Embrace Divergence for Richer Insights: A Multi-document Summarization Benchmark and a Case Study on Summarizing Diverse Information from News Articles

Kung-Hsiang Huang[1*] **Philippe Laban**[2] **Alexander R. Fabbri**[2]
Prafulla Kumar Choubey[2] **Shafiq Joty**[2] **Caiming Xiong**[2] **Chien-Sheng Wu**[2]
[1]University of Illinois Urbana-Champaign [2]Salesforce AI Research
[1]khhuang3@illinois.edu
[2]{plaban, afabbri, pchoubey, sjoty, cxiong, wu.jason}@salesforce.com

Abstract

Previous research in multi-document news summarization has typically concentrated on collating information that all sources agree upon. However, the summarization of diverse information dispersed across multiple articles about an event remains underexplored. In this paper, we propose a new task of summarizing diverse information encountered in multiple news articles encompassing the same event. To facilitate this task, we present a data collection schema for identifying diverse information and curated a dataset named DIVERSESUMM. The dataset includes 245 news stories, with each story comprising 10 news articles and paired with a human-validated reference. Next, to enable consistent automatic evaluation, we conduct a comprehensive analysis to pinpoint the position and verbosity biases when utilizing Large Language Model (LLM)-based metrics for evaluating the coverage and faithfulness of summaries. Through correlation analyses, we outline the best practices for effectively using automatic LLM-based metrics on the DIVERSESUMM dataset. Finally, we study how LLMs summarize multiple news articles by analyzing which type of diverse information LLMs are capable of identifying. Our analyses suggest that despite the extraordinary capabilities of LLMs in single-document summarization, the proposed task remains a complex challenge for them mainly due to their limited coverage, with GPT-4 only able to cover under 40% of the diverse information on average.[1]

1 Introduction

In the realm of news reporting, each event is often chronicled by multiple sources, providing a rich tapestry of perspectives and insights. The sheer volume of articles available via news aggregators, as noted by Laban et al. (2023), can overwhelm readers, leading to fatigue (Lee and Chyi, 2015). This has fueled the demand for more digestible multi-source summaries. However, as highlighted by existing multi-document summarization studies (Over and Yen, 2004; Owczarzak and Dang, 2011; Fabbri et al., 2019), these often only reflect consensus information and neglect the breadth of differing viewpoints. To address this, we propose the **Multi-document Diversity Summarization (MDDS)** task, aimed at faithfully illuminating the diverse information presented in multiple sources.

Following Laban et al. (2022), we formalize diverse information as *questions and answers where numerous sources can answer the same question, and the corresponding answers extracted from different news articles exhibit a variety of opinions or perspectives*. For robust and objective evaluation, we opted for a QA representation for references, aligning with the granularity and reliability advantages emphasized in prior work on summarization evaluation (Krishna et al., 2023; Liu et al., 2023c; Arumae and Liu, 2019). An example of diverse information is shown in Figure 1.

Using this formulation, we propose a reference annotation methodology to identify and gather diverse information dispersed across multiple articles about the same story. Our approach is a pipeline based on GPT-3.5-Turbo (OpenAI, 2023a), which generates questions concerning the story likely to pull varied responses from different sources. The subsequent answers extracted from each news article are then clustered into groups. We employ a post-processing step that removes invalid questions and answers. Finally, all questions and answers are validated by human annotators. The resulting dataset contains 245 news story clusters, where each story contains 10 news articles and an average of 2.49 questions, with each question associated with 3.41 answer clusters on average. This dataset is named **DIVERSESUMM**.

We conduct a series of experiments to under-

[*]Work done while interning at Salesforce AI Research.
[1]The code and data have been made publicly available: https://github.com/salesforce/DiverseSumm.

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 570–593
June 16-21, 2024

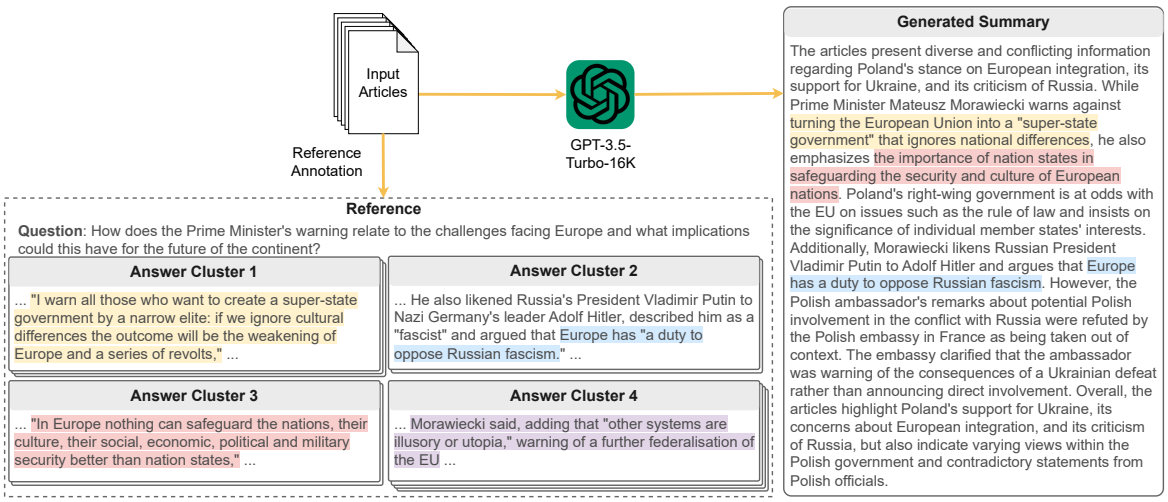

Figure 1: An example from our DIVERSESUMM dataset and a summary generated by GPT-3.5-Turbo-16K. To depict the process succinctly, only 4 news answer clusters from the reference are displayed. In this instance, the reference contains a single question with various answers extracted from each news article. In general, a news event may contain multiple reference questions, each of which can correspond to multiple answer clusters. The summary produced by GPT-3.5-Turbo-16K encompasses 3 of the answer clusters shown, but does not cover Answer Cluster 4.

stand the relevancy and challenges of our task in the era of LLMs and how future work should evaluate models on our task. Our fine-grained human evaluation results identify that even the most advanced LLM, GPT-4, only covers about 37% of diverse information with optimally designed prompts (see Appendix C.2). This highlights the significant challenge of effectively incorporating diverse information from multiple sources and the efficacy of our dataset as a rigorous LLM benchmark. Furthermore, we assess GPT-4 as an evaluator, given the impracticality of extensive human evaluations and its high correlation with human ratings (Liu et al., 2023b). Based on the correlation and bias analysis of GPT-4 evaluations, we provide recommendations for its application in assessing coverage and faithfulness of LLMs on our task. Our key findings are outlined in Table 1.

