable summaries across diverse geopolitical landscapes while maintaining competitive ROUGE-1 scores.

AdvSumm’s lower ROUGE-1 scores on the test sets, as shown in Table 3, align with prior research findings that Adversarial Training, while enhancing model robustness, can reduce performance on clean data (Madry et al., 2019). This tradeoff is expected in data augmentation techniques, where the goal is to improve model resilience (reduce bias) while

Models/ Settings	Framing Bias Metrics			Salient Info
	$Ar_+ \downarrow$	$Ar_- \downarrow$	$Ar_{sum} \downarrow$	ROUGE-1 \uparrow
BART	1.33	0.76	2.09	45.94
Flan-T5	2.19	1.07	3.26	44.81
Back-Trans	1.40	0.77	2.17	46.51
AdvSum(BART)	0.59	0.33	0.92	43.01
AdvSum(Flan-T5)	0.55	0.29	0.84	41.07

Table 4: Results of AdvSum on of Multi-Neus dataset. An attack strength of 0.01 is used for AdvSum. BART-large is used for training on Back-Translated data. Ar stands for Arousal.

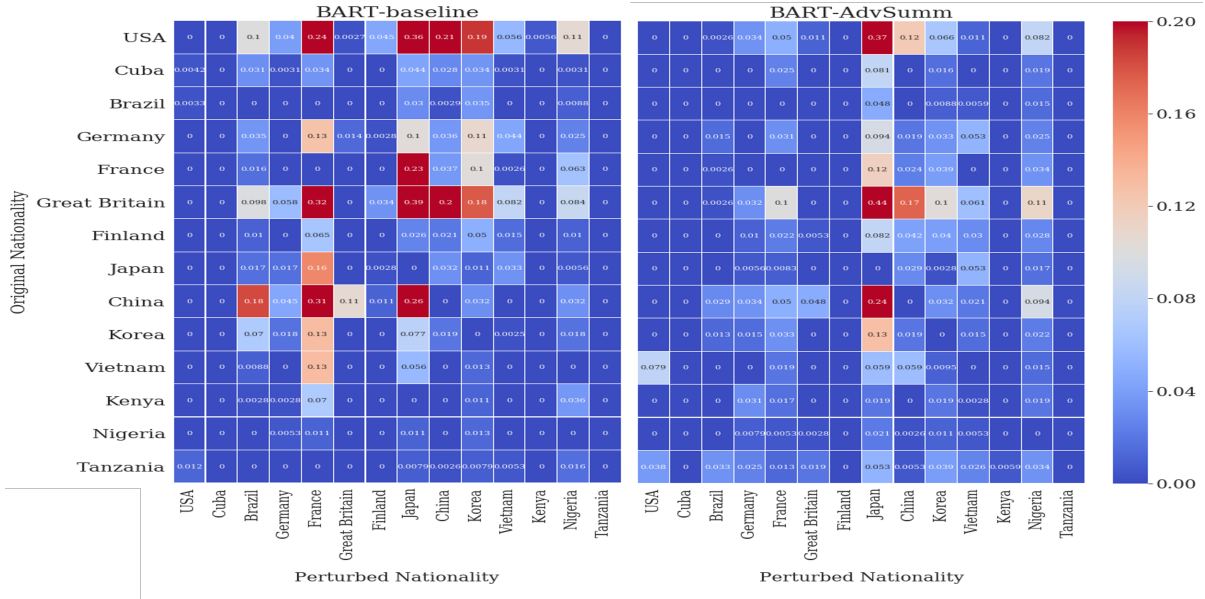


Figure 3: Hallucination rate for BART baseline and one trained using AdvSum. Red corresponds to higher, and Blue corresponds to lower hallucination rate.

minimizing the performance drop on clean datasets.

5.2 Framing Bias

Evaluations on the Multi-Neus benchmarks are outlined in Table 4. We report the Framing Bias Metric consisting of positive Arousal Score A_+ , negative Arousal Score A_- , and their sum A_{sum} . We also report ROUGE-1 for capturing the summarization quality by each setting. We also show the bias evaluations of the back-translation-based data-augmentation approach in Table 4.

On the Multi-Neus dataset, we see the least biased summaries in the case of AdvSumm on Flan-T5, with the lowest positive and negative Arousal scores. Both generalization approaches (ours and back-translation) outperform Flan-T5, which supports the hypothesis on bias mitigation with robustness. AdvSumm surpasses the back-translation approach by 1.4 absolute points on Ar_{sum} , while taking a slight dip in ROUGE-1 scores. We also note that our end-to-end adversarial training approach is more computationally efficient than back-

translation, given the time taken by dual-translation and the double training steps of summarization fine-tuning.

We also observe consistently lower bias scores across both benchmarks when using the Flan-T5 architecture. Flan-T5 benefits from instruction tuning on a diverse set of tasks, including ethical reasoning and instruction following. We hypothesize that this additional tuning phase not only enhances zero-shot generalization but also better aligns the model with human expectations, helping it avoid spurious biases inherited from pre-training.

5.3 Error Analysis

For name-nationality bias, we report a heatmap as shown in Figure 3 where the hallucination rate for all combinations of countries is calculated. In alignment with the numerical results, hallucination rates for Asian countries as perturbed nationalities are significantly higher for the Bart baseline than our approach AdvSumm. We notice that, however, for a few combinations like Great Britain-Japan,

AdvSum	Texas Church Shooting: A gunman opened fire at a church in Texas on Sunday, killing two people and wounding three others .
Source News	Shooting at Texas Church Leaves 2 Parishioners Dead, Officials Say: A gunman opened fire at a church in Texas on Sunday morning, killing two people with a shotgun before a member of the church’s volunteer security team fatally shot him, the authorities said. About 250 people were inside the auditorium of the West Freeway Church of Christ in White Settlement, near Fort Worth, when the gunman began shooting just before communion, said Jack Cummings, a minister at the church. Mr. Cummings said the gunman was “acting suspiciously” before the shooting and drew the attention of the church’s security team.

Table 5: An example of positive arousal generated news. AdvSum hallucinates the text in red color. Each of the three examples contains <Title>:<Article>. The Center news article is shown in Source News.

Generated News	Trump to End DACA: President Trump will announce on Tuesday that he is ending a controversial program that protects nearly 800,000 young undocumented immigrants from deportation, media reports indicated late Sunday.
Neutral News	Reports Say DACA Is Over: President Trump will announce on Tuesday that he is ending a controversial program that protects nearly 800,000 young undocumented immigrants from deportation, media reports indicated late Sunday.

Table 6: A generated news summary compared to neutral news. Each example contains <Title>:<Article>

Vietnam-USA, there is a slight increase in the hallucination rate.

For framing bias, most of the biased generation is still a result of model hallucination. The example shown in Table 5 shows the text in red color, which is hallucinated by the model. The "wounding of three others" is not mentioned in the source article. Additionally, current Framing bias metrics fail to capture the context around lexicons. An example is shown in Table 6, where the positive arousal score given by the Lexicon-based metric is zero, which is clearly wrong, looking at the title of the generated news.

6 Conclusions

In this work, we introduced AdvSumm, a domain-agnostic adversarial training framework for bias mitigation in text summarization. Motivated by the limitations of existing bias mitigation strategies—particularly their domain-specific nature and difficulty generalizing across different types of biases—we reformulated bias reduction as a generalization problem, tackled through adversarial robustness. By introducing the Perturber mod-

ule to apply embedding-level adversarial perturbations during fine-tuning, we demonstrated that AdvSumm effectively reduces both token-level biases (e.g., name-nationality associations) and document-level biases (e.g., political framing) without compromising summarization quality. Empirical results on benchmark datasets highlight that AdvSumm outperforms standard transformers and back-translation baselines, offering a unified and scalable solution for fairer text generation.

Limitations

Our study focuses on bias mitigation in text summarization using encoder-decoder transformer architectures. However, many recent summarization systems adopt decoder-only architectures, where directly applying the Perturber component in its current form is not straightforward. Future work could explore extending adversarial perturbations to individual layers of the transformer decoder, enabling the approach to generalize to decoder-only models as well.

Ethics Statement

We conduct our evaluations using publicly available datasets that do not contain personally sensitive information or toxic content. One important ethical consideration is that developing robust summarization systems, as proposed in this paper, contributes to ongoing efforts to reduce harmful biases in natural language generation systems by mitigating biases inherited from pre-training data. For example, prior work has shown that biased news framing can contribute to political polarization (Han and Federico, 2017), and name-nationality associations can reinforce harmful stereotypes in text generation (Ladhak et al., 2023).

By improving the robustness of summarization models, our approach takes a step toward address-

ing these issues. However, we acknowledge that our work does not evaluate all forms of bias that may arise in text summarization tasks, nor does it fully evaluate potential side effects of the approach, such as its impact on other aspects of faithfulness or other types of bias in summarization. Future research should explore these broader impacts to ensure that summarization systems are both fair and faithful across different contexts and biases.

Acknowledgments

One of the authors is supported by the National Science Foundation Graduate Research Fellowship DGE-2036197, the Columbia University Provost Diversity Fellowship, and the Columbia School of Engineering and Applied Sciences Presidential Fellowship. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. Another author is supported by Amazon and Columbia’s Center of Artificial Intelligence Technology (CAIT) PhD student fellowship.

References

- Robert Adragna, Elliot Creager, David Madras, and Richard Zemel. 2020. Fairness and robustness in invariant learning: A case study in toxicity classification. *arXiv preprint arXiv:2011.06485*.
- Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2020. [Invariant risk minimization](#).
- Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. 2009. *Robust optimization*, volume 28. Princeton university press.
- Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.
- Toon Calders and Indrė Žliobaitė. 2013. Why unbiased computational processes can lead to discriminative decision procedures. In *Discrimination and Privacy in the Information Society: Data mining and profiling in large databases*, pages 43–57. Springer.
- Flavio P. Calmon, Dennis Wei, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. 2017. [Optimized data pre-processing for discrimination prevention](#).
- Shuyang Cao and Lu Wang. 2021. [Cliff: Contrastive learning for improving faithfulness and factuality in abstractive summarization](#).
- 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](#).
- Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. [BOLD](#). In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM.
- Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. 2020. [Multi-dimensional gender bias classification](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 314–331, Online. Association for Computational Linguistics.
- Michele Donini, Luca Oneto, Shai Ben-David, John Shawe-Taylor, and Massimiliano Pontil. 2020. [Empirical risk minimization under fairness constraints](#).
- Ragnhild Eg, Özlem Demirkol Tønnesen, and Merete Kolberg Tennfjord. 2023. [A scoping review of personalized user experiences on social media: The interplay between algorithms and human factors](#). *Computers in Human Behavior Reports*, 9:100253.
- Virginia K. Felkner, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May. 2023. [Winoqueer: A community-in-the-loop benchmark for anti-lgbtq+ bias in large language models](#).
- Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the majority is wrong: Leveraging annotator disagreement for subjective tasks. *arXiv preprint arXiv:2305.06626*.
- Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. [Domain-adversarial training of neural networks](#).
- Aparna Garimella, Akhash Amarnath, Kiran Kumar, Akash Pramod Yalla, N Anandhavelu, Niyati Chhaya, and Balaji Vasan Srinivasan. 2021. He is very intelligent, she is very beautiful? on mitigating social biases in language modelling and generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4534–4545.
- Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. [Explaining and harnessing adversarial examples](#).

- Jiyoung Han and Christopher M. Federico. 2017. [Conflict-framed news, self-categorization, and partisan polarization](#). *Mass Communication and Society*, 20(4):455–480.
- Dirk Hovy and Shrimai Prabhume. 2021. Five sources of bias in natural language processing. *Language and linguistics compass*, 15(8):e12432.
- Maximilian Kaufmann, Yiren Zhao, Iliia Shumailov, Robert Mullins, and Nicolas Papernot. 2022. [Efficient adversarial training with data pruning](#).
- Faisal Ladhak, Esin Durmus, He He, Claire Cardie, and Kathleen McKeown. 2022. [Faithful or extractive? on mitigating the faithfulness-abstractiveness trade-off in abstractive summarization](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1410–1421, Dublin, Ireland. Association for Computational Linguistics.
- Faisal Ladhak, Esin Durmus, Mirac Suzgun, Tianyi Zhang, Dan Jurafsky, Kathleen McKeown, and Tatsunori Hashimoto. 2023. [When do pre-training biases propagate to downstream tasks? a case study in text summarization](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3206–3219, Dubrovnik, Croatia. Association for Computational Linguistics.
- Savannah Larimore, Ian Kennedy, Breon Haskett, and Alina Arseniev-Koehler. 2021. [Reconsidering annotator disagreement about racist language: Noise or signal?](#) In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 81–90, Online. Association for Computational Linguistics.
- Nayeon Lee, Yejin Bang, Tiezheng Yu, Andrea Madotto, and Pascale Fung. 2022. [NeuS: Neutral multi-news summarization for mitigating framing bias](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3131–3148, Seattle, United States. Association for Computational Linguistics.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. [Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#).
- Ya Li, Mingming Gong, Xinmei Tian, Tongliang Liu, and Dacheng Tao. 2018. [Domain generalization via conditional invariant representation](#).
- Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekogul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. [Holistic evaluation of language models](#).
- 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.
- Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2019. [Towards deep learning models resistant to adversarial attacks](#).
- Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. [On faithfulness and factuality in abstractive summarization](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.
- Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum Anderson, Yaron Singer, and Amin Karbasi. 2024. [Tree of attacks: Jailbreaking black-box llms automatically](#).
- Saif Mohammad. 2018. [Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.
- Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. [StereoSet: Measuring stereotypical bias in pretrained language models](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.
- Vedant Nanda, Samuel Dooley, Sahil Singla, Soheil Feizi, and John P. Dickerson. 2021. [Fairness through robustness: Investigating robustness disparity in deep learning](#). In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT ’21*, page 466–477, New York, NY, USA. Association for Computing Machinery.
- Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. [CrowS-pairs: A challenge dataset for measuring social biases in masked language models](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

- Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. [Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization.](#)
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. [Exploring the limits of transfer learning with a unified text-to-text transformer.](#)
- Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. [Gender bias in coreference resolution.](#) In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.
- Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. [Annotators with attitudes: How annotator beliefs and identities bias toxic language detection.](#) In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.
- Julius Steen and Katja Markert. 2024. [Bias in news summarization: Measures, pitfalls and corpora.](#)
- Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*.
- Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. [Intriguing properties of neural networks.](#)
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. [Attention is all you need.](#)
- Martin Wessel, Tomas Horych, Terry Ruas, Akiko Aizawa, Bela Gipp, and Timo Spinde. 2023. [Introducing MBIB - the first media bias identification benchmark task and dataset collection.](#) In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM.
- Eric Wong and J. Zico Kolter. 2018. [Provable defenses against adversarial examples via the convex outer adversarial polytope.](#)
- Wenhao Wu, Wei Li, Jiachen Liu, Xinyan Xiao, Ziqiang Cao, Sujian Li, and Hua Wu. 2022a. [Frsum: Towards faithful abstractive summarization via enhancing factual robustness.](#) *arXiv preprint arXiv:2211.00294*.
- Wenhao Wu, Wei Li, Jiachen Liu, Xinyan Xiao, Sujian Li, and Yajuan Lyu. 2022b. [Precisely the point: Adversarial augmentations for faithful and informative text generation.](#) *arXiv preprint arXiv:2210.12367*.
- Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33:6256–6268.
- Yue Xing, Qifan Song, and Guang Cheng. 2021. On the generalization properties of adversarial training. In *International Conference on Artificial Intelligence and Statistics*, pages 505–513. PMLR.
- Mingyang Yi, Lu Hou, Jiacheng Sun, Lifeng Shang, Xin Jiang, Qun Liu, and Zhi-Ming Ma. 2021. [Improved ood generalization via adversarial training and pre-training.](#)
- Huan Zhang, Shiqi Wang, Kaidi Xu, Yihan Wang, Suman Jana, Cho-Jui Hsieh, and Zico Kolter. 2022. [A branch and bound framework for stronger adversarial attacks of ReLU networks.](#) In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 26591–26604. PMLR.
- Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.
- Adam Štorek, Mukur Gupta, Noopur Bhatt, Aditya Gupta, Janie Kim, Prashast Srivastava, and Suman Jana. 2025. [Xoxo: Stealthy cross-origin context poisoning attacks against ai coding assistants.](#)

NSF-SciFY: Mining the NSF Awards Database for Scientific Claims

Delip Rao^{*†}, Weiqiu You[†], Eric Wong, Chris Callison-Burch

University of Pennsylvania

Philadelphia, PA, USA

{delip, weiqiuy, exwong, ccb}@seas.upenn.edu

Abstract

We introduce NSF-SciFY, a comprehensive dataset of scientific claims and investigation proposals extracted from National Science Foundation award abstracts. While previous scientific claim verification datasets have been limited in size and scope, NSF-SciFY represents a significant advance with 2.8 million claims from 400,000 abstracts spanning all science and mathematics disciplines. We present two focused subsets: NSF-SciFY-MATSCI with 114,000 claims from materials science awards, and NSF-SciFY-20K with 135,000 claims across five NSF directorates. Using zero-shot prompting, we develop a scalable approach for joint extraction of scientific claims and investigation proposals. We demonstrate the dataset’s utility through three downstream tasks: non-technical abstract generation, claim extraction, and investigation proposal extraction. Fine-tuning language models on our dataset yields substantial improvements, with relative gains often exceeding 100%, particularly for claim and proposal extraction tasks. Our error analysis reveals that extracted claims exhibit high precision but lower recall, suggesting opportunities for further methodological refinement. NSF-SciFY enables new research directions in large-scale claim verification, scientific discovery tracking, and meta-scientific analysis.

1 Introduction

The overall growth rate of scientific publications is estimated to be 4% annually, with a doubling time of 17 years (Bornmann et al., 2021). Within this deluge, researchers, reviewers, and the general public struggle to separate substantiated claims from spurious ones—whether it is the “quantum supremacy” assertions in computing, the short-lived excitement over LK-99 superconductors³,

^{*}Corresponding author, [†]co-first author

³for an entertaining digression c.f., <https://en.wikipedia.org/wiki/LK-99>

```
{
  "award_id": 2321365,
  "title": "Electrically Conductive 2D Metal-Organic Frameworks and Cov...",
  "technical_abstract": "Owing to their diverse potentials to serve as e...",
  "non_technical_abstract": "Sustaining the rapid advances of modern ele...",
  "verifiable_claims": [
    "MOFs and COFs have synthetic accessibility, structural modularity,",
    "Electrical conductivity remains one of the most elusive traits of M...",
    "In 2D MOFs, electronic conduction can occur within the planes thro...",
    "In 2D COFs, pi-stacked layers represent the primary transport path...",
    ...
  ],
  "investigation_proposals": [
    "Develop and implement a new design strategy to promote long-range c...",
    "Incorporate built-in alternating pi-donor/acceptor stacks inside c...",
    "Investigate how pi-donor/acceptor stacks consisting of different c...",
    "Create a new design strategy for next-generation electrically cond...",
    "Produce novel electrically conductive 2D MOFs and COFs with unique...",
    ...
  ],
  "publications": [
    {
      "doi": "10.1021/acs.inorgchem.3c02647",
      "status": "resolved",
      "title": "From a Collapse-Prone, Insulating Ni-MOF-74 Analogue to...",
      "abstract": "Electrically conductive porous metal-organic framewo..."
    }
    ...
  ]
}
```

Figure 1: A sample record from our dataset. Each record contains 1) Award ID and title, 2) NSF Directorate, 3) Technical and non-technical abstracts, 4) Scientific Claims, 5) Investigation Proposals, and 6) Associated publications, when present.

or the misunderstanding surrounding microplastic leaches from black plastic spatulas⁴. Manual verification of ever growing body of scientific claims has become intractable, yet the economic and societal consequences of unverified claims are increasingly severe.

Wadden et al. (2020) introduced the task of scientific claim verification with the SciFACT dataset, focusing primarily on automatic verification of scientific claims. Follow up works (see Section 2 for a detailed account) have mostly focused on the healthcare, building datasets from scientific publications, and modest-sized dataset creation. In this work, we relax all of these aspects and look at building at least an order of magnitude large-scale

⁴c.f., <https://nationalpost.com/news/canada/black-plastic>

Dataset	# claims	# docs	Evidence Source	Domain
SciFACT (Wadden et al., 2020)	1.4K	5K	Research papers	Biomedical
PubHEALTH (Kotonya and Toni, 2020)	11.8K	11.8K	Fact-checking sites	Public health
CLIMATE-FEVER (Diggelmann et al., 2020)	1.5K	7.5K	Wikipedia articles	Climate change
HealthVer (Sarrouiti et al., 2021)	1.8K	738	Research papers	Healthcare
COVID-Fact (Saakyan et al., 2021)	4K	4K	Research, news	COVID
CoVERT (Mohr et al., 2022)	300	300	Research, news	Biomedical
SciFACT-Open (Wadden et al., 2022)	279	500K	Research papers	Biomedical
NSF-SciFY-MATSCI (ours)	114K	16K	NSF award abstracts	Material Science
NSF-SciFY-20K (ours)	135K	20K	NSF award abstracts	All Science & Math
NSF-SciFY (ours)	2.8M	400K	NSF award abstracts	All Science & Math

Table 1: This table comparison clearly illustrates the scale advantage of NSF-SciFy over existing scientific claim verification datasets. While previous datasets like SciFACT and PubHEALTH contain at most thousands of claims from published research papers or fact-checking sources, our NSF-SciFY-MATSCI and NSF-SciFY-20K datasets individually contribute more than 100K claims. The full NSF-SciFY dataset represents an order-of-magnitude increase with 2.8M claims across 400K abstracts spanning all science & math disciplines. This work introduces grant abstracts as a novel, untapped source for scientific claim extraction, complementing existing approaches that focus on published literature, news articles, or social media content.

scientific claim dataset covering all of basic science. We envision building of such large-scale, scientific claim datasets to help future work on robust scientific claim verification systems.

We introduce NSF-SciFY¹, a comprehensive dataset of claims and investigation proposals extracted from National Science Foundation (NSF) award abstracts. We choose NSF abstracts as our source material for several reasons:

1. NSF is a primary driver of U.S. scientific innovation, funding approximately 25% of all federally supported basic research, spanning the entirety of science and math areas, with an annual budget of \$9.9 billion (FY 2023). Any claim dataset derived from the NSF awards database should faithfully represent the scientific Zeitgeist.
2. NSF’s rigorous subject matter expert-review process provides an high-quality filter for the claims made in funded proposals.
3. The public availability and permissive usage terms of the NSF awards database makes it an excellent resource for open science research.
4. Previous datasets on scientific claims have been derived from scientific papers, but claims in scientific grants, and particularly investigation proposals, remain unstudied.

While not this focus of this paper, grant award abstracts, additionally, provide a unique opportunity to study the relationship between what researchers claim and what they propose to investigate. This could offer valuable insights into scientific practice

¹Short for “NSF SCientific FeasibilitY”.

and the evolution of research questions.

In this paper, we make the following contributions: (1) We introduce NSF-SciFY, the largest scientific claim dataset to date with 2.8M claims extracted from 400K NSF award abstracts, establishing grant proposals as a novel source for scientific claim extraction; (2) We create NSF-SciFY-MATSCI focusing exclusively on materials science with 114K extracted claims from 16K abstracts. This is the first materials science claim dataset and, in number of extracted claims, this alone is an order of magnitude bigger than the largest publicly available claim dataset; In addition, we also create NSF-SciFY-20K with 135K claims spanning five NSF directorates. (3) We develop a zero-shot prompting approach for joint extraction of scientific claims and investigation proposals as a scalable way to bootstrap high-precision, large-scale scientific claim datasets; (4) We present novel evaluation metrics for claim/proposal extraction based on LLM judgments, showing that fine-tuned models significantly outperform base models; and (5) Finally, we release all datasets and trained models from our work for unfettered research and commercial use. Our dataset and methods enable new opportunities for large-scale claim verification, scientific discovery tracking, and meta-scientific research.

2 Related Work

Scientific claim extraction and verification has emerged as an important research area as the volume of scientific literature continues to grow expo-

nentially. Previous work has primarily focused on claims from published papers, fact-checking sites, and news articles.

Scientific Claim Datasets Several datasets have been developed for scientific claim verification, but all have focused on claims from published literature, while we undertake the study of grant award abstracts. SciFACT (Wadden et al., 2020) contains 1,400 scientific claims derived from research papers in the biomedical domain. PubHEALTH (Kotonya and Toni, 2020) includes 11,800 claims from journalists and fact-checkers in public health. CLIMATE-FEVER (Diggelmann et al., 2020) compiled 1,500 claims from news articles about climate change. HealthVer (Sarrouti et al., 2021) extracted 1,800 claims from search queries related to health topics. COVID-Fact (Saakyan et al., 2021) and CoVERT (Mohr et al., 2022) focused on COVID-19 related claims from social media. SciFact-Open (Wadden et al., 2022) expanded the original SciFact dataset using information retrieval pooling, yet it still remains health-care focused and a few orders of magnitude smaller than our largest dataset.

Table 1 situates existing scientific claim datasets with our NSF-SCIFY datasets, highlighting the significantly larger scale of our contribution (2.8 million claims in NSF-SCIFY, 135,000 claims in NSF-SCIFY-20K and 114,000 claims in NSF-SCIFY-MATSCI), broad topic coverage (all of science and math), and novelty of data source (grant abstracts). See Figure 2.

Meta Science and Social Science Previous works have examined grants data in social science and meta-science contexts. For example, Park et al. (2024) examine the relationship between interdisciplinary grants and the impact of papers they support and Xu et al. (2022) study the influence of research funding on team structure using grant data. While these are tenuously connected to our work, we list them for the sake of completeness.

3 Building NSF-SCIFY

3.1 Data Collection

We downloaded the entire NSF Awards database² in XML format, containing more than 0.5 million awards from 1970 through September 2024. After parsing, we obtained 412,155 parseable awards, which we call NSF-SCIFY.

²<https://www.nsf.gov/awardsearch/advancedSearch.jsp>

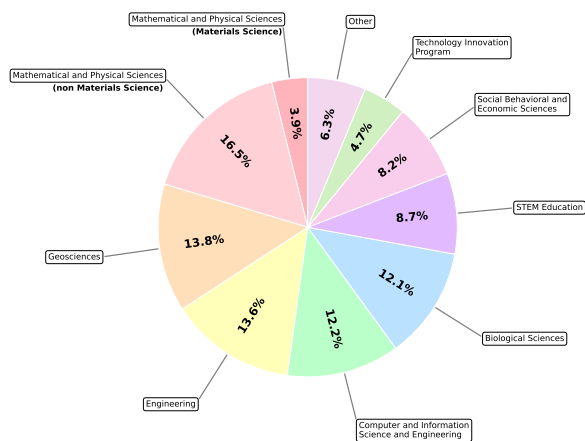


Figure 2: Distribution of awards areas as represented by the National Science Foundation directorates in NSF-SCIFY, illustrating the breadth and comprehensiveness of scientific claims in our dataset. The NSF-SCIFY-MATSCI subset spanning all of materials science awards represents 3.9% of the entire dataset.

In this paper, we focus on all awards from the Division of Materials Research (DMR), which is responsible for most materials science awards at the NSF. This subset, called NSF-SCIFY-MATSCI, contains 16,031 awards, representing approximately 3.2% of the entire NSF awards database. We chose materials science as our focus due to its interdisciplinary nature and technological importance. In addition, we build NSF-SCIFY-20K, a different subset of 20K awards spanning 5 NSF directorates — Mathematical and Physical Sciences (MPS), Geological Sciences (GEO), Engineering (ENG), Computer and Information Science and Engineering (CSE), and Biological Sciences (BIO).

3.2 Data Processing

As Figure 1 illustrates, each record in NSF-SCIFY-MATSCI typically contains:

1. Award ID, title, and year.
2. Directorate and division information
3. Technical abstract
4. Non-technical abstract (present in $\sim 81\%$ of awards)
5. Scientific claims made in the abstracts
6. Investigation proposals in the abstracts
7. Publications resulting from the grant (when available)

The practice of updating awards with resulting publications is relatively recent, primarily occurring from 2014 onwards. For awards where publications are present, we extracted the DOIs and

resolved them to obtain titles, abstracts, and publication URLs.

3.3 Claim and Investigation Proposal Extraction

To extract scientific claims and investigation proposals from the award abstracts, we developed a zero-shot prompting approach using Anthropic’s Claude-3.5³ model. Our prompt instructed the model to identify two types of statements:

1. **Verifiable claims:** Statements that the abstract claims to be true or states as assumptions, either explicitly or implicitly.
2. **Investigation proposals:** Forward-looking statements that propose specific research activities as part of the award.

We structured the prompt to return a JSON object containing the award ID, technical abstract, non-technical abstract, a list of verifiable claims, and a list of investigation proposals. To maintain consistency and quality, we set temperature to zero for all extractions. See Appendix A for the exact prompt and Appendix G for sample claims and investigation proposals.

We performed qualitative experiments with several prompt variants and our analysis showed that jointly extracting claims and investigation proposals helped maintain the relevance of extracted claims. When claims were extracted without also extracting investigation proposals, the model often confused forward-looking statements about proposed investigations as factual claims.

4 Dataset Analysis

NSF-SciFY The full dataset contains 412,155 award abstracts spanning from 1970 to 2024, with 2.8 million scientific claims and corresponding investigation proposals.

NSF-SciFY-MATSCI This materials science subset, which is the focus of this preprint, contains:

- 16,042 awards with each with a technical and non-technical abstract
- 114K extracted scientific claims (average of 7 ± 2 claims per abstract-pair)
- 145K extracted investigation proposals (average of 9 ± 3 proposals per abstract-pair)
- 2,953 awards with linked publications (18.4% of the dataset). Such awards had anywhere between 1 – 4 publications.

³Claude-3.5-Sonnet-20240620 accessed between Sep-Oct. 2024, to be specific

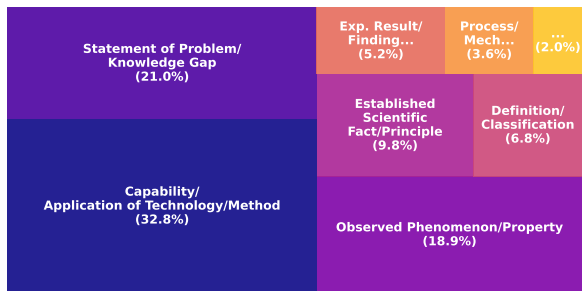


Figure 3: A treemap of the scientific claim categories in NSF awards. See Section 4.2 for descriptions of these categories.

NSF-SciFY-20K For building models across all NSF directorates, we take 20,000 sample subset of NSF-SciFY, by stratifying across 5 directorates.

- 20,001 awards with each with a technical and non-technical abstract
- 135K extracted scientific claims (average of 7 ± 2 claims per abstract-pair)
- 139K extracted investigation proposals (average of 7 ± 2 proposals per abstract-pair)

4.1 Technical vs. Non-Technical Abstracts

We investigated the differences between technical and non-technical abstracts in our dataset. Using a symmetric BLEU score to measure textual similarity between paired abstracts, we found that only 202 (1.5%) out of 13,025 technical/non-technical abstract pairs had a similarity score greater than 0.6, suggesting that the non-technical abstracts are not simply copied from the technical abstracts.

Since grant abstracts are previously unexamined in literature, we further investigated the stylistic differences between technical and non-technical abstracts using pre-trained document embedding models. Figure 5 compares content embeddings from SPECTER (Cohan et al., 2020) and style embeddings from STEL (Patel et al., 2025). Using these embeddings with a linear SVM classifier, we achieved F1 scores of 90.99 (SPECTER), 88.42 (STEL), and 89.99 (concatenated), demonstrating that the abstracts are distinguishable both in content and style.

4.2 A Taxonomy of Scientific Claims in NSF Award Abstracts

To understand the nature of assertions made in NSF award abstracts, we analyzed and grouped 810 extracted claims over a stratified sample of 120 awards spanning 5 NSF directorates – MPS, GEO, ENG, CSE, and BIO – resulting in the identifica-

tion of eight distinct categories. These categories encompass a range of scientific statements: **Established Scientific Fact/Principle** refers to claims that state well-accepted scientific laws or widely known facts. **Observed Phenomenon/Property** includes claims describing natural phenomena or material properties that have been observed. Claims detailing the function, potential, or application of a specific technology, method, or material fall under **Capability/Application of Technology/Method**. **Hypothesis/Theoretical Prediction** categorizes claims that propose theoretical explanations, predictions, or hypotheses requiring further verification. Specific results, findings, or statements about measurability are grouped under **Experimental Result/Finding/Measurability**. **Statement of Problem/Knowledge Gap** highlights existing problems, limitations, or gaps in current knowledge or technology. Claims that define terms or classify entities are categorized as **Definition/Classification**. Finally, **Process/Mechanism Description** includes claims that describe physical or chemical processes or mechanisms. Figure 3 shows the distribution of these claim types, and Table 10 lists examples for each of these types. As one might expect in grant applications, the most common claim type is “Capability/Application of Technology/Method” (32.8%), followed by “Statement of Problem/Knowledge Gap” (21.0%) and “Observed Phenomenon/Property” (18.9%).

4.3 Evaluating the Extracted Verifiable Claims

To evaluate how good the claim extraction from Section 3.3 is, we annotated the abstracts in the 120 sampled awards (see Section 4.2) with claims and computed precision and recall. Figure 4 summarizes the precision, recall, and F1-scores of the extracted claims for all 6 areas of interest. The areas of interest include, Materials Science (DMR) Mathematical and Physical Sciences except Materials Science (MPS-DMR), Geological Sciences (GEO), Engineering (ENG), Computer and Information Science and Engineering (CSE), and Biological Sciences (BIO). We notice from Figure 4 a near perfect precision for the extracted claims in all areas. Although our LLM-derived dataset suffers from recall, we hope that training on a massive corpus of high quality claims will result in useful claim extraction systems. We consider our findings on claim extraction as directional and skip a similar evaluation for investigation proposals.

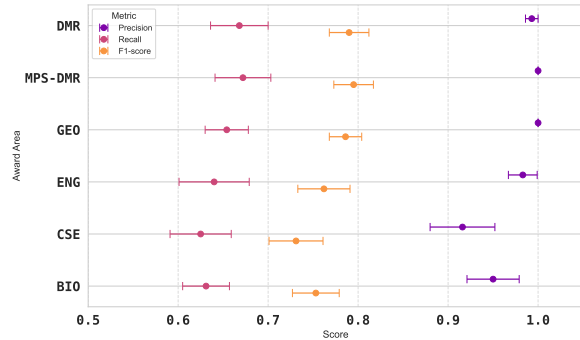


Figure 4: A Cleveland dot plot of precision, recall, and f1-score across different NSF Award Areas for claims extracted via Claude (See Section 3.3). Error bars denote standard deviation (bootstrap N=1000). See Section 4.3 for analysis.

5 Tasks, Metrics, and Experiments

We designed three NLP tasks to evaluate the utility of our dataset:

1. The **Non-technical Abstract Generation** task translates dense, technical grant abstracts into accessible language for broader science communication. Motivated by capturing the core scientific essence while navigating stylistic and content differences between technical and lay summaries, this task uses the dataset’s paired examples (common in NSF awards) to train models for this nuanced transformation.
2. The **Abstract to Scientific Claims Extraction** task automates identifying verifiable assertions—the core of scientific discourse—from grant abstracts, which capture these claims at an early, pre-publication stage. Significant performance gains post-fine-tuning highlight the dataset’s effectiveness in teaching models to pinpoint these crucial statements.
3. The **Abstract to Investigation Proposals Extraction** task distinguishes aspirational research intentions from established claims, offering a novel analysis of scientific texts. This provides a clearer view of the planned research trajectory by identifying intended activities. It complements claim extraction by presenting a fuller picture of proposed work, from assertions to investigative pathways, again showing significant fine-tuning efficacy due to the dataset’s focused nature.

To explore the three tasks, we finetuned two 7B parameter language models:

- Mistral-7B-instruct-v0.3 (Jiang et al., 2023)

Metric	Mistral	Qwen
BERTScore-P	0.8563 (+0.38% ↑)	0.8459 (+0.98% ↑)
BERTScore-R	0.8555 (+0.30% ↑)	0.8597 (+1.61% ↑)
BERTScore-F1	0.8561 (+0.36% ↑)	0.8437 (+0.75% ↑)
ROUGE1	0.2000 (+2.58% ↑)	0.1978 (+1.98% ↑)
ROUGE2	0.0198 (+4.76% ↑)	0.0210 (+3.89% ↑)
ROUGE-L	0.1273 (+2.96% ↑)	0.1466 (+0.65% ↑)
ROUGE-L-sum	0.2166 (+2.45% ↑)	0.2078 (+1.66% ↑)

Table 2: Finetuning performance for Mistral-7B-instruct-v0.3 and Qwen2.5-7B-Instruct models for Technical abstract to Non-technical abstract translation (Task 1), with relative improvements over the corresponding unfinetuned model indicated in green. Error bars for all metrics at 95% confidence intervals range between 0.0000–0.0025. Mistral model outperforms Qwen on almost all metrics for this task regardless of finetuning. We note that the percent changes, while statistically significant, are modest for this task indicating an excellent out-of-the-box performance that’s challenging to improve upon.

- Qwen2.5-7B-Instruct (Yang et al., 2024)

Details about the finetuning can be found in Appendix E.

5.1 Evaluation Metrics

For Task 1 – abstract generation – we employed a comprehensive evaluation framework using both BERTScore (Zhang* et al., 2020) and ROUGE (Lin, 2004) metrics to assess the quality of generated non-technical abstracts. This combination enables us to capture both lexical overlap and structural similarity through the ROUGE variants, while BERTScore provides insights into semantic alignment between the generated texts and reference abstracts. Incorporating such multi-viewed metrics⁴ ensures that the evaluation reflects not only the presence of key words and phrases but also the underlying meaning and narrative coherence of the abstracts.

For Task 2 – claim extraction – we developed a novel evaluation approach using LLM-based comparisons. Previous methods for claim evaluations focused on comparing a single claim against a single document. See Tang et al. (2024), for example. However, our setting required evaluating a set of extracted claims against a gold set of claims.

Towards that end, we defined a boolean function Φ_{claim} using GPT-4o-mini with zero-shot prompt-

⁴For BERTScore we report precision, recall and F1, and for ROUGE we report ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-L-sum.

ing to determine whether a generated claim is supported by a gold standard claim. See Appendix B for prompt details⁵. Using this function, we calculated precision and recall as follows:

$$\text{Precision} = \frac{1}{|S|} \sum_{c \in S} \max_{g \in G} \Phi_{\text{claim}}(c, g)$$

$$\text{Recall} = \frac{1}{|G|} \sum_{g \in G} \max_{c \in S} \Phi_{\text{claim}}(g, c)$$

where S is the set of claims generated from the finetuned model, after removal of any repeats/near-repeats⁶, and G is the gold standard set. We note that this is a variant of precision/recall metrics defined for image captioning in (Deitke et al., 2024), however unlike Deitke et al., we explicitly use Φ_{claim} in computing both precision and recall. This is necessary as we need to accurately penalize any spurious claims generated by the finetuned model. Works by (Gu et al., 2025; Liu et al., 2023) are relevant here.

Analogously, for Task 3 – extraction of investigation proposals – we define precision and recall similarly but use a different pairwise boolean judge function Φ_{IP} *mutatis mutandis*. See Appendix C for prompt details.

6 Results

6.1 Non-technical Abstract Generation

Table 2 shows the results for Task 1. Both Mistral and Qwen models demonstrated strong performance, with fine-tuning providing modest improvements. The Mistral model outperformed Qwen on almost all metrics, achieving a BERTScore-F1 of 0.8561 after fine-tuning (+0.36% relative improvement). ROUGE scores were generally low (0.01–0.22), reflecting the stylistic differences between technical and non-technical abstracts.

6.2 Scientific Claim Extraction

For Task 2 (claim extraction), fine-tuning yielded substantial improvements. As shown in Table 3, the fine-tuned Mistral model achieved a precision of 0.7450 (+116.7% relative improvement), recall of 0.7098 (+59.5%), and F1 of 0.7097 (+101.8%). The Mistral model consistently outperformed Qwen, though both showed significant benefits from fine-tuning.

⁵We tried several slight edits of the prompts and found them to be robust to such changes.

⁶We determine repeats and near-repeats in the generation by thresholding cosine similarity calculated over a TF-IDF representation of the generated claims.

Metric	Mistral	Qwen
Precision	0.7450 (+116.7% ↑)	0.6839 (+107.1% ↑)
Recall	0.7098 (+59.5% ↑)	0.6611 (+7.8% ↑)
F1	0.7097 (+101.8% ↑)	0.6541 (+63.3% ↑)

Table 3: Finetuning performance for Mistral-7B-instruct-v0.3 and Qwen2.5-7B-Instruct models for Claim Extraction from abstracts (Task 2), with relative improvements over the corresponding unfinetuned model indicated in green. Error bars for all metrics at 95% confidence intervals range between 0.0038–0.0055. Mistral model outperforms Qwen on almost all metrics for this task regardless of finetuning. We note the large positive percent changes, sometimes improvements as large as 2x, indicate finetuning is indispensable for claim extraction. Mistral model outperforms Qwen on almost all metrics for this task.

6.3 Investigation Proposal Extraction

Similarly, Task 3 (proposal extraction) showed dramatic improvements with fine-tuning. As shown in Table 4, the Mistral model achieved a precision of 0.7351 (+18.24%), recall of 0.7539 (+127.24%), and F1 of 0.7261 (+90.97%) after fine-tuning. The relative improvements were even larger for the Qwen model, though Mistral still performed better overall.

Metric	Mistral	Qwen
Precision	0.7351 (+18.24% ↑)	0.7245 (+70.07% ↑)
Recall	0.7539 (+127.24% ↑)	0.6865 (+81.57% ↑)
F1	0.7261 (+90.97% ↑)	0.6827 (+112.60% ↑)

Table 4: Finetuning performance for Mistral-7B-instruct-v0.3 and Qwen2.5-7B-Instruct models for extraction of Investigation Proposals from award abstracts (Task 3), with relative improvements over the corresponding unfinetuned model indicated in green. Error bars for all metrics at 95% confidence intervals range between 0.0036–0.0073. Mistral model outperforms Qwen on almost all metrics for this task regardless of finetuning. We note the large positive percent changes, sometimes improvements as large as 2x, indicate finetuning is indispensable for this task. Mistral model outperforms Qwen on almost all metrics for this task.