Our contributions are: (1) We introduce the Multi-document Diversity Summarization task that challenges models to identify diverse information across news articles and propose a reference annotation scheme to construct the DIVERSESUMM dataset. (2) We conduct extensive human evaluations to understand LLMs' ability to tackle our task and demonstrate that even GPT-4 struggle to achieve high coverage. (3) We conduct bias and correlation analysis on different GPT-4-based evaluation protocols to provide recommendations on using GPT-4-based metrics on our task. These guidelines are used to assess the coverage bias in various LLMs to understand how they summarize diverse information, highlighting the remaining challenges.

2 Task

The MDDS task revolves around a cluster of K news articles all centered around the same news event. To maintain a balance between task feasibility and challenge, we have opted to set K at a value of 10. The primary aim of our task is to generate a natural-language summary that effectively captures the diverse information presented within this cluster of news articles. To facilitate this process, our data collection pipeline, as elaborated in §3, produces references for each news cluster. These references take the form of question-answers (QAs), and their validity is established through human validation. The QAs must satisfy two properties: (1) the valid question must be answered by a sufficient number of sources, and (2) the answers associated with a valid question must present diverse opinions or perspectives.

In this work, the assessment of the generated summaries centers on two key facets: faithfulness and coverage. The faithfulness aspect evaluates the extent to which the summary aligns with the factual content present in the source articles. On the other hand, the coverage aspect gauges the inclusivity of information by considering how many answers within the reference are effectively addressed in the summary. We set our primary focus on these two aspects instead of other qualities, such as compression ratio and coherence, because recent work has shown that faithfulness and coverage

RQ1: How proficient are LLMs in summarizing diverse information from multiple news articles about an event?
- While LLMs can generate faithful summaries, they often lack adequate coverage. - Given the challenge of multi-document diverse summarization, our dataset serves as a rigorous benchmark for LLMs.
RQ2: What are the pitfalls and best practices when leveraging GPT-4 as the evaluation metric for our task?
- As a pairwise evaluator, GPT-4 shows a bias for the second summary. - Used as a single-answer grader, GPT-4 is prone to verbosity bias and prefers shorter summaries. - Likert-scale grading balances budget with correlation to human judgment for faithfulness evaluation. - Both granular evaluation methods correlate well with human judgment for coverage.
RQ3: Do LLMs exhibit coverage bias when performing MDDS?
- LLMs usually focus on summarizing the initial and final input articles, often overlooking the middle ones. - LLMs struggle to comprehensively address "How" and "What" type questions. - Long-context LLMs excel at covering frequent answers, while standard LLMs are proficient at summarizing infrequent ones. - Increasing model size improves LLMs' coverage of diverse information.

Table 1: Summary of research questions and key findings of our study.

are two major summarization challenges faced by models based on pre-trained transformers (Cao and Wang, 2021; Tang et al., 2022; Huang et al., 2023; Qiu et al., 2024).

3 Data Collection

This section details the DIVERSESUMM data collection pipeline, delineating its automated diverse information discovery from articles and the human validation stage that ensures data integrity.

3.1 Automatic Data Curation

Our data collection framework surfaces diverse information across news articles by asking questions about a news story, extracting answers from each news article, clustering the answers based on semantics, and filtering invalid questions and answers that are invalid. Our method extends the Discord Questions data generation pipeline (Laban et al., 2022) with four major modifications aimed at improving data quality:

(1) We perform question generation in a two-stage fashion, which increases the number of questions that result in diverse answers extracted from different articles. (2) Our question-answering component extracts answers from the context of the entire article, instead of extracting from each paragraph independently, significantly improving the recall of answers. (3) We perform a post-processing step to remove answers that do not make sense and QA-pairs that do not form diverse information. (4) Our method is based on GPT-3.5-Turbo [2], allowing for collection of higher-quality data.

Data Source We create DIVERSESUMM by gathering news stories and corresponding events from Google News, a news aggregator that collects news articles from various sources for a given news story. Each news story in Google News corresponds to around 40 news articles. We picked 400 news stories on the recent section of Google News. Most articles were published during March 2023, hence beyond the knowledge cut-off date of GPT-3.5-Turbo, which is September 2021.

Question Generation Upon collecting news stories, our next step is to ask questions about each news story that satisfy two properties: (1) *Availability of response*: this property ensures that any question deemed valid for the task should be one that many source articles can answer, hence indicating its centrality to the news event being reported. It is about the presence of answers across the corpus rather than their content. (2) *Diversity of answers*: this property focuses on the content of the responses rather than their presence. It stipulates that the answers to a valid question should exhibit a range of perspectives or opinions when extracted from different sources/articles. This is the heart of our approach to capturing the diversity of viewpoints represented in news articles.

We validate a query if at least 30% of the sources answer it and it results in assorted responses. To assess the efficiency of various methods of Question Generation (QG), we manually reviewed 10 news stories. We extend the Discord Question framework (Laban et al., 2022) by replacing their QG component with GPT-3.5-Turbo for its better performance over smaller models. For each news narrative, we heuristically select a medium-length article to prompt GPT-3.5-Turbo, generating 20 questions each, after which answers are extracted from all sources using the QA method outlined subsequently. The analysis reveals that of the 200 questions generated via this method, only 42 questions sufficiently cover all source articles, with

[2]We used the `gpt-3.5-turbo-0613` variant.

a mere 10 questions satisfy the two requirements mentioned above, indicating the single-article input's limited recall.

To enhance question coverage, we incorporate multiple representative articles into GPT-3.5-Turbo. We hypothesize that the answer clusters identified by a RoBERTa-based QA pipeline (Laban et al., 2022) provide a decent degree of diversity. Consequently, we identified representative articles through a heuristic method: a question corresponding to the median number of answer clusters was chosen. Within the associated articles, we opted for a medium-length article. This process produces a set of representative articles for the chosen questions corresponding to a news story. Prompting GPT-3.5-Turbo with these articles yielded 20 questions.

On a manual assessment of the aforementioned 10 news stories, this novel approach increased the number of questions linked with sufficient answers and valid questions, to 85 (+102.4%) and 19 (+90.0%), respectively. This indicates the proposed QG strategy's efficacy, significantly increasing the generation of valid questions compared to the prior method (Laban et al., 2022), and justifies our hypothesis mentioned in the previous paragraph.