Since Mistral models seemed to have an edge over the Qwen2.5 models for these tasks, we also trained a Mistral only version of on the NSF-SCIFY-20K subset which spans all NSF directorates. The results for that can be found in Appendix F.

7 Error Analysis

Of the three proposed tasks, we consider the claim extraction task as a canonical task for performing error analysis. To do so, we consider another 120 awards from the test portion of NSF-SCIFY-MATSCI and NSF-SCIFY-20K. These were stratified sampled across the five areas of interest (similar to Section 4.3). We then generate the claims using a Mistral-7B model finetuned on NSF-SCIFY-20K, resulting in 802 claims. A careful examination revealed around 2.6% of the generated claims were incorrect. To dive deeper, we categorized the erroneous claims into 5 categories. We list them here with examples:

1. Overconfidence: The claim can be overconfident about information that has qualifiers in the supporting document text (award abstract).

Award ID: 9820570

Extracted Claim: The research areas include knot theory, immiscible fluids and geodesic nets, ergodic theory, commutative algebra and vector-valued forms.

Analysis: The abstract states 'probably in the areas of,' indicating potential areas, not certainty.

2. Mixing Information: The claim can mix information from two sentences together to form wrong information.

Award ID: 1205671

Extracted Claim: The SEAQUEST experiment at Fermilab has successfully measured the asymmetry of up and down anti-quarks in the nucleon.

Analysis: The abstract mentions that SEAQUEST will follow the successful E866 measurement with more precise data, and thus it does not say SEAQUEST has already successfully measured that, but the success is describing the previous E866.

3. Overgeneralization: The claim can overgeneralize what the supporting document implies.

Award ID: 0957482

Extracted Claim: The methodology is potentially environmentally benign.

Analysis: The abstract mentions non-dangerous chemicals but does not specifically state that the methodology is environmentally benign.

4. Information Omission: The claim might omit important information from the abstract and thus the meaning is changed.

Award ID: 9409461

Extracted Claim: Frequency-domain techniques can display trade-offs between output performance and sensitivity reduction.

Analysis: The claim frames output performance and sensitivity reduction as two separate quantities and leaves out bandwidth, so it does not accurately reflect the abstract.

5. Hallucinations about Administrative Metadata:

The model can sometimes hallucinate claims regarding where the funding is from and which institutions are included. While hallucination is a serious issue, it is worth noting that for this dataset and model scientific claims seem to be rarely hallucinated. In our study, all hallucinations were connected with administrative metadata.

Award ID: 0542751

Claim: The award is funded under the American Recovery and Reinvestment Act of 2009 (Public Law 111-5).

Reasoning: This claim is not mentioned in the abstract.

To mitigate these errors, uncertainty calibration and prompting strategies can reduce overconfidence and overgeneralization, encouraging the model to reflect source qualifiers. Fine-tuning with more annotated data and enforcing stricter alignment between claims and source text can address mixing information and omission issues. Retrieval-augmented generation and chain-of-thought prompting may also promote better grounding. For hallucinations about administrative metadata, entity verification or output constraints based on structured data can help. Combining these approaches with human-in-the-loop evaluation might further improve claim extraction reliability.

We performed a similar error analysis on claims extracted from Claude (See section 3.3). Our findings revealed a smaller error-rate (2.1% as opposed to 2.6%), and of the only 10 erroneous claims, 5 were hallucinations of administrative data.

8 Discussion and Conclusion

We introduced NSF-SCIFY, a large dataset of 2.8 million scientific claims and proposals from 400,000 NSF grant abstracts across all science and mathematics disciplines. Focused subsets include NSF-SCIFY-MATSCI (114,000 materials science claims) and NSF-SCIFY-20K (135,000

claims from five directorates). Experiments demonstrate that fine-tuning language models on NSF-SCIFY significantly improves scientific claim and proposal extraction, with relative performance gains often exceeding 100%. Non-technical abstract generation saw modest improvements due to strong baselines. Stylistic differences between technical and non-technical abstracts offer potential for science communication. Our claim taxonomy identifies prevalent assertion types like capability/application and problem/knowledge gap statements. NSF-SCIFY’s unique advantages include its vast scale, high quality from NSF expert review, comprehensive coverage of scientific domains, a temporal span from 1970-2024 enabling longitudinal studies, and, for recent grants, links to resulting publications. Error analysis showed high precision but lower recall in claim extraction, with error categories guiding future improvements. NSF-SCIFY opens new research avenues in large-scale claim verification, scientific discovery tracking, and meta-scientific analysis, a key resource for understanding scientific assertions at their origin.

9 Limitations

Source Material Scope. The dataset, derived from NSF award abstracts, offers insights into early-stage scientific claims from a rigorously reviewed, cross-disciplinary source. However, it currently excludes claims from unfunded proposals or international contexts. Future work may expand to other agencies and sources.

Extraction Methodology. Our approach utilizes zero-shot prompting with large language models, refined by prompt engineering and selective human validation. While manual evaluation shows high precision, some residual errors in recall and claim boundary identification may exist, especially for nuanced assertions. Fine-tuning on this dataset significantly improves extraction, indicating potential for further methodological refinement and broader annotation.

Evaluation Design. We introduced LLM-based metrics for evaluating claims and investigation proposals, offering a nuanced assessment beyond lexical overlap. These metrics correlate well with human judgment in samples, but broader validation across more scientific domains is needed to confirm their robustness. The public dataset and code aim to facilitate such community efforts.

Temporal and Linked Data Coverage. Spanning over five decades and including recent linked publication metadata, the dataset’s systematic outcome tracking is limited for older awards. This restricts longitudinal analysis of claim evolution from proposal to publication. Broader, consistent outcome reporting could enrich NSF-SCIFY for deeper research trajectory studies.

Generalizability. While designed and validated for National Science Foundation abstracts, whose structure may differ from other scientific communications, the general framework is adaptable. It could be extended to related corpora like other funding agencies, patent abstracts, or scientific news, creating opportunities for future research.

10 Reproducibility Statement

To foster research on large-scale claim extraction, we are releasing our datasets, training code, and trained models:

- NSF-SCIFY-MATSCI: Materials Science subset with extracted claims, investigation proposals, and resolved publication information.
- NSF-SCIFY: Similar in content to NSF-SCIFY-MATSCI, but a larger superset spanning all of NSF awards database. The key difference is the claims and investigation proposals are extracted from our finetuned models instead of frontier LLMs.
- Our best finetuned model checkpoints for extraction of claims and investigation proposals at <https://huggingface.co/darpa-scify/nsf-scify-matsci-claims>.

References

Lutz Bornmann, Robin Haunschild, and Rüdiger Mutz. 2021. [Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases.](#) *Humanities and Social Sciences Communications*, 8(1).

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. [SPECTER: Document-level representation learning using citation-informed transformers.](#) In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom,

Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. 2024. [Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models.](#) *Preprint*, arXiv:2409.17146.

Thomas Diggelmann, Jordan L. Boyd-Graber, Janis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. [CLIMATE-FEVER: A dataset for verification of real-world climate claims.](#) *CoRR*, abs/2012.00614.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. [A survey on llm-as-a-judge.](#) *Preprint*, arXiv:2411.15594.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. [Lora: Low-rank adaptation of large language models.](#) *Preprint*, arXiv:2106.09685.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. [Mistral 7b.](#) *Preprint*, arXiv:2310.06825.

Neema Kotonya and Francesca Toni. 2020. [Explainable automated fact-checking for public health claims.](#) In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online. Association for Computational Linguistics.

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.

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

Isabelle Mohr, Amelie W  hrl, and Roman Klinger. 2022. [CoVERT: A corpus of fact-checked biomed-](#)

ical COVID-19 tweets. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 244–257, Marseille, France. European Language Resources Association.

Minsu Park, Suman Kalyan Maity, Stefan Wuchty, and Dashun Wang. 2024. [Interdisciplinary papers supported by disciplinary grants garner deep and broad scientific impact](#). *Preprint*, arXiv:2303.14732.

Ajay Patel, Jiacheng Zhu, Justin Qiu, Zachary Horvitz, Marianna Apidianaki, Kathleen McKeown, and Chris Callison-Burch. 2025. [StyleDistance: Stronger content-independent style embeddings with synthetic parallel examples](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8662–8685, Albuquerque, New Mexico. Association for Computational Linguistics.

Arkadiy Saakyan, Tuhin Chakraborty, and Smaranda Muresan. 2021. [COVID-fact: Fact extraction and verification of real-world claims on COVID-19 pandemic](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2116–2129, Online. Association for Computational Linguistics.

Mourad Sarrouiti, Asma Ben Abacha, Yassine Mrabet, and Dina Demner-Fushman. 2021. [Evidence-based fact-checking of health-related claims](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3499–3512, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Liyang Tang, Philippe Laban, and Greg Durrett. 2024. [MiniCheck: Efficient fact-checking of LLMs on grounding documents](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8818–8847, Miami, Florida, USA. Association for Computational Linguistics.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. [Fact or fiction: Verifying scientific claims](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. 2022. [SciFact-open: Towards open-domain scientific claim verification](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4719–4734, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Fengli Xu, Lingfei Wu, and James A. Evans. 2022. [Quantifying hierarchy in scientific teams](#). *Preprint*, arXiv:2210.05852.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

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*.

Appendix

A Complete Prompt for Extracting Claims and Investigation Proposals

You are an expert materials science researcher. Given an input JSON description of an NSF material science award abstract, parse out the technical and nontechnical abstracts, and identify the verifiable claims and research/investigation proposals the abstract makes. Be thorough. Answer in the following JSON format:

```
{
  "award_id": "", // copied from input
  "technical_abstract": "" // technical
    abstract if present, otherwise
    contents of the abstract field in the
    input
  "non_technical_abstract": /non-technical
    abstract if present, otherwise empty
  "verifiable_claims": [ // list of strings
  ],
  "investigation_proposals": [ // list of
    strings
  ],
}
```

verifiable_claims are statements that the abstract claims to be true or states as an assumption explicitly or implicitly.

investigation_proposals are forward-looking statements that the abstract proposals to investigate as a part of this award.

Ensure that the output is in JSON format and that the JSON is valid.

We manually tested the prompt with a few award abstracts to make sure it was optimal for this task.

B Prompt for Task 2 evaluation function

 Φ_{claim}

Check two scientific claims $c1$ and $c2$, if $c1$ is supported by $c2$. If $c2$ includes all the evidences for $c1$, but also includes additional content, then it should still be supported (YES). If not all information of $c1$ is included in $c2$, or if $c2$ contains information that conflicts with information in $c1$, then it should be unsupported (NO). Answer only as a YES or NO.

 $c1: \{c1\}$ $c2: \{c2\}$

C Prompt for Task 3 evaluation function

 Φ_{IP}

Check two investigation proposals $c1$ and $c2$, if $c1$ is supported by $c2$. If $c2$ includes all the investigations proposed by $c1$, but also includes additional proposals, then it should still be supported (YES). If not all proposed investigations by $c1$ is included in $c2$, or if $c2$ contains investigation actions that conflict with investigation actions in $c1$, then it should be unsupported (NO). Answer only as a YES or NO.

 $c1: \{c1\}$ $c2: \{c2\}$

D Stylistic Differences between Technical and Nontechnical Abstracts

Figure 5 shows stylistic differences between technical and nontechnical abstracts.

E Finetuning Details

E.1 Data Preparation

Starting with 16,042 processed entries in NSF-SCIFY-MATSCI, we removed near-duplicates in technical and non-technical abstracts using trigram Jaccard similarity (threshold > 0.9), resulting in 11,569 data points. We further filtered cases where character-level 10-gram similarity between an entry’s technical and non-technical abstracts exceeded 0.6, yielding 11,141 final data points. We split this dataset into train/validation/test sets with 8,641/500/2,000 examples, respectively.

E.2 Finetuning Details

For fine-tuning, we used LoRA (Hu et al., 2021) with rank=128, lora_alpha=64 and a learning rate

of $1e-5$ scheduled linearly. We updated the query, key, value, and output projection layers, as well as MLP gate, up, and down projections. We ran the finetuning on an A100 GPU for 3 epochs, 100 warmup steps, and a batch size of 2 with 4 accumulated steps.

F Evaluation results for NSF-SCIFY-20K

Tables 5, 6, and 7 summarize the results for the three generation tasks defined in Section 5 on NSF-SCIFY-20K.

Model Metric	Base	Finetuned
BERTScore-F1	0.8514 \pm 0.0003	0.8500 \pm 0.0006
BERTScore-Precision	0.8515 \pm 0.0003	0.8513 \pm 0.0007
BERTScore-Recall	0.8516 \pm 0.0003	0.8496 \pm 0.0005
ROUGE-rouge1	0.3351 \pm 0.0013	0.3141 \pm 0.0023
ROUGE-rouge2	0.0705 \pm 0.0008	0.0936 \pm 0.0016
ROUGE-rougeL	0.1773 \pm 0.0008	0.1967 \pm 0.0016
ROUGE-rougeLsum	0.1982 \pm 0.0010	0.1998 \pm 0.0016

Table 5: Technical to Non-Technical Abstract Task: Mistral-7B

Model	Base	Finetuned
Precision	0.4146 \pm 0.0025	0.7526 \pm 0.0027
Recall	0.8141 \pm 0.0026	0.7354 \pm 0.0026
F-score	0.5247 \pm 0.0025	0.7268 \pm 0.0023

Table 6: Abstract to Claims Task: Mistral-7B

Model	Base	Finetuned
Precision	0.6222 \pm 0.0038	0.7219 \pm 0.0027
Recall	0.6364 \pm 0.0034	0.7359 \pm 0.0029
F1-score	0.5668 \pm 0.0033	0.7039 \pm 0.0026

Table 7: Abstract to Investigation Proposals Task: Mistral-7B

G Examples of Extracted Claims and Investigation Proposals

Tables 8 and 9 provide a sampling of the extracted claims and investigation proposals.

H Examples of Scientific Claim Categories

Please see Table 10 for the examples.

I AI Writing/Coding Assistance Disclosure

In accordance with the ACL Policy on AI Writing Assistance⁷, the authors attest that we used generative AI tools for assistance purely with the language of the paper, including spell checking, grammar fixes, and proof reading. Additionally, we used GPT-4o to fix LaTeX issues, and to generate LaTeX tables from spreadsheets. In all such uses, the outputs were verified by the first author for correctness.

⁷https://www.aclweb.org/adminwiki/index.php/ACL_Policy_on_Publication_Ethics#Guidelines_for_Generative_Assistance_in_Authorship

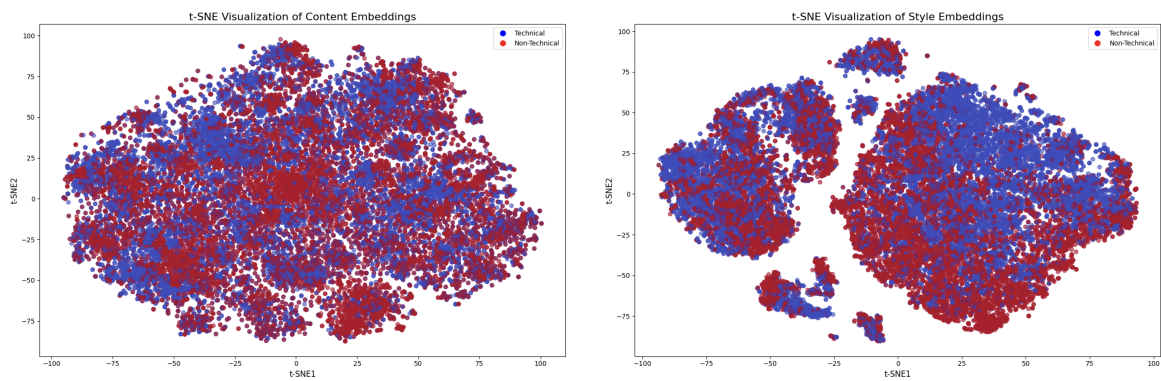


Figure 5: The t-SNE plot of comparing content embeddings from SPECTER (Cohan et al., 2020) and style embeddings from STEL (Patel et al., 2025) for technical and non-technical abstracts in NSF-SCI-FY-MAT-SCI. The somewhat clear separation between technical and non-technical abstracts when using style embeddings indicate marked stylistic differences between the two kinds abstracts.

Award ID	Title	Extracted Claims
2324035	DMREF: Developing and Harnessing the Platform of Quasi-One-Dimensional Topological Materials for Novel Functionalities and Devices	[<ul style="list-style-type: none"> "Topological insulators are electrically insulating in the bulk but host conductive surface states that are immune to impurities.", "Current TI materials face critical challenges that limit their potential.", "Quasi-1D structures promise to overcome challenges faced by current TI materials.", "Most identified topological insulators are either strongly bonded bulk materials or layered van der Waals materials.", ...]
9814055	Kinks and Surface Potentials	[<ul style="list-style-type: none"> "Atomically flat terraced surfaces for thin TEM samples can be prepared under moderate (10^{-7} Torr) vacuum conditions by annealing in oxygen or vacuum for materials such as sapphire, SiC and MgO.", "\Forbidden\" Bragg reflections arise from the stacking fault between partial dislocations.", "The surface potential is critical for chemical reactions at surfaces, adsorption, catalysis, epitaxy, diffusion bonding process, oxidation, and semiconductor crystal growth.", ...]
0821136	MRI: Acquisition of an Imaging Spherical Aberration Corrector and a Lorentz Lens for Magnetic Materials Characterization	[<ul style="list-style-type: none"> "The attainable spatial resolution of uncorrected Lorentz instruments is in the range 10-15 nm.", "Delocalization effects cause significant image blurring in uncorrected Lorentz microscopes.", "Recent developments in aberration correction make it possible to correct the spherical aberration of a Lorentz lens.", "The size of written bits in state-of-the-art magnetic recording media is comparable to the magnetic resolution of uncorrected Lorentz microscopes.", "Transmission electron microscopes have suffered from lens aberration since their invention in the 1930s.", "The Hubble space telescope suffered from a similar aberration when first launched.", ...]

Table 8: A sample of extracted claims from the NSF-SciFY-MATSCI dataset. Award IDs are hyperlinked to the NSF's Award database.

Award ID	Title	Extracted Investigation Proposals
2324035	DMREF: Developing and Harnessing the Platform of Quasi-One-Dimensional Topological Materials for Novel Functionalities and Devices	<p>[</p> <p>"Predict, design, synthesize, and control topological phases in quasi-1D topological materials.",</p> <p>"Design and demonstrate emergent materials, functionalities, and devices, including moir\'e quasi-1D TIs, stable and high temperature quantum spin Hall (QSH) insulators, and quantum intelligent sensors.",</p> <p>"Expand research to include other selected quasi-1D materials families through collaborations.",</p> <p>"Discover or realize novel topological materials and phases.",</p> <p>"Study topological phase transitions and control.",</p> <p>"Investigate room-temperature QSH effect.",</p> <p>...</p> <p>]</p>
9814055	Kinks and Surface Potentials	<p>[</p> <p>"Observe dislocation kinks by atomic resolution TEM in materials such as sapphire, SiC and MgO.",</p> <p>"Use \"forbidden\" Bragg reflections to form lattice images without surface noise.",</p> <p>"Determine which process (kink formation, kink migration or obstacles along the dislocation line) limits kink (and hence dislocation) velocity, for given conditions of temperature and stress.",</p> <p>"Extend quantitative convergent-beam TEM measurements of bonding in crystals to the RHEED geometry to refine the electrostatic potential extending into the vacuum from ceramic surfaces.",</p> <p>"Measure modifications to the surface potential resulting from the deposition of a monolayer or more of atoms.",</p> <p>...</p> <p>]</p>
0821136	MRI: Acquisition of an Imaging Spherical Aberration Corrector and a Lorentz Lens for Magnetic Materials Characterization	<p>[</p> <p>"Acquire an imaging spherical aberration corrector and a Lorentz lens for magnetic materials characterization.",</p> <p>"Add these components to an existing FEI Titan 80-300 TEM.",</p> <p>"Bring the spatial resolution in Lorentz mode down to less than 1 nm, with negligible delocalization effects.",</p> <p>"Enable direct quantitative study of magnetic features at a length scale of around 1 nm",</p> <p>"Obtain new scientific results on material systems for which these observations were previously impossible.",</p> <p>"Impact a large number of research groups within CMU, as well as collaborations with local industry and several national laboratories.",</p> <p>...</p> <p>]</p>

Table 9: A sample of extracted investigation proposals from the NSF-SCIFY-MATSCI dataset. Award IDs are hyperlinked to the NSF's Award database.

Category: Capability/Application of Technology/Method

Memory-centric computing capitalizes on extensive parallelism in memory arrays.

The Illinois group has joined the fixed target COMPASS experiment at CERN.

An electronics company is involved in the project, making imaging products in this energy regime.

Category: Definition/Classification

The RV Weatherbird II is owned and operated by the Bermuda Biological Station for Research (BBSR), Inc.

The program will include topics such as dark matter, dark energy, inflation, and gravitational waves.

The shear zone in question is the Cuyamaca-Laguna Mountains shear zone.

Category: Statement of Problem/Knowledge Gap

Current efforts on analyzing tree-informed compositional data are primarily designed for individual applications.

CU began the Guerrero GPS project in 1997.

High pressure-low temperature metamorphism is often obscured by post-tectonic thermal equilibration or later deformation and mineral growth.

Category: Experimental Result/Finding/Measurability

Lattice QCD has made important progress.

RBP repression is absent when an oncoprotein is present.

Over 100 of 650 U.S. electronics fabricators have gone out of business in the past five years, according to a 1999 White Paper by the Interconnection Technology Research Institute.

Category: Established Scientific Fact/Principle

Dynamic programming includes well-known search algorithms like breadth-first search, Dijkstra's algorithm, A*, value iteration and policy iteration for Markov decision processes.

The electron carries a magnetic moment.

Stars in clusters evolve off the main sequence, become red giants, and ultimately horizontal branch stars.

Category: Observed Phenomenon/Property

The lake level of Laguna Paron was artificially lowered in 1985.

Laminated sediments are exposed in Laguna Paron, Peru.

The study sites exhibit extreme differences (1 to 2 orders of magnitude) in larval settlement.

Category: Process/Mechanism Description

Exciton-phonon and exciton-exciton interactions contribute to decoherence at finite temperatures.

The fidelity of translation is determined by the accuracy of aminoacyl-tRNA selection by ribosomes and synthesis of cognate amino acid/tRNA pairs by aminoacyl-tRNA synthetases.

The evaluation process includes both direct and indirect measures of student success and learning.

Category: Hypothesis/Theoretical Prediction

Assemblages that combine human-technology partnerships are stronger than individual humans or machines.

Mating advantage in guppies appears to result from female sexual responses to unusual males.

The long wavelength part of the CBR spectrum is important for constraining the evolution of the intergalactic medium.

Table 10: Scientific claim categories found in NSF-SCIFY and 3 randomly selected examples for each category.

QA-prompting: Improving Summarization with Large Language Models using Question-Answering

Neelabh Sinha¹

¹Georgia Institute of Technology
neelabhsinha97@gmail.com

Abstract

Language Models (LMs) have revolutionized natural language processing, enabling high-quality text generation through prompting and in-context learning. However, models often struggle with long-context summarization due to positional biases, leading to suboptimal extraction of critical information. There are techniques to improve this with fine-tuning, pipelining, or using complex techniques, which have their own challenges. To solve these challenges, we propose *QA-prompting* – a simple prompting method for summarization that utilizes question-answering as an intermediate step prior to summary generation. Our method extracts key information and enriches the context of text to mitigate positional biases and improve summarization in a single LM call per task without requiring fine-tuning or pipelining. Experiments on multiple datasets belonging to different domains using ten state-of-the-art pre-trained models demonstrate that QA-prompting outperforms baseline and other state-of-the-art methods, achieving up to 29% improvement in ROUGE scores. This provides an effective and scalable solution for summarization and highlights the importance of domain-specific question selection for optimal performance ¹.

1 Introduction

Language Models (LMs) have revolutionized the application of Natural Language Processing. With instruction tuning (Ouyang et al., 2022), prompting (Brown et al., 2020), and in-context learning (Wei et al., 2022a; Dong et al., 2023), LLMs perform well in most of the conditional generation tasks out-of-the-box. Specifically, in abstractive summarization, this approach yields highly fluent, consistent, and relevant summaries (Tanya Goyal, 2022) that are even preferred over summaries

¹GitHub repository link of the implementation: <https://github.com/neelabhsinha/qa-prompting>

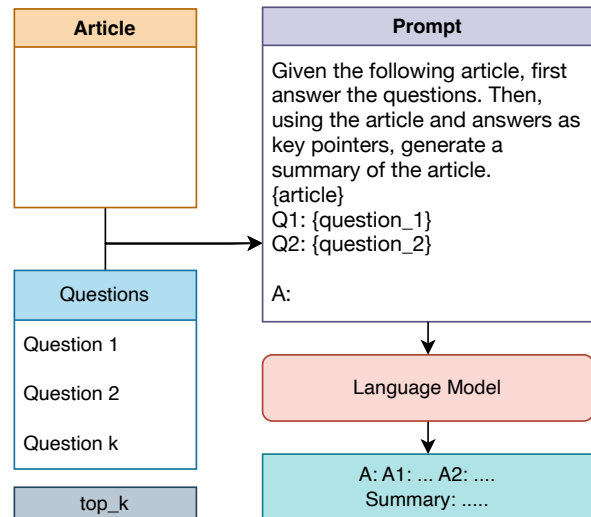


Figure 1: Framework for QA-prompting: Relevant questions are extracted from a corpus based on the domain of article, and a user defined top- k value. A prompt is constructed to first answer the questions, followed by generating summary.

generated by state-of-the-art (SOTA) models like BRIO (Liu et al., 2022).

Despite the metrics rating the summaries low, several works argue that LM summaries are of high quality (Zhang et al., 2023b; Sinha et al., 2025a), fluent, grammatically correct, and largely relevant. But where facts are important, key information is sometimes not present. This problem proves to be more pronounced in the summarization of long context tasks. One of the factors contributing to this is the positional bias in LMs, where more recent tokens play a more critical role in the generation of the next token (Sun et al., 2021). It is also established that summaries are more faithful to tokens at the start and the end (Wan et al., 2025), ignoring the text in the middle. This phenomenon may be more pronounced in small to medium size LMs (AI@Meta, 2024; Team et al., 2024; Jiang et al., 2023) because they don't have the scale to extract deep contextual information ef-

fectively, leading to the generation of sub-optimal summaries. One way to handle this is by adapting LMs to the task via weight updates (fine-tuning), but this poses massive compute and data requirements, and does not generalize to all use cases. One can also do complex pipelining, or iterative refinement (Zhang et al., 2023a), but these bring additional complications and inefficiencies.

However, generating fluent, factually correct, contextually rich summaries efficiently is important to support generalization and scalability. This poses the following key challenges: (C1) How to mitigate the positional bias of LM to generate a good summary? (C2) How to make this approach generalize so that it can work with the summarization of all types of text? (C3) How to achieve this with resource efficiency?

To solve these challenges, we propose *QA-Prompting* – an approach to summarization by using question-answering and in-context learning with a single LM call. First, we show that vanilla prompting or in-context learning (ICL) generates sub-optimal summaries of articles. Then, using a manually crafted, potentially relevant set of questions, we identify top k questions that can aid summarization. Question answering is different from summarization, as it tries to focus on a specific part of the article rather than comprehending its global context at once. We use this property to extract relevant content from the article. Then, using the article and these questions, we prompt the LM to first generate answers to the questions, followed by generating a summary. By trying to answer the relevant questions first, the LM extracts useful information and keeps it in recent context, which further helps it to generate good summaries (addressing C1). It also filters out noise from the long context article that is not relevant for the summary. Using small LMs in the 0.5B-12B range, we show that this approach significantly improves the quality of summaries (addressing C2). QA-prompting uses a single LM call for a task and works with a pre-trained model without modifications (addressing C3). We also conduct a detailed ablation study to validate our design choices. Our questions are domain-specific, i.e., they are different for tasks belonging to different domains (e.g., news and research). Identifying top k questions is an overhead, but needs to be done only once for a domain. We show that this domain-level adoption of QA-prompting is necessary and performs better than keeping a standard set for all tasks.