Question Answering Similar to QG, we create an evaluation set for assessing the performance of question answering (QA) on our collected data, which contains two news stories, each paired with six human-generated valid questions. We compared various QA models, including a RoBERTa-based model (Liu et al., 2019) and two GPT-3.5-Turbo variants. One GPT-3.5-Turbo variant processes paragraphs independently, akin to RoBERTa, while its article-level counterpart extracts answers from the entire news article. Upon inspecting the outputs, we found that RoBERTa demonstrated higher precision, but the article-level GPT-3.5-Turbo variant excelled in recall (64.6%) against RoBERTa's (43.8%). Given the ease of filtering excessive answers compared to recovering missed answers, we opted for the article-level GPT-3.5-Turbo for all subsequent experiments.

Answer Consolidation For answer consolidation, we conduct a similar small-scale analysis to understand the performance of different answer clustering methods. We do not find significant advantages of the method based on GPT-3.5-Turbo compared to prior approaches; hence, we use the RoBERTa-based method (Laban et al., 2022) as our answer consolidation model.

Figure 2: Dataset statistics regarding the number of questions and answer clusters.

Post-processing To ensure task feasibility, we downsize the articles by selecting articles that have higher coverage of answers such that each news story is now associated with at most 10 articles. To expedite the process of human validation illustrated in §3.2, we utilized GPT-3.5-Turbo to filter non-sensical answers and non-diverse QA-pairs. Questions that are no longer associated with adequate answers due to the filtering are removed. Similarly, news stories that do not have any valid questions because of the filtering will be removed as well. The LLM prompts used in this subsection can be found in Appendix C.1.

3.2 Human Validation

To address any invalid QA-pairs that slipped past our post-processing procedure and enhance data quality, we recruited human annotators to validate the post-processed QAs. They are tasked to verify whether an answer addresses the corresponding question and ensure at least one article contains such an answer. More about this process is detailed in Appendix B.2. The resulting DIVERSESUMM dataset contains 245 news stories, each containing 10 articles. The distribution of the number of questions per news story and the number of answer clusters per question are shown in Figure 2. The distribution of question types and the topic of these news stories are shown in Appendix E.

4 Analysis

We address the research questions from §1, first evaluating how well diverse information from mul-

Model	Faithfulness (%)	Coverage (%)
Extract then summarize		
GPT-4	95.63	**36.58**
Vicuna-7B	78.42	13.36
Directly summarize		
GPT-3.5-Turbo-16K	**98.44**	35.66
LongChat-7B-16K	92.49	30.04

Table 2: Performance of different LLMs on our task. The faithfulness score and coverage score are determined by averaging the binary ratings provided by human evaluators.

tiple sources is summarized by LLMs (§4.1), then examining LLM behavior during this summarization (§4.3) using the most reliable LLM-based evaluation protocols we found (§4.2).

4.1 RQ 1: How proficient are LLMs in summarizing diverse information from multiple news articles?

To understand LLMs' performance on MDDS, we conduct human evaluation on summaries produced by four representative LLMs, GPT-4 (OpenAI, 2023b), GPT-3.5-Turbo-16K (OpenAI, 2023b), Vicuna-7B (Chiang et al., 2023), LongChat-7B-16K (Li et al., 2023).[3] *Long-context* LLMs, GPT-3.5-Turbo-16K and LongChat-7B-16K, handle texts up to 16K tokens and can perform direct summarization by taking all articles as input. *Standard LLMs*, GPT-4 and Vicuna-7B, are limited to 8K and 2K tokens, respectively; hence, we split summarization into two stages: selecting the most salient N sentences from each article and summarizing these sentences.[4] To elicit a high-coverage summary of diverse information, we manually optimize the prompts. Details of the prompts used for summarization in our experiments can be found in Appendix C.2. Following Krishna et al. (2023), we conduct evaluations at a finer granularity. Faithfulness is judged per sentence, whereas coverage is determined by how many reference QA pairs are covered by each summary. The resultant scores for each LLM were averaged from evaluations per summary sentence and reference QA pair, respectively. Evaluation details, such as worker qualification and user interface, are in Appendix B.3.

The human evaluation results are presented in Table 2. We observe that all four LLMs in general achieve high faithfulness but insufficient coverage of diverse information. This suggests that the proposed task is challenging even for state-of-the-art LLMs, and highlights that DIVERSESUMM serves as a challenging test bed for LLMs.

[3]We use `gpt-4-0613`, `gpt-3.5-turbo-16k-0613`, `vicuna-7b-v1.3` and `longchat-7b-16k`.

[4]We chose $N = 5$.

Aspect	First (%)	Second (%)	Consistency (%)
Coverage	1.63	**17.55**	60.10
Faithfulness	1.32	**13.27**	61.94

Table 3: Position bias analysis of swapping two summaries produced by two systems. Consistency is calculated as the percentage of cases in which the evaluator (i.e., GPT-4) provides coherent outcomes upon swapping the order of two summaries. First/Second indicates the percentage of cases in which a judge demonstrates a preference for the first/second summary. Overall, GPT-4 prefers the summary placed in the second position.

Aspect	Protocol	Original (%)	Extended (%)
Faithfulness	Single	41.44	20.58
	Pairwise	**0.20**	**0.00**
Coverage	Single	53.46	16.33
	Pairwise	**1.12**	**0.82**

Table 4: Verbosity bias analysis using GPT-4 as the evaluator. Single (i.e., single-answer grading) results in significant verbosity bias as we can see shorter summaries (i.e., Original) are preferable to longer summaries (i.e., Extended). Such bias can be significantly mitigated if pairwise comparison is used instead.

4.2 RQ 2: What are the pitfalls and best practices when leveraging GPT-4 as the evaluation metric for our task?

To facilitate the analysis and discussion of our next research question, we rely on LLM-based evaluation metrics to conduct various analyses, given their superior correlation with human judgments (Liu et al., 2023b) and the high cost of human annotation. For this research question, we aim to provide the best practices when using GPT-4 as the evaluator for the MDDS task by conducting bias and correlation analyses.

We focus on two major biases: position bias (i.e., whether the LLM evaluator favors certain positions over others) and verbosity bias (i.e. whether the LLM evaluator prefers shorter or longer texts). For all the experiments conducted in this analysis, we investigated summaries produced by GPT-4, GPT-3.5-Turbo, Vicuna-7B, and LongChat-7B-16K. The details of our prompts for the below experiments can be found in Appendix C.3.