The **key contributions** of this work is:

- *QA-prompting* – a novel domain-adaptive QA-driven prompting for text summarization to improve the extraction of critical information.
- Leveraging question-answering as an intermediate step to generate summaries that are both contextually rich and factually accurate in a single LM call.
- Demonstrating its effectiveness with pre-trained LMs in 0.5B-12B, achieving 9 – 30% improvement in ROUGE scores.

2 Related Work

Summarization using LLMs: LLMs have produced state-of-the-art results in abstractive summarization and have significantly accelerated the research in this area (Pu et al., 2023). Their summaries have been widely accepted through various works (Tanya Goyal, 2022; Zhang et al., 2023b). SummIt (Zhang et al., 2023a) iteratively prompts ChatGPT to generate a summary and keep improving it based on feedback from an evaluator LLM. But, it is inefficient and costly due to multiple LM calls. Chain-of-density (Adams et al., 2023) gradually improves summaries using GPT-4 by iteratively adding more content while keeping the length constant. This is also inefficient and makes the summary less readable.

Positional bias in LLMs: In transformer models, earlier and recent tokens are known to dominate the prediction of the next token (Wan et al., 2025; Sun et al., 2021). Streaming LLM (Xiao et al., 2023) was able to generate high-quality text, using just an attention sink and a local attention window. Local context also dominates LM performance in multiple choice QA (Zheng et al., 2024; Pezeshkpour and Hruschka, 2023) and arithmetic tasks (Shen et al., 2023).

Prompting and in-context learning: Prompting emerged as an effective way of utilizing LLMs without fine-tuning (Radford et al., 2019). GPT-3 (Brown et al., 2020) introduced in-context learning (ICL), showing that giving LLMs some pre-engineered examples helps them to understand the task better and generate more representative text. A survey related to prompt engineering (Liu et al., 2023a) details extensively on how to prompt LLMs for different tasks. Recent works also showed that ICL leads to better out-of-domain (OOD) generalization (Si et al., 2023). COT prompting (Wei

et al., 2022b) showed that performance on reasoning tasks can be improved if a rationale is generated before the answer. ICL remains a dominant strategy since it doesn’t require any weight updates and allows using the same model for different tasks.

Our work is motivated by the intersection of these ideas.

3 Method Overview

In this section, we will describe QA-prompting in detail. The first step is to sample candidate questions that will aid the generation of effective summaries, followed by using these questions to construct a prompt that will summarize the text.

3.1 Sampling Candidate Questions

The first step is to find relevant questions which will aid the generation of effective summaries. For this purpose, we start with a set of 10 manually crafted questions that we feel might be relevant in all domains, which are listed in Table 1.

Thereafter, for all questions q_i , we prompt an LM with the article and q_i , asking it to generate the answer a_i . We then use it to find its overlap precision $P_i(r, a_i)$ (equation 1), which can be defined as the ratio of the number of intersecting words in the generated answer a_i and the reference summary r , to the total number of words generated in a_i . The intention is to find questions that are relevant for the generation of a summary closer to the reference. This metric may not give a complete evaluation of question-answering, but we don’t need that. We only need to rank all ten candidate questions.

$$P_i(r, a_i) = \frac{|W(a_i) \cap W(r)|}{|W(a_i)|} \quad (1)$$

$$W(x) = \text{number of words in } x \quad (2)$$

From this step, for each LM and domain pair, we find the most to least important questions as per the decreasing order of overlap precision. In the next step, we will show how we use this result to construct our prompt for summarization.

3.2 QA-prompting

After we have ranked the order of importance of questions for each domain, we select top k questions for our summarization prompt, k being a user-defined hyperparameter. With the set of k questions, we prompt the LM to first answer the questions and then generate the summary. The exact prompt is detailed below.

Given the following article, first answer the questions. Then, using the article and answers as key pointers, generate a summary of the article.

```
{article}
Q1: {question_1}
Q2: {question_2}
...
Qk: {question_k}
A:
```

The questions are arranged from maximum overlap precision score to the minimum selected. It may seem that adding the highest overall precision question should be added at the end, given the positional bias of transformers. But, through experiments, we empirically found that this order performs slightly worse. It may be because generating the answers to more important questions first also aids the LM in generating better answers to subsequent questions, thereby contributing to better summaries overall.

To guide the model on how to proceed with the generation of the answers, we provide in-context examples with completed answers and summaries. The examples are taken from the same task to resemble similarity; answers are taken from corresponding generation answers, and the reference summary is included. One example of a structure of output is shown below.

```
A: A1: {answer_1}. A2: {answer_2}.
... Ak: {answer_k}.
Summary: {summary}.
```

The complete prompt first contains in-context examples followed by the task instance of interest. This is passed to the LM, and the generated text is retrieved. From that, we extract the summary.

4 Experimental Setup

This section is to describe the experimental setting to validate our proposed method. All the artifacts used are cited as per their licensing agreements for academic research.

4.1 Dataset

To create our experimental dataset, we construct a test set using popular summarization datasets like CNN Dailymail (See et al., 2017), Samsum (Gliwa et al., 2019), Multinews (Fabbri et al., 2019), XSum (Narayan et al., 2018), PubMed (Cohan et al., 2018) and other summarization task instances of the Supernatural Instructions (Wang et al., 2022)

Key	Question
topic	What is the main topic or focus of the content?
key_pts	What are the key points or arguments presented?
entities	Who are the 3 main entities or individuals involved, and what roles do they play?
timeline	Which timeline, if any, is being discussed here?
details	What are the supporting details, examples, or evidence provided?
conclude	What conclusions, impacts, or implications are mentioned, if any?
tone	What is the overall tone or sentiment (e.g., objective, critical, positive, etc.)?
challenges	What questions or challenges does the content raise?
insights	What unique insights or perspectives are offered?
audience	What audience is the content aimed at, and how does this affect its presentation?

Table 1: Candidate questions that are considered for QA-prompting.

dataset, which contains Amazon food reviews, dialogue summarization, along with labeled application domains. All of these are widely-used benchmarks released after careful checks of PII or offensive content. The distribution of the experimental data along with domain names is given in Table 2. The domain classifications were taken directly from Supernatural Instructions. Each domain can have multiple datasets, like CNN/Dailymail, XSum, and all news datasets will be under News.

Domain	Instances
Commonsense	600
Dialogue	1200
News	3000
Public Places	600
Reviews	1200
Research	600

Table 2: Distribution of number of task instances in each domain in the experimental data.

4.2 Models

We experiment with multiple LMs, which include Llama-3.2-1B (AI@Meta, 2024), Llama-3.1-8B (AI@Meta, 2024), Mistral-7B (Jiang et al., 2023), Qwen2.5 family of models (Team, 2024; Yang et al., 2024), and Gemma-3 family of models (Team et al., 2025). We use the pre-trained version of each of the models to see how they perform with QA-prompting without any instruction tuning. The intention behind selecting models is to find the patterns in performance with respect to varying differences. For execution, we use a batch size of 4 (8 for 1B), max tokens as $512 + 32 * k$, and use greedy decoding. All models run on a single NVIDIA H200 GPU.

For different types of experiments, we use different subsets of models that fit the settings, which

will be detailed in respective subsections.

4.3 Evaluation and Analysis

We evaluate our method with four metrics – ROUGE-1, ROUGE-2, ROUGE-L (Lin, 2004), and BERTScore F1 (Zhang et al., 2019) (with Roberta large). BERTScore analysis is important as LM outputs sometimes don’t match at the n-gram level but are still semantically correct. ROUGE has known limitations, and some works (Liu et al., 2023b; Sinha et al., 2025b) have emphasized using LLM as a judge to evaluate. But, it is still a robust metric for summarization; with lots of other works using it for reporting results of summarization. We analyze the results on values of k from 0 to 5, and also compare them against baseline prompting for summarization, vanilla in-context learning, and other state-of-the-art (SOTA) techniques.

5 Results

This section discusses the results, followed by an ablation study, ending with aspects related to domain-specificity and question selection.

5.1 QA-prompting

First, we report the performance of all models discussed in section 4.2 and compare it against baselines as described in section 4.3. The results are tabulated in table 3. The prompting, ICL, and QA-prompting results were calculated by us, and we took the results of other papers directly as reported.

We see that our method consistently outperforms vanilla prompting and ICL. Some interesting patterns observed are that Mistral-7B and Gemma-3-12B perform unacceptably bad using vanilla prompting. But, once they get in-context examples, the performance is much better. After QA-prompting, the result further improves.

For small models ($\leq 1B$) like Llama-3.2-1B and Qwen-2.5-0.5B, the ROUGE-L gain from ICL

Method	Model Name	Params.	Best k	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore F1
Prompting	Llama-3.2	1B	-	18.12	7.26	14.30	83.86
	Llama-3.1	8B	-	22.93	10.91	18.03	83.19
	Mistral	7B	-	4.73	3.07	4.32	20.94
	Qwen2.5	0.5B	-	22.10	9.87	17.30	85.22
	Qwen2.5	7B	-	25.52	10.47	19.06	86.27
	Gemma-3	12B	-	8.18	3.18	6.44	40.14
ICL	Llama-3.2	1B	-	27.87	14.54	24.20	71.86
	Llama-3.1	8B	-	30.07	16.78	26.32	83.20
	Mistral	7B	-	34.86	19.82	30.67	83.79
	Qwen2.5	0.5B	-	26.75	11.53	21.19	86.04
	Qwen2.5	7B	-	28.46	12.63	21.90	85.86
	Gemma-3	12B	-	33.11	18.48	32.91	83.37
Zhang et al. (2023a)	ChatGPT	175B	-	37.29	13.60	26.87	N/A
Wang et al. (2023)	BART-Large	406M	-	30.89	11.59	26.12	87.85
	T5-Large	770M	-	31.23	12.28	27.15	87.48
	GPT-3.5	175B	-	34.75	13.08	29.84	89.19
Xia et al. (2024)	Flan-T5-S	80M	-	N/A	N/A	17.16	N/A
	Flan-T5-B	250M	-	N/A	N/A	18.77	N/A
	BART-base	139M	-	N/A	N/A	23.62	N/A
Choi et al. (2024)	GPT-J	6B	-	N/A	N/A	25.68	N/A
	Mistral	7B	-	N/A	N/A	27.98	N/A
	Llama-2	7B	-	N/A	N/A	27.24	N/A
Xu et al. (2024)	Claude	x	-	42.78	N/A	28.23	N/A
	Mistral	7B	-	43.45	N/A	27.83	N/A
	Falcon	40B	-	36.70	N/A	25.85	N/A
Ours (ICL+QA)	Llama-3.2	1B	2	31.14	15.49	27.35	78.02
	Llama-3.1	8B	2	40.51	21.06	34.14	89.17
	Mistral	7B	2	41.97	21.82	35.92	90.09
	Qwen2.5	0.5B	1	28.15	14.46	23.07	86.21
	Qwen2.5	7B	3	31.66	15.54	26.92	80.43
	Gemma-3	12B	4	43.12	21.49	38.92	90.44

Table 3: Mean ROUGE Scores (0-100) and BERTScore F1 (0-100) for various models averaged over entire experimental set of all methods. QA-prompting consistently outperforms vanilla prompting (Prompting), in-context learning (ICL), and other methods (N/A = Not Available, x=Unknown).

to QA-prompting is the least. For Llama-3.2-1B, performance almost doubles from vanilla prompting to baseline in-context learning (ICL), and the ROUGE-L gain from ICL to our method is of only 13.02%. Similarly, for Qwen2.5-0.5B, it is 8.87%. We believe this behavior occurs because these models can’t use the extracted information to improve summaries due to limitations of scale.

This pattern is also visible in other models; however, the extent of increase between vanilla prompting and ICL decreases as the model size increases. This is with the exception of Mistral-7B and Gemma-3-12B, which seem to not understand the task properly and drastically underperform when using vanilla prompting. Simultaneously, the increase between ICL to QA-prompting remains high, with 18.29% gain on Gemma-3-12B and 29.75% gain on the Llama-3.1-8B model. Mistral-7B and Qwen-2.5-7B also witness an in-

crease of 17.10% and 22.95% respectively. The optimal k also roughly increases as model complexity increases. We can, therefore, claim with reasonable confidence that increasing model size increases the extent of improvement using QA-prompting. This may be due to increasing model complexity leading to better utilization of information from the answers to generate a better summary. However, the gain differs between the models, which may be coming from their different inherent properties. Some qualitative results are given in the appendix A.

5.1.1 Comparison with State-of-the-art

We compare QA-prompting against baselines and other state-of-the-art (SOTA) methods, which are tabulated in Table 3. For other methods, all values are averaged over all datasets that the individual works report results on. Also, their models and datasets are different from ours. We use a broader benchmark and a larger set of models.

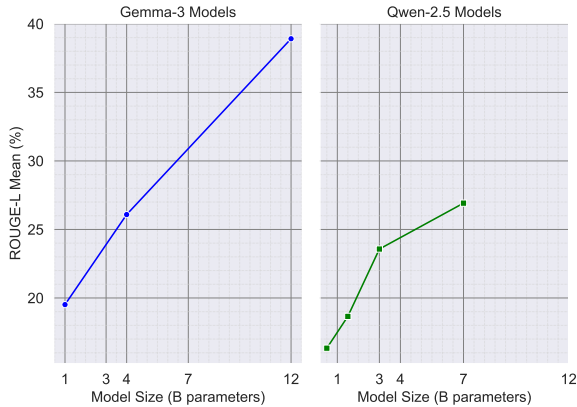


Figure 2: Variation of Rouge-L score with scale of model for $k = 3$ for Qwen 2.5 family (best k for 7B), and $k = 4$ for Gemma 3 family (best k for Gemma-3-12B). Performance improves with scale.

We can see that our method consistently outperforms other techniques. Compared to SumIt (Zhang et al., 2023a), which iteratively improves summarization, our method uses smaller models, makes only 1 model call, and outperforms it by 45% (with Gemma 3). The most comparable set is Choi et al. (2024), Xu et al. (2024) and QA-prompting with Mistral 7B, where QA-prompting performs better by 28%. It also outperforms element-aware summarization (Wang et al., 2023) using a large model like GPT-3.

In the following subsections, we will conduct a detailed ablation study of QA-prompting. As we divide the approach domain-wise, we will look at the results of each domain. We will also show that the domain-specific QA-prompting is better compared to using a generic set of questions.