Position Bias Position bias is most relevant to the pairwise comparison protocol. While previous work has shown that GPT-4 does exhibit

Criteria	Reference	Evaluated Texts	Rating Method	Evaluator	Rating	Correlation (%)
Faithfulness	Article	Summaries	Pairwise (both ways)	GPT-4	Win-Tie-Lose	**26.68**
	Article	Summary	Single-answer grading	GPT-4	Likert	<u>21.18</u>
	Article	Summary	Single-answer grading	GPT-4	Binary	18.54
	Articles	Summary	Single-answer grading	GPT-3.5-Turbo-16K	Likert	-7.44
	Articles	Summary	Single-answer grading	GPT-3.5-Turbo-16K	Binary	-3.70
	Articles	Summary sentence	Single-answer grading	GPT-3.5-Turbo-16K	Likert	15.58
	Articles	Summary sentence	Single-answer grading	GPT-3.5-Turbo-16K	Binary	-12.30
Coverage	QA pairs	Summaries	Pairwise (both ways)	GPT-4	Win-Tie-Lose	32.00
	QA pairs	Summary	Single-answer grading	GPT-4	Likert	**36.75**
	QA pairs	Summary	Single-answer grading	GPT-4	Binary	22.57
	QA pair	Summary	Single-answer grading	GPT-4	Likert	29.05
	QA pair	Summary	Single-answer grading	GPT-4	Binary	<u>35.83</u>

Table 5: Summary-level correlation between different LLM-based evaluation protocols and human judgments computed using Kendall's Tau. The best and second best protocol for each criterion are marked in boldface and underlined, respectively. The recommended evaluation protocols are highlighted.

position bias when used to assess text quality in conversational-focused tasks (Wang et al., 2023; Zheng et al., 2023), none of the prior studies have investigated whether such bias is also observed when evaluating faithfulness or coverage. To analyze position bias, we task GPT-4 with assessing a pair of summaries generated by two LLMs on which one is better, and then swap the positions of these two summaries and query GPT-4 again. We compute the percentage of times GPT-4 prefers the first or second summaries.

When GPT-4 compared pairs of LLM-generated summaries to evaluate faithfulness and coverage, a strong position bias surfaced, favoring the second entry (Table 3). Position bias was particularly pronounced when assessing similar-quality summaries (see Figure 23a). Hence, we deduce that **GPT-4 is unreliable when utilized as a pairwise evaluator in the MDDS task with respect to faithfulness and coverage**. Interestingly, this outcome contradicts Zheng et al. (2023), implying that **the position of bias for LLM-based evaluators could vary across different tasks**. A breakdown of the position bias analysis can be found in Appendix D.

Verbosity Bias To assess the verbosity bias of GPT-4 as an evaluator, we create extended summaries that maintain the semantic meaning. We achieve this by duplicating the original summaries, following Zheng et al. (2023). Ideally, a fair evaluator should provide identical faithfulness and coverage scores for both the original and extended summaries. We employed two experimental designs: pairwise comparison and single-answer grading on a Likert scale of 5.

The results of our verbosity bias analysis can be found in Table 4. We see that when **using the single-answer grading protocol, GPT-4 has a strong preference over shorter summaries, whether it is assessing faithfulness or coverage**. This conclusion was unexpected, particularly as we anticipated GPT-4 to favor longer summaries when determining coverage. Additionally, we noted that **verbosity bias is significantly lessened when using the pairwise comparison protocol**, which also comes with a much higher computational cost.

Correlation Analysis Upon examining the biases, we explore LLM-based evaluation protocols for their alignment with human judgments, varying reference granularity and rating models, including the use of GPT-3.5-Turbo-16K for efficiency in faithfulness assessment. For the pairwise comparison, since we had already established the prevalence of its significant position bias, we conducted the comparison both ways by swapping the summaries and then aggregating the results. As shown in Table 5, the both-way pairwise comparison protocol highly correlate with human judgment, mitigating verbosity and position biases, but was computationally demanding. In contrast, single-answer document-summary grading was efficient and fairly accurate. Notably, some GPT-3.5-Turbo-16K protocols negatively correlate with human assessment, indicating that **even though state-of-the-art long-context LLMs have a wide context window, their capacity to reason through extensive text effectively is occasionally unsatisfactory**.

In terms of coverage, we observed that both coarse-grained (QA-pairs) and fine-grained (single QA) evaluation protocols can establish a reasonably high correlation with human judgments provided we use appropriate rating methods (i.e., Likert scale for the former and binary rating for the latter). Either protocol proves suitable, contingent upon the level of granularity required for analysis.

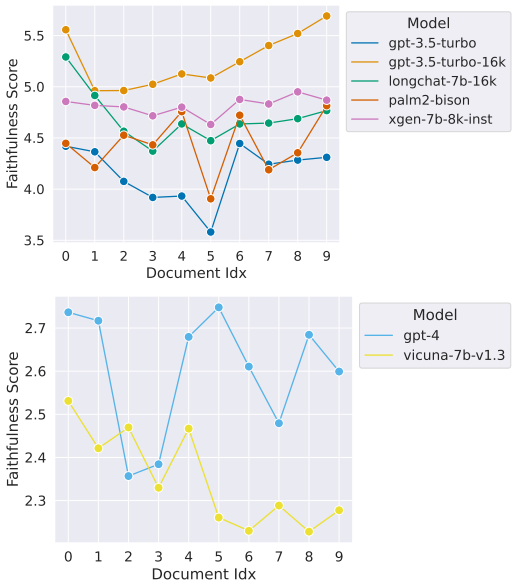

Figure 3: Faithfulness scores w.r.t. the index of the news article in the input prompt for LLMs. We see that LLMs with higher faithfulness (top), regardless of the way it summarize the article, tend to summarize from the starting or ending articles, while such a pattern is not observed for LLMs of low faithfulness (bottom).

Evaluation Recommendations For faithfulness evaluation, if budget is not a concern, it is recommended to use both-way pairwise comparisons given its high correlation with human judgments and least bias (The average cost for this evaluation protocol on our dataset is around $200 for each pair of models.). Otherwise, Likert scale single-answer grading with GPT-4 is the optimal alternative. For coverage evaluation, Likert scale single-answer grading has the highest correlation with human judgments.

4.3 RQ 3: Do LLM exhibit coverage bias when performing MDDS?

With the insights drawn from our analysis of the previous research questions, we are able to effectively conduct experiments to answer what type of information LLMs tend to summarize. We break down this research question into three sub-questions, with each focus on different aspects: focusing on article position, question type, and answer frequency. Since the evaluation is automatically conducted using GPT-4, we additionally consider the following LLMs for analysis: GPT-3.5-Turbo, XGen-7B-8K-Inst (Nijkamp et al., 2023), and Palm2-Bison (Ghahramani, 2023). The results are discussed in the following paragraphs.

Do LLMs tend to summarize articles at particular positions? The faithfulness score can serve as a measure to gauge how much content in an article's summary is drawn from each input news article. Higher faithfulness indicates greater information extraction from corresponding articles. We compute the faithfulness score between the generated summaries and each corresponding article using GPT-4 based on the article-summary Likert-scale single-answer grading protocol. In Figure 3, a prominent U-shape pattern for faithful LLMs (top) suggests that **faithful LLMs tend to summarize content from the first and last articles, while giving less attention to the middle articles**, aligning with findings from Liu et al. (2023a) on QA tasks. However, lower-faithfulness LLMs (bottom) show no clear pattern.[5]

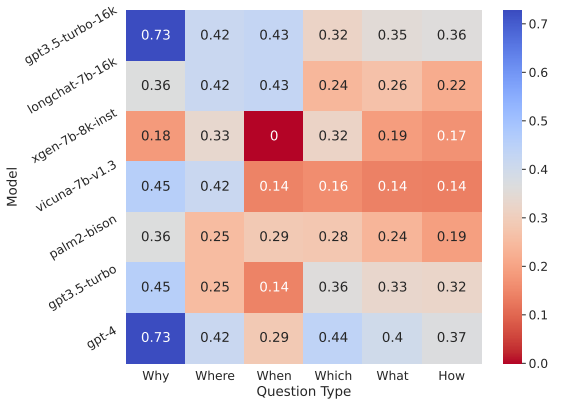

Figure 4: Average coverage scores with regard to different question types for different LLMs. Blue indicates a higher coverage, while red represents a lower coverage.

What diverse information do LLMs best identify and summarize? To understand categories of diverse information that LLMs are more inclined to summarize, we analyzed coverage by question type, with each binary coverage score mapping a summary to reference answers using GPT-4 with the QA-summary binary single-answer grading protocol. Then, we aggregate these answers based on the respective question types and calculate the averages, as depicted in Figure 4. Results show that questions starting with "Why" and "Where" tend to have better coverage, likely due to the direct presence of related answers in the source articles. Conversely, **LLMs encounter challenges in adequately covering answers for "How" and "What" type questions.** These question types delve into implications and require the model to establish connections between events, making them more intri-

[5] GPT-4's lower faithfulness scores arise from their summaries containing article indexes, which are not presented to the evaluators during the evaluation process.

Reference	Generated Summary
Question: Why do some people welcome the idea of "pumping the brakes" on AI development? **Answers**: ✓ bad actors will use the technology improperly ✓ the need for safety limits and regulation of AI tools	OpenAI CEO Sam Altman believes that artificial intelligence technology, such as the company's AI language model ChatGPT, has the potential to reshape society and improve our lives significantly. However, Altman also expresses real concerns about the dangers and misuse of AI. He worries that AI could be used for large-scale disinformation and offensive cyberattacks. Altman also highlights the importance of involving regulators and society in the development and regulation of AI systems...
Question: What are the broader implications of insulin prices on the overall healthcare system? **Answers**: ✗ ..., there is no guarantee that companies will continue to choose people over profits. ✗ Legally capping the price of insulin would bring a welcome financial relief to millions of Americans and save lives. ✗ ... policies to make diabetes treatment more affordable. ✗ ... it can hurt patients on fixed incomes and others less able to afford their insulin.	Multiple pharmaceutical companies, including Eli Lilly, Novo Nordisk, and Sanofi, have announced price reductions for their insulin products in response to growing pressure to lower the cost of this life-saving medication. The reductions range from 70% to 78% and are aimed at both Medicare patients and individuals with private insurance. The price cuts come as a result of the Inflation Reduction Act, which caps insulin prices for Medicare patients at $35 per month. However, some critics argue that these reductions only cover a portion of the insulin market and that individuals without insurance or with high-deductible plans may still struggle to afford the medication...

Table 6: Two instances in our DIVERSESUMM dataset and corresponding summaries generated by GPT-3.5-Turbo-16K. References and summaries are truncated due to space limits. The references in these two examples contain different types of questions. In the first instance, GPT-3.5-Turbo-16K successfully identifies the answers, demonstrating its proficient comprehension skills. However, in the second instance, the model fails to provide a high-coverage summary. This likely signifies its struggle with complex reasoning tasks that certain types of questions demand.

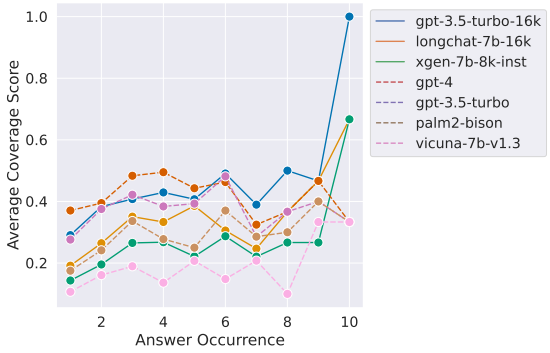

Figure 5: Average coverage scores with regard to answer frequency for different LLMs. Solid lines denote long-context LLMs, while dotted lines indicate standard LLMs. Answer occurrence represents the number of articles containing a given answer. For example, an answer occurrence of 10 means that all 10 input articles contain such an answer.

Model	Size	Coverage Score
Llama-2	7B	2.29
Llama-2	13B	2.53
Llama-2	70B	2.81
Vicuna-v1.5-16K	7B	2.00
Vicuna-v1.5-16K	13B	2.02

Table 7: Coverage score with regard to LLMs of varying sizes. The coverage scores are computed using the single-answer Likert-scale evaluation protocol with question-and-answer pairs as the reference.

cate to address. Two examples of different types of questions are demonstrated in Table 6.

Do LLMs have a tendency to summarize frequent information? We are intrigued by how the frequency of a piece of information influences the behavior of LLMs when summarizing multiple articles. Our data collection approach has facilitated this analysis, as answers extracted from each article have been systematically grouped. This enables us to easily determine the occurrence of a specific answer by calculating the number of articles containing that particular answer within its cluster. We compute the average coverage scores by aggregating answers based on their frequency of occurrence. The results, illustrated in Figure 5, reveal a notable trend: frequent answers (i.e., those found in a higher number of articles) tend to be covered more. Additionally, we found that **long-context LLMs exhibit greater proficiency in covering frequent answers, while standard LLMs appear to excel at summarizing infrequent answers**. This distinction is evident in the comparison between the performance of GPT-4 and GPT-3.5-Turbo-16K.

Does the size of LLMs correlate with their coverage of diverse information? To run this analysis, we need to ensure that factors other than the size of the model do not influence the results. Hence, we conduct experiments with LLMs in the same family. These include a family of *standard* LLMs, Llama-

2 (Touvron et al., 2023), with a maximum token length of 4K, as well as a family of *long-context* LLMs, Vicuna-v1.5-16K, which can handle up to 16K tokens. To measure the coverage scores, we utilized the evaluation protocol that shows the highest correlation with human judgment, as shown in Table 5. This consisted of a single-answer Likert-scale grading scheme, using question-and-answer pairs as the reference, and GPT-4 serving as the evaluator. As shown in Table 7, we found that increasing the model size enhances the coverage scores for both Llama-2 and Vicuna-v1.5-16K. This indicates that **more parameters improve LLM's ability to identify diverse information**.

5 Related Work

5.1 Multi-document Summarization

Conventional approaches to multi-document summarization (MDS) can be categorized into three types: extractive (Radev et al., 2000; Gillick and Favre, 2009; Lin and Bilmes, 2011; Hong and Nenkova, 2014; Peyrard and Eckle-Kohler, 2016; Cheng and Lapata, 2016; Narayan et al., 2018; Liu et al., 2018), abstractive (McKeown and Radev, 1995; Radev and McKeown, 1998; Barzilay et al., 1999; Zhang et al., 2018; Fabbri et al., 2019), and multi-sentence compression (Ganesan et al., 2010; Banerjee et al., 2015; Chali et al., 2017; Nayeem et al., 2018).

Recently, large language models (LLMs) have demonstrated significant advantages over conventional approaches in generating summaries of high faithfulness and quality. Studies have used LLMs to generate summaries of multiple documents by first extract important sentences from each article and then summarize them (Bhaskar et al., 2023) or iteratively improve summary quality with the guidance of a checklist (Zeng et al., 2023).

5.2 MDS Datasets

In previous studies, several popular MDS datasets have been examined. These datasets include DUC (Over and Yen, 2004; Dang, 2005) and TAC (Dang et al., 2008; Owczarzak and Dang, 2011), which are smaller in scale with approximately 50 and 100 article clusters, respectively. MULTINEWS (Fabbri et al., 2019) is the first large-scale MDS dataset in the news domain, containing 56K article clusters, with an average of fewer than 3 news articles per cluster. AUTO-HMDS (Zopf, 2018) is a multi-lingual MDS dataset focused on the Wikipedia domain, comprising 7.3K article clusters. WCEP (Gholipour Ghalandari et al., 2020) is another Wikipedia domain dataset, where each cluster may contain up to 100 articles. MULTI-XSCIENCE (Lu et al., 2020) and MS^2 (DeYoung et al., 2021) are two scientific domain MDS datasets. The above MDS datasets task models with summarizing consensus information, our work differentiates itself by focusing on summarizing diverse information across the articles.

6 Conclusion

We introduce a novel task of Multi-document Diverse Summarization that focuses on effectively summarizing diverse information from multiple news articles discussing the same news story. To facilitate this task, we construct a dataset, DIVERSESUMM, using our proposed QA-based pipeline. Through meticulous human evaluation, we have demonstrated that although LLMs exhibit a high level of faithfulness in tackling our task, achieving a high coverage rate remains particularly challenging, even with the most advanced LLM like GPT-4. This underscores both the challenges and opportunities of MDDS.

Furthermore, we have conducted an extensive analysis of bias and its correlation with human assessments across a range of evaluation protocols. Leveraging the insights obtained from these experiments, we propose a set of recommendations that outline the most effective protocols for evaluating model performance within our task domain. Our paper also delves into a comprehensive study that investigates LLMs' tendency to summarize various types of information. The outcomes of these analyses offer valuable insights into the behaviors exhibited by different LLMs when they engage with the challenge of summarizing diverse information. By presenting these resources and research findings, we hope to inspire and motivate future endeavors in the realm of comprehending and summarizing the intricate nuances present in diverse news articles concerning the same news event.

7 Ethical Considerations

In §3 and §4.1, we engaged annotators for data annotation and human evaluation. We prioritized fair compensation for our participants, with details provided in Appendix A. To foster an ethical working environment, we allowed participants to set their own pace, facilitated open communication

for any concerns, and provided the option to withdraw from the project at any time without repercussions. Additionally, we took measures to ensure the anonymity of the data annotations by avoiding the inclusion of any personally identifiable information.

8 Limitation

This study contributes significantly to the field of multi-document summarization by providing a larger and more comprehensive dataset than those available in previous research. However, there are several limitations that must be acknowledged.

Firstly, despite our best efforts to curate a large enough dataset, it still represents a relatively small fraction of the vast array of news content available online. This limitation is intrinsic to the task at hand, given the financial implications of human annotation and the complexity of multi-document summarization necessitates that annotators thoroughly read and understand multiple articles, which exponentially increases the time and cost associated with the annotation process compared to single-document summarization.

Moreover, while we carried out thorough LLM-based evaluations, we did not investigate the exact influence of different prompts on the LLM's performance. Even though we have tried our best to manually optimize the prompts, the lack of analysis on prompt sensitivity could lead to slightly different outcomes.

Furthermore, as our dataset encompasses online news articles, the study may not adequately capture the complexity of summarizing documents from diverse domains. News articles often follow a particular structure, which might not be prevalent in other kinds of multi-document contexts, such as academic papers or legal documents. Consequently, the generalizability of our findings and the utility of the dataset beyond the news domain demands further analysis.

References

Kristjan Arumae and Fei Liu. 2019. Guiding extractive summarization with question-answering rewards. 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 2566–2577, Minneapolis, Minnesota. Association for Computational Linguistics.

Siddhartha Banerjee, Prasenjit Mitra, and Kazunari Sugiyama. 2015. Multi-document abstractive summarization using ILP based multi-sentence compression. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pages 1208–1214. AAAI Press.

Regina Barzilay, Kathleen R. McKeown, and Michael Elhadad. 1999. Information fusion in the context of multi-document summarization. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 550–557, College Park, Maryland, USA. Association for Computational Linguistics.

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

Shuyang Cao and Lu Wang. 2021. CLIFF: Contrastive learning for improving faithfulness and factuality in abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6633–6649, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yllias Chali, Moin Tanvee, and Mir Tafseer Nayeem. 2017. Towards abstractive multi-document summarization using submodular function-based framework, sentence compression and merging. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 418–424, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 484–494, Berlin, Germany. Association for Computational Linguistics.

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

Hoa Trang Dang. 2005. Overview of duc 2005. In *Proceedings of the document understanding conference*, volume 2005, pages 1–12. Citeseer.

Hoa Trang Dang, Karolina Owczarzak, et al. 2008. Overview of the tac 2008 update summarization task. In *TAC*.

Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Lu Wang. 2021. MSˆ2: Multi-document summarization of medical studies. In *Proceedings of the 2021 Conference on Empirical Meth-*

ods in Natural Language Processing, pages 7494–7513, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. 2010. Opinosis: A graph based approach to abstractive summarization of highly redundant opinions. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 340–348, Beijing, China. Coling 2010 Organizing Committee.

Zoubin Ghahramani. 2023. Introducing palm 2. Google AI Research Blog.

Demian Gholipour Ghalandari, Chris Hokamp, Nghia The Pham, John Glover, and Georgiana Ifrim. 2020. A large-scale multi-document summarization dataset from the Wikipedia current events portal. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1302–1308, Online. Association for Computational Linguistics.

Dan Gillick and Benoit Favre. 2009. A scalable global model for summarization. In *Proceedings of the Workshop on Integer Linear Programming for Natural Language Processing*, pages 10–18, Boulder, Colorado. Association for Computational Linguistics.

Kai Hong and Ani Nenkova. 2014. Improving the estimation of word importance for news multi-document summarization. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 712–721, Gothenburg, Sweden. Association for Computational Linguistics.

Kung-Hsiang Huang, Siffi Singh, Xiaofei Ma, Wei Xiao, Feng Nan, Nicholas Dingwall, William Yang Wang, and Kathleen McKeown. 2023. SWING: Balancing coverage and faithfulness for dialogue summarization. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 512–525, Dubrovnik, Croatia. Association for Computational Linguistics.

Klaus Krippendorff. 1970. Estimating the reliability, systematic error and random error of interval data coded by several independent judges.

Kalpesh Krishna, Erin Bransom, Bailey Kuehl, Mohit Iyyer, Pradeep Dasigi, Arman Cohan, and Kyle Lo. 2023. LongEval: Guidelines for human evaluation of faithfulness in long-form summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1650–1669, Dubrovnik, Croatia. Association for Computational Linguistics.

Philippe Laban, Chien-Sheng Wu, Lidiya Murakhovs'ka, Xiang Chen, and Caiming Xiong. 2022. Discord questions: A computational approach to diversity analysis in news coverage. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5180–5194, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Philippe Laban, Chien-Sheng Wu, Lidiya Murakhovs'Ka, Xiang 'Anthony' Chen, and Caiming Xiong. 2023. Designing and evaluating interfaces that highlight news coverage diversity using discord questions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Angela M Lee and Hsiang Iris Chyi. 2015. The rise of online news aggregators: Consumption and competition. *International Journal on Media Management*, 17(1):3–24.

Dacheng Li, Rulin Shao, Anze Xie, Ying Sheng, Lianmin Zheng, Joseph E. Gonzalez, Ion Stoica, Xuezhe Ma, and Hao Zhang. 2023. How long can open-source llms truly promise on context length?

Hui Lin and Jeff Bilmes. 2011. A class of submodular functions for document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 510–520, Portland, Oregon, USA. Association for Computational Linguistics.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023a. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.

Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. *CoRR*, abs/1801.10198.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yixin Liu, Alex Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023c. 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.

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.

Kathleen McKeown and Dragomir R. Radev. 1995. Generating summaries of multiple news articles. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Ranking sentences for extractive summarization with reinforcement learning. 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 1747–1759, New Orleans, Louisiana. Association for Computational Linguistics.

Mir Tafseer Nayeem, Tanvir Ahmed Fuad, and Yllias Chali. 2018. Abstractive unsupervised multi-document summarization using paraphrastic sentence fusion. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1191–1204, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Erik Nijkamp, Tian Xie, Hiroaki Hayashi, Bo Pang, Congying Xia, Chen Xing, Jesse Vig, Semih Yavuz, Philippe Laban, Ben Krause, Senthil Purushwalkam, Tong Niu, Wojciech Kryscinski, Lidiya Murakhovs'ka, Prafulla Kumar Choubey, Alex Fabbri, Ye Liu, Rui Meng, Lifu Tu, Meghana Bhat, Chien-Sheng Wu, Silvio Savarese, Yingbo Zhou, Shafiq Rayhan Joty, and Caiming Xiong. 2023. Long sequence modeling with xgen: A 7b llm trained on 8k input sequence length. Salesforce AI Research Blog.

OpenAI. 2023a. Chatgpt.

OpenAI. 2023b. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Paul Over and James Yen. 2004. An introduction to duc-2004. *National Institute of Standards and Technology*.

Karolina Owczarzak and Hoa Trang Dang. 2011. Overview of the tac 2011 summarization track: Guided task and aesop task. In *Proceedings of the Text Analysis Conference (TAC 2011), Gaithersburg, Maryland, USA, November*.

Maxime Peyrard and Judith Eckle-Kohler. 2016. A general optimization framework for multi-document summarization using genetic algorithms and swarm intelligence. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 247–257, Osaka, Japan. The COLING 2016 Organizing Committee.

Haoyi Qiu, Kung-Hsiang Huang, Jingnong Qu, and Nanyun Peng. 2024. Amrfact: Enhancing summarization factuality evaluation with amr-driven training data generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*.

Dragomir R. Radev, Hongyan Jing, and Malgorzata Budzikowska. 2000. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In *NAACL-ANLP 2000 Workshop: Automatic Summarization*.

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

Xiangru Tang, Arjun Nair, Borui Wang, Bingyao Wang, Jai Desai, Aaron Wade, Haoran Li, Asli Celikyilmaz, Yashar Mehdad, and Dragomir Radev. 2022. CONFIT: Toward faithful dialogue summarization with linguistically-informed contrastive fine-tuning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5657–5668, Seattle, United States. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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

Qi Zeng, Mankeerat Sidhu, Hou Pong Chan, Lu Wang, and Heng Ji. 2023. Meta-review generation with checklist-guided iterative introspection. *CoRR*, abs/2305.14647.

Jianmin Zhang, Jiwei Tan, and Xiaojun Wan. 2018. Towards a neural network approach to abstractive multi-document summarization. *CoRR*, abs/1804.09010.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

Markus Zopf. 2018. Auto-hMDS: Automatic construction of a large heterogeneous multilingual multi-document summarization corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

A Are our findings in §4.2 still reproducible after a GPT-4 update every two months?

While it's a valid concern that the evolution of GPT models could impact the reproducibility of our findings, it's important to note that the principles highlighted in this research are not necessarily tied to the specific version of the GPT model itself, but rather how these language models work conceptually. The potential biases and evaluation techniques of GPT-4 we discuss can likely be applied or adapted to newer versions as well.

Naturally, with the release of an updated model, a new set of tests would be ideal to validate whether these findings hold. But this is true of any research in changing and evolving fields and does not detract from the value of our current findings. If anything, our research forms a foundation to more effectively assess future iterations of the GPT models in terms of evaluating coverage and faithfulness.

B Human Annotation

In this section, we illustrate the details of our human annotation process.

B.1 Worker Qualification

We established specific preliminary criteria for the recruitment of MTurk workers who possess strong performance histories. These criteria include having a HIT approval rate of 99% or higher, having approved a minimum of 10,000 HITs, and being located within the United Kingdom, Canada, and the United States.

Furthermore, apart from these preliminary criteria, eligible workers are required to pass three rounds of qualification tests centered around the faithfulness evaluation task, which is illustrated in Table 2. To streamline the qualification process, the authors manually annotate 3 HITs. Each HIT comprises ten news articles and four summaries generated by four different models. During each qualification round, annotators are presented with one of these annotated samples. Workers whose annotations do not exhibit a sufficiently high level of agreement with our annotations are excluded from the selection process.

Ultimately, 16 annotators who successfully passed all three rounds of qualification tests were selected. All the human evaluations and annotations are conducted by these 16 annotators. Additionally, every HIT has been meticulously designed to ensure that annotators can achieve an equivalent

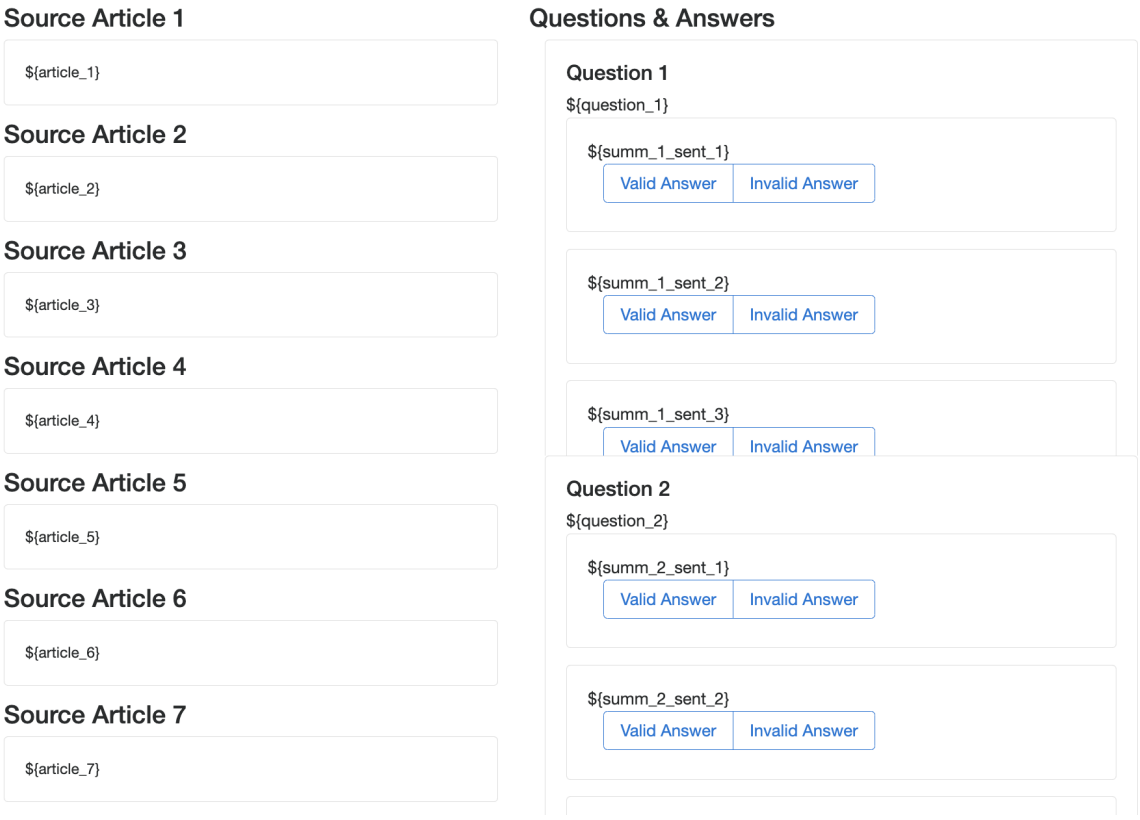

Figure 6: Annotation interface for filtering invalid QA pairs.

hourly pay rate of $20 provided they work continuously.

B.2 Annotating QAs

When annotating QA pairs, annotators are presented with the post-processed results detailed in §3.1. Below, we show the guidelines and the annotation interface presented to the annotators...

Guideline In this task, you will evaluate the validity of several answers with regard to the corresponding questions. To correctly solve this task, follow these steps:

- Carefully read the questions, answers, and the source articles.
- For each answer, check it against the question and the list of source articles.
- An answer is **Valid** if and only if (1) it **addresses the question, AND** (2) **at least one article contains such information** (It does not have to be word by word. It is sufficient that the information presented in the answer can be found in at least one article).

Warning: Annotations will be checked for quality against control labels, **low-quality work will be rejected**.

Valid answer: The validity depends on if the information in the answer is mentioned/supported by any source articles, not if the exact words are stated in the source articles. A valid answer should also provide a response that addresses the question it is paired with. Answer not addressing the question or suggesting no information should be marked as **Invalid Answer**. Examples of **Invalid Answer** are shown below:

- Question: What are the foreign impact of ...? Answer: The domestic influence of ...
- The article does not provide a clear answer to ...
- ... is not discussed in the article.
- As a language model, I cannot ...

Interface The annotation interface for filtering invalid QA pairs is presented in Figure 6.

B.3 Coverage Evaluation

Guideline In this task, you will evaluate the *coverage* of several statements with regard to the corresponding summaries. The statements are derived from news articles. To correctly solve this task, follow these steps:

- Carefully read the statements and the summaries.
- For each statement, check it against the corresponding summary.
- A statement is **Covered** if and only if it is mentioned or supported by the corresponding summary. (**It does not have to be word by word**. It is sufficient that the information presented in the statement can be found in the corresponding summary).

Warning: Annotations will be checked for quality against control labels, **low-quality work will be rejected**.

Covered Statement: The coverage depends on if the information in the statement is mentioned/supported by the corresponding summary, not if the exact words are stated in the corresponding summary. Some summaries may contain article number. Please