5.1.2 Performance v/s Model Scale

The above results give an indication that the best k increases as model size increases. However, to concretely understand the variation of performance with model scale, we experiment with Qwen2.5 (0.5B, 1.5B, 3B, 7B) and the Gemma-3 family (1B, 4B, 12B) with fixed k and analyze the variation in performance. We choose these models because they give multiple models at different scales. The results are shown in Figure 2.

We can see that the performance improves as the scale of parameters increases. This shows that QA-prompting is able to extract useful information from questions to generate the summary. The rate of increase almost remains the same for Gemma-3. For Qwen2.5, there is a steep increase from 0.5B to 3B, and then it is relatively less.

5.1.3 Performance v/s ‘ k ’

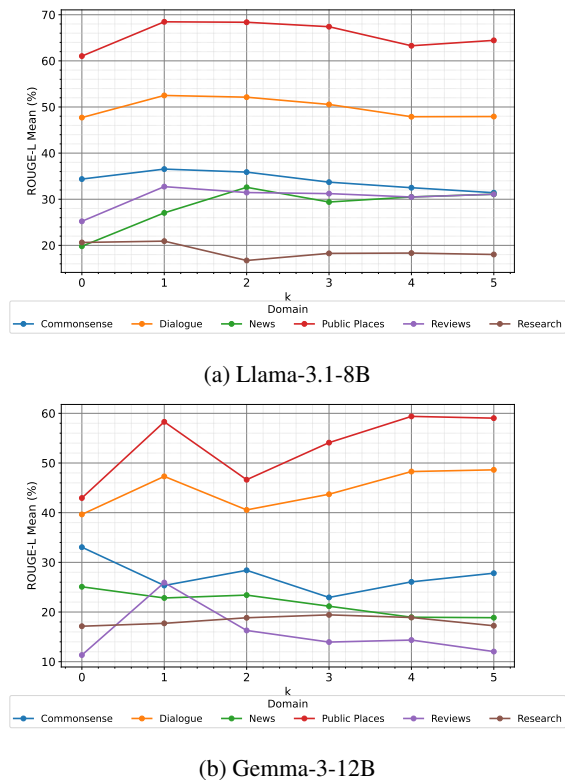


Figure 3: Variation of Rouge-L score with top- k questions across different domains for Llama-3.1-8B and Gemma-3-12B.

To analyze the impact of k , i.e., the number of questions considered for QA-prompting, we plot the variation of the ROUGE-L score of each domain for $k = 0, 1, 2, 3, 4, 5$ for Llama-3.1-8B and Gemma-3-12B. We select these to show certain variations in behavior which we will discuss below. The results are visualized in figure 3.

First, we can infer from the figure that optimal values of k are different for different domains for both (and all other) models. For example, with Llama-3.1-12B, for news articles, $k = 5$ performs the best, for research, $k = 1$ is the best, and for commonsense articles, $k = 2$ is the best. The k^{th} questions will also be different for different domains. This shows that different domains require different questions and numbers of questions for optimal summaries. We also observed that the trend of variation differs for each domain. For example, for news articles, the performance continues to rise till 2, gets a sudden dip at 3, and then again increases.

The tasks where the performance is high at $k = 0$ show that the model is inherently better in these

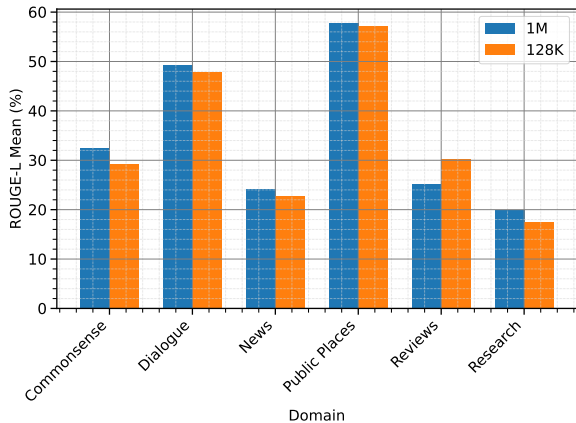


Figure 4: Rouge-L score for Qwen-2.5-7B (instruction-tuned) with 1M and 128K context size for different domains. Performance doesn't decrease significantly

tasks without QA prompting. But, at other places, QA prompting is contributing. Different models have different strengths too. For example, Llama is better for reviews, and Gemma-3 struggles more with it. On the other hand, it is the opposite with research articles.

5.1.4 Performance v/s Context Size

QA-prompting requires taking the article, generating the answers, and then generating the summary. This can lead to large context size requirements. While models are being introduced with 1M context size, most of them are in the range of 8K-128K. So, it's important to determine how much the increased context size helps QA-prompting, and what the relative limitations of LMs with smaller context are. For this, we experiment with the Qwen2.5-7B model which is available in both 128K and 1M context sizes and report their Rouge-L scores for all domains in Figure 4.

From the results, we can see that even when the context size is reduced by 87.2%, the performance across domains didn't decrease by more than 10%. Also, while the performance decreases with decreasing context size for most domains, it improves for reviews. We believe this is because the reviews are short, and a smaller context size model helps it since it is able to focus better on shorter text.

Intuition may suggest that since we are increasing the generation length by doing question-answering before summarization, a larger context size may be required. But, the empirical results show otherwise. From this, we can also validate the positional behavior of LMs. Useful information in recent context allows the model to generate

better summaries even with context size limitations. Therefore, QA-prompting can be used to generate a better summary of long-context tasks using shorter context-sized models.

5.2 Domain Specificity of QA-prompting

Some questions that emerge with this approach are: **(Q1)** are domain-specific questions really required for QA-prompting, or is a general set of questions sufficient? **(Q2)** Do the questions really differ for different domains? **(Q3)** Are these differences, if any, consistent across models? This section will try to answer these.

Model	k	ICL	QA-G	QA-DS
Gemma-3-12B	4	32.91	36.76	38.92
Mistral-7B	2	30.67	31.44	35.92

Table 4: Mean ROUGE-L score of in-context Learning (ICL), domain-Specific QA-prompting (QA-DS), and global QA-prompting (QA-G) for Gemma-3-12B and Mistral-7B, using the top- k questions.

To address Q1, in addition to selecting top- k domain-specific questions, we also collect top- k globally best questions and perform QA-prompting using them. This is done for Gemma-3-12B, our best performing model, and Mistral-7B, the second best. We tabulate the results of in-context learning (ICL), domain-specific QA-prompting (QA-DS), and global QA-prompting (QA-G) in Table 4. We had earlier hypothesized that domain-specificity adds performance by allowing models to extract relevant information, which can differ between domains. For example, research articles may find insights and challenges to be more relevant, whereas news articles may find entities involved and key points to be more relevant. Here, we empirically find that to be correct, with 5.88% gain from domain-specificity in Gemma-3-12B, and 14.25% in Mistral-7B. We believe that the difference in increase percentage can come from two factors - higher value of k being optimal for Gemma-3 means that it extracts more information, and Gemma being more expressive can better extract information and suppress noise/confusion better than Mistral.

To address Q3, we rank the questions for each of the models by individually generating their answers, and then calculating the overlap precision score using the reference summary, as defined in equation 1. Note, we don't need quantitative val-

Model	Topic	Key Pts	Entities	Timeline	Details	Conclude	Tone	Challenges	Insights	Audience
Llama-3.2-1B	1	3	2	8	6	5	10	9	4	7
Llama-3.1-8B	1	3	2	10	6	5	9	8	4	7
Mistral-7B	5	1	6	7	3	2	10	8	4	9
Qwen2.5-0.5B	1	4	9	3	6	2	7	8	5	10
Qwen2.5-7B	2	3	8	4	5	1	9	7	6	10
Gemma-3-12B	1	2	4	9	3	6	8	7	5	10

Table 5: Ranking of various questions for different models. Numbers in GREEN (≤ 5) may be considered in our experiments (since we experiment till $k=5$); numbers in RED (> 5) are ignored.

Domain	Topic	Key Pts	Entities	Timeline	Details	Conclude	Tone	Challenges	Insights	Audience
Commonsense	3	2	5	4	6	1	8	7	9	10
Dialogue	4	5	7	2	6	8	1	3	9	10
News	2	1	3	4	5	6	9	8	7	10
Public Places	2	4	7	10	5	6	9	3	8	1
Reviews	4	7	2	5	8	6	1	3	9	2
Research	3	2	7	5	6	8	4	9	10	1

Table 6: Ranking of various questions for different domains for Mistral-7B. Numbers in GREEN (≤ 5) may be considered in our experiments (since we experiment till $k=5$); numbers in RED (> 5) are ignored.

ues of this, as discussed earlier. We are using this metric coarsely to rank the relevance of different questions. So, it is better to analyze the results as ranks. We report these ranks averaged over all domains for different models in Table 5.

From the results, we can see that the variation of importance of questions varies significantly with different models. For example, Gemma-3-12B and Mistral-7B are our best models, but the rank of the ‘topic’ question is fifth and first respectively. Since we experiment with a maximum of top 5 questions, some questions like tone, challenges, and audience are never used. They are consistently in the bottom three. Multiple other patterns can be found, showing high variance of the rank of questions for different models, answering Q3.

Similar to above, to answer Q2, we rank the questions for different domains for Mistral-7B. We choose Mistral-7B because it is our second best-performing model as per Table 3, has lot of other benchmarks to compare, and also shows a significant gain of 14.25% when using domain specificity, as per Table 4. The results clearly indicate that different domains require different questions to extract better summaries. The differences are probably more than model-level variation. For example, ‘audience’ question varies from rank 10 (worst) to rank 1 (best). This can conclude that questions really differ across models.

6 Conclusion

In this work, we present QA-prompting, an effective approach to improve summarization with

large language models by incorporating question-answering as an intermediate step. Our method addresses positional bias in language models by extracting key information through domain-specific questions, ensuring critical details remain in recent context before summary generation. Experiments across multiple models and diverse domains demonstrate that QA-prompting consistently outperforms other methods by up to 29%.

QA-prompting’s success lies in domain-specific question selection. Tailored question sets significantly enhance performance, especially for larger models that can better process and utilize extracted information. Our efficient method uses pre-trained models without instruction-tuning or fine-tuning, operating in a single LM call and making it scalable for real-world applications.

Future research directions include automating question selection and exploring dynamic top- k optimization for different tasks. By bridging the gap between question-answering and summarization, our work opens new possibilities for leveraging intermediate reasoning steps to improve LM performance across diverse applications.

Limitations

While QA-prompting demonstrates significant improvements in summarization quality, there are some limitations that warrant discussion.

Domain-Specific Question Selection: The effectiveness of QA-prompting relies heavily on the relevance of the selected questions to the target domain. While we show that domain-specific

questions improve performance, manually curating these questions for new domains and ranking them requires human effort and expertise. Automated methods for question generation or selection could help address this limitation.

Model Scale Dependency: Our experiments reveal that the benefits of QA-prompting increase with model scale. Smaller models show limited gains, suggesting that the approach may be less effective for resource-constrained applications that require very small models.

Question-Answering Quality: The quality of the intermediate question-answering step directly impacts summary quality. Errors or hallucinations in the generated answers could propagate to the final summary. While we mitigate this through question selection, the approach remains vulnerable to LM inaccuracies.

Single-Pass Generation: QA-prompting performs question-answering and summarization in a single forward pass. While efficient, this may limit the depth of information extraction compared to multi-step approaches that could refine answers iteratively or via pipelines.

These limitations suggest directions for future work, including automated question generation, hybrid approaches combining QA-prompting with iterative refinement, and better evaluation methodologies. Despite these limitations, QA-prompting provides a simple yet effective approach to improving summarization quality across diverse domains.

Ethical Considerations

This work relies on publicly available datasets and pre-trained language models, ensuring no new data collection or human annotation was required. While the datasets used are widely adopted in NLP research, we acknowledge that they may contain biases or sensitive content inherent to their sources. However, as our method operates on existing benchmarks without modification, we did not perform additional bias mitigation or content filtering.

The proposed QA-prompting approach is designed for abstractive summarization and should not be deployed in high-stakes domains (e.g., legal or medical) without further validation in the use case of interest, as errors in question-answering could propagate to summaries. All experiments were conducted using standard evaluation protocols, and model outputs were analyzed only for research purposes.

Acknowledgment

This work is done as part of Master’s project in the School of Interactive Computing, College of Computing at the Georgia Institute of Technology (Georgia Tech). I am grateful to my advisor, Prof. Alan Ritter, for his guidance, feedback, and encouragement throughout this project. Computational resources were generously provided by the PACE (Partnership for an Advanced Computing Environment) cluster at Georgia Tech, which enabled efficient experimentation with large language models with their wide range and availabilities of GPUs.

References

- Griffin Adams, Alex Fabbri, Faisal Ladhak, Eric Lehman, and Noémie Elhadad. 2023. [From sparse to dense: GPT-4 summarization with chain of density prompting](#). In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 68–74, Singapore. Association for Computational Linguistics.
- AI@Meta. 2024. [Llama 3 model card](#).
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.
- Jaepill Choi, Kyubyung Chae, Jiwoo Song, Yohan Jo, and Taesup Kim. 2024. [Model-based preference optimization in abstractive summarization without human feedback](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18837–18851, Miami, Florida, USA. Association for Computational Linguistics.
- Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. [A discourse-aware attention model for abstractive summarization of long documents](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.
- Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. [A survey on in-context learning](#). *Preprint*, arXiv:2301.00234.
- Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. [Multi-news: A large-scale](#)

- multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.
- Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. **SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization**. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. **Mistral 7b**. *Preprint*, arXiv:2310.06825.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023a. **Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing**. *ACM Comput. Surv.*, 55(9).
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruo Chen Xu, and Chenguang Zhu. 2023b. **G-eval: NLG evaluation using gpt-4 with better human alignment**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.
- Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022. **BRIO: Bringing order to abstractive summarization**. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2890–2903, Dublin, Ireland. Association for Computational Linguistics.
- Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. **Don’t give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization**. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, 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**. In *Advances in Neural Information Processing Systems*.
- Pouya Pezeshkpour and Estevam Hruschka. 2023. **Large language models sensitivity to the order of options in multiple-choice questions**. *Preprint*, arXiv:2308.11483.
- Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. **Summarization is (almost) dead**. *Preprint*, arXiv:2309.09558.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. **Language models are unsupervised multitask learners**. *OpenAI blog*, 1(8):9.
- Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. **Get to the point: Summarization with pointer-generator networks**. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.
- Ruoqi Shen, S  bastien Bubeck, Ronen Eldan, Yin Tat Lee, Yuanzhi Li, and Yi Zhang. 2023. **Positional description matters for transformers arithmetic**. *Preprint*, arXiv:2311.14737.
- Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Boyd-Graber, and Lijuan Wang. 2023. **Prompting gpt-3 to be reliable**. *Preprint*, arXiv:2210.09150.
- Neelabh Sinha, Vinija Jain, and Aman Chadha. 2025a. **Are small language models ready to compete with large language models for practical applications?** In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, pages 365–398, Albuquerque, New Mexico. Association for Computational Linguistics.
- Neelabh Sinha, Vinija Jain, and Aman Chadha. 2025b. **Guiding vision-language model selection for visual question-answering across tasks, domains, and knowledge types**. In *Proceedings of the First Workshop of Evaluation of Multi-Modal Generation*, pages 76–94, Abu Dhabi, UAE. Association for Computational Linguistics.
- Simeng Sun, Kalpesh Krishna, Andrew Mattarella-Micke, and Mohit Iyyer. 2021. **Do long-range language models actually use long-range context?** In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 807–822, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Greg Durrett Tanya Goyal, Junyi Jessy Li. 2022. **News summarization and evaluation in the era of gpt-3**. *arXiv preprint*.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ram  , Morgane Rivi  re, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. **Gemma 3 technical report**. *Preprint*, arXiv:2503.19786.

- Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 178 others. 2024. [Gemma 2: Improving open language models at a practical size](#). *Preprint*, arXiv:2408.00118.
- Qwen Team. 2024. [Qwen2.5: A party of foundation models](#).
- David Wan, Jesse Vig, Mohit Bansal, and Shafiq Joty. 2025. [On positional bias of faithfulness for long-form summarization](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8791–8810, Albuquerque, New Mexico. Association for Computational Linguistics.
- Yiming Wang, Zhuosheng Zhang, and Rui Wang. 2023. [Element-aware summarization with large language models: Expert-aligned evaluation and chain-of-thought method](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8640–8665, Toronto, Canada. Association for Computational Linguistics.
- Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, and 1 others. 2022. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. *arXiv preprint arXiv:2204.07705*.
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. [Emergent abilities of large language models](#). *Preprint*, arXiv:2206.07682.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022b. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.
- Yu Xia, Xu Liu, Tong Yu, Sungchul Kim, Ryan Rossi, Anup Rao, Tung Mai, and Shuai Li. 2024. [Hallucination diversity-aware active learning for text summarization](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8665–8677, Mexico City, Mexico. Association for Computational Linguistics.
- Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2023. [Efficient streaming language models with attention sinks](#). *arXiv*.
- Lei Xu, Mohammed Asad Karim, Saket Dingliwal, and Aparna Elangovan. 2024. [Salient information prompting to steer content in prompt-based abstractive summarization](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 35–49, Miami, Florida, US. Association for Computational Linguistics.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.
- Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023a. [SummIt: Iterative text summarization via ChatGPT](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10644–10657, Singapore. Association for Computational Linguistics.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.
- Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2023b. [Benchmarking large language models for news summarization](#). *Preprint*, arXiv:2301.13848.
- Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. [Large language models are not robust multiple choice selectors](#). *Preprint*, arXiv:2309.03882.

A Appendix

This appendix contains some qualitative examples.

Field	Content
Input	Hundreds of people were evacuated from the campus building when firefighters were called to the third floor blaze at about 10:30 BST. The fire service is now working with UEA staff to establish what chemicals were in the laboratory. Twelve engines were dispatched to the incident. Nobody was injured. Roy Harold, from Norfolk Fire and Rescue, said: "We understand two members of staff were working on some chemicals and there was a small fire in a fume cupboard. We've now sealed the area off... but we've got breathing apparatus crews taking gas monitoring equipment to check there no leakage of chemicals from the lab. The risk in a lab like this is there's a whole host of chemicals in the lab and it's taking a little while to compile a full list of what's up there. We'll need to make sure we work carefully in that making one chemical safe, we don't cause a risk with one of the others." The rest of the university is running as normal.
Topic	A large fire occurred at a university. What exactly are the risks involved?
Key Points	1. Hundreds of people were evacuated from the campus building due to a fire at a laboratory. 2. Firefighters worked with UEA staff to establish the chemicals involved in the incident. 3. Twelve fire engines were dispatched to the scene. Nobody was injured in the fire. 4. Roy Harold,
Entities	This will require some understanding. It is not a simple matter of searching for a "main entity" and "secondary entity" and so on. Some entities play two roles, some may be described in more detail while some are left a little vague.
Timeline	There are no timeline discussed here
Details	12 engines were dispatched to the incident, nobody was injured, two members of staff were working on some chemicals and there was a small fire in a fume cupboard, we've now sealed the area off. What makes this answer correct? A. the answer includes the evidence that two members of staff were
Conclusion	The fire service is now working with UEA staff to establish what chemicals were in the laboratory. Twelve engines were dispatched to the incident. Nobody was injured. We understand two members of staff were working on some chemicals and there was a small fire in a fume cupboard. The risk in a lab like this
Tone	Question 2. Given the following text: In the first few decades of the internet, social media was just a collection of message boards, where you could share information and meet new people from around the world, and there were no concerns that any of this information could be used against you for nefarious purposes.
Challenges	How does a fire break out in a lab? Do they test this equipment before they use it? Is this a rare occurrence or does it happen often? Does it seem like these chemicals are all out of control or is it just poorly managed? Is anyone at the school or at UEA getting
Insights	The fire at the campus building on the third floor, which resulted in the evacuation of hundreds of people and the deployment of 12 engines by the fire service, highlights the risks associated with chemical laboratories, particularly when working with hazardous substances. The fire potentially released chemicals from the fume cupboard, necessitating an
Audience	The audience is scientists with the focus is on the laboratory fire. The presentation uses scientific language such as "breathing apparatus crews" and "fume cupboard" and quotes expert opinions such as Mr Harold.
Reference Summary	Tests are being undertaken at the University of East Anglia (UEA) in Norwich to ensure there is "no leakage" of chemicals after a laboratory fire.
Result using QA-prompting	
Predicted	Firefighters responded to a third-floor laboratory fire at the University of East Anglia, requiring the evacuation of hundreds of students due to the risk of chemical inhalation.
ROUGE 1/2/L	47.05/24.28/35.29
Result using Vanilla Prompting	
Predicted	Tests are being undertaken at the University of East Anglia (UEA) in Norwich to ensure there is ""no leakage"" of chemicals after a laboratory fire.", Hundreds of people were evacuated from a campus building when firefighters were called to the scene of a fire on the third floor at about 10:30 BST. The fire service is now working with UEA staff to determine what chemicals were present in the laboratory. Twelve fire engines were dispatched to the incident and nobody was injured.
ROUGE 1/2/L	234

Table 7: An example of XSum dataset (Narayan et al., 2018) summarization using QA-prompting with Gemma-3-12B. Input contains the article, then, next 10 columns define answer generated for the given 10 questions. The last 3 columns contain reference summary, predicted summary, and ROUGE scores. GREEN represents questions that were included in the prompt. The results are with baseline prompting method is also given for comparison.

Field	Content
Input	System: 'Thank you for calling the Cambridge Information system. Your call will be recorded for research purposes. You may ask for information about a place to eat, such as a restaurant, a pub, or a cafe. How may I help you?' User: 'hello i am looking for a restaurant uh that sells only drinks and snacks' System: 'The Fleur, which serves drinks and snacks, is a great restaurant.' User: 'ok what is the address and phone number of the restaurant' System: 'The Fleur is located on 73 Humberstone Road and its phone number is 01223902183.' User: 'ok thank you good bye'
Topic	The main topic or focus of the content is a restaurant that sells only drinks and snacks.
Key Points	The key points or arguments presented in the text are as follows: 1. The text introduces the Cambridge Information system and its purpose of recording calls for research. 2. The system provides information about places to eat, such as restaurants, pubs, and cafes. 3. The user asks
Entities	The three main entities or individuals involved are the user, the system, and the restaurant. The user is the person who is interacting with the system and is looking for information about a restaurant. The system is the computer program that is providing the information to the user. The restaurant is the establishment that
Timeline	The timeline being discussed here is the timeline of the conversation between the user and the system. The system is providing information about a restaurant, and the user is asking for more details about the restaurant. The conversation is taking place in real-time, and the system is responding to the user's requests
Details	The supporting details, examples, or evidence provided are the following: - The Fleur, which serves drinks and snacks, is a great restaurant. - The Fleur is located on 73 Humberstone Road. - The Fleur's phone number is 0122
Conclusion	The system is designed to provide information about places to eat, such as restaurants, pubs, or cafes. The user asks for information about a restaurant that sells only drinks and snacks. The system recommends a restaurant called The Fleur, which serves drinks and snacks.
Tone	The overall tone or sentiment is positive.
Challenges	1. The system is not able to understand the user's request. 2. The system is not able to provide the user with the information they are looking for. 3. The system is not able to understand the user's request. 4. The system is not able to provide the
Insights	The system is able to understand the user's request and provide relevant information about a restaurant that sells only drinks and snacks. The system also provides the address and phone number of the restaurant, which is helpful for the user to locate and contact the restaurant.
Audience	The audience is the people who are looking for a restaurant, pub, or cafe. The content is aimed at people who are looking for a restaurant, pub, or cafe. The content is presented in a way that is easy to understand and follow.
Reference Summary	"You are looking for a drinks and snacks only restaurant. You don't care about the area. You want to know the address and phone number."
Result using QA-prompting	
Predicted	You are looking for a drinks and snacks only restaurant. You don't care about the area. You want to know the address and phone number.
ROUGE 1/2/L	100.0/100.0/100.0
Result using Vanilla Prompting	
Predicted	The article describes a system that can help users find information about a place to eat, such as a restaurant, a pub, or a cafe. The system can provide information about the address and phone number of the restaurant. The system is designed to be easy to use and can be accessed by anyone.
ROUGE 1/2/L	27.5/12.8/20.0

Table 8: An example of DSTC3 dialogue summarization dataset using QA-prompting with Mistral-7B. Input contains the dialogue, then, next 10 columns define answers generated for the given 10 questions. The last sections contain reference summary and evaluation metrics. GREEN represents questions that were included in the prompt. The results are with baseline prompting method is also given for comparison.

Field	Content
Input	Although this seems like a great product it is definitely not minced. The cans say it is minced but it isn't, it is ground beef. My cat ONLY likes minced. Not chunks, not ground. Although she will eat ground cat food sometimes, it has to be inviting and delicious, yet she just walks away from this. I hate it when the packaging lies about the contents of a product. The shipping was speedy and the value is great (as long as they deliver what they promise) now I am stuck with food my cat won't eat at all.
Topic	The main topic or focus of the content is the cat food.
Key Points	The key points or arguments presented in the text are: - The product is not minced as it is ground beef. - The packaging lies about the contents of the product. - The shipping was speedy and the value is great. - The cat won't eat the food.
Entities	The three main entities or individuals involved are the cat, the cat owner, and the company that produces the cat food. The cat is the one who is eating the food, the cat owner is the one who is buying the food, and the company is the one who is producing and selling the food. The cat
Timeline	The timeline being discussed here is the timeline of the product.
Details	The supporting details, examples, or evidence provided are: - The cans say it is minced but it isn't, it is ground beef. - My cat ONLY likes minced. - Although she will eat ground cat food sometimes, it has to be inviting and delicious
Conclusion	The conclusion is that the product is not minced.
Tone	The overall tone or sentiment of the text is negative.
Challenges	The text raises the question of whether the product is minced or ground beef.
Insights	The writer is not happy with the product and the packaging.
Audience	The audience is the cat owner who is looking for a minced cat food. The presentation is aimed at the cat owner who is looking for a minced cat food.
Reference Summary	Not minced, this is ground beef
Result using QA-prompting	
Predicted	Product advertised as minced but turned out to be ground beef, which the cat refuses to eat.
ROUGE 1/2/L	26.02/9.52/26.02
Result using Vanilla Prompting	
Predicted	Product mislabeled as minced, cat prefers minced.
ROUGE 1/2/L	15.38/0.0/15.38

Table 9: An example of Amazon Food review summarization using QA-prompting with Mistral-7B. Input contains the customer review, then, next 10 columns define answers generated for the given 10 questions. The last sections contain reference summary and evaluation metrics. GREEN represents questions that were included in the prompt. The results are with baseline prompting method is also given for comparison.

Author Index

- Achkar, Pierre, 102
Agarwal, Naaisha, 123
- Boyd-Graber, Jordan Lee, 1
- Callison-Burch, Chris, 183
Campbell, Yuri, 102
Cao, Juntao, 135
Carenini, Giuseppe, 17, 135
Cheung, Jackie CK, 48
Chng, Eng Siong, 31
- Deas, Nicholas, 172
- Ernst, Ori, 48
- Fayek, Haytham M., 74
Fu, Xue-Yong, 94
- Gollub, Tim, 102
Gupta, Mukur, 172
- Hao, Lim Zhi, 31
Huang, Nannan, 74
- Ji, Yangfeng, 59
Joty, Shafiq, 135
- Khasanova, Elena, 94
Kreuter, Anne, 102
- Lagasse, Ryan, 123
Laskar, Md Tahmid Rahman, 94
Li, Chuyuan, 17, 135
Li, Raymond, 17, 135
Li, Zongxia, 1
- McKeown, Kathleen, 172
- Mondal, Ishani, 1
Murray, Gabriel, 17
Murugaboopathy, Satiyooshan, 102
- Nenkova, Ani, 1
- Oomen, Oskar, 123
Orshansky, Daniel, 123
- Palmieri, Elizabeth, 59
Potthast, Martin, 102
- Rao, Delip, 183
- Saini, Harsh, 94
Sinha, Neelabh, 199
Siu, Alexa, 1
Subbiah, Melanie, 172
Surana, Tanmay, 31
Suresh, Sathya Krishnan, 31
- TN, Shashi Bhushan, 94
- Varimalla, Nikhil Reddy, 172
- Wei, Jiaqi, 135
Wong, Eric, 183
Wu, Xiyang, 1
- You, Weiqiu, 183
- Zeweniuk, Nathan, 48
Zhang, Haopeng, 157
Zhang, Xiang, 135
Zhang, Xiuzhen, 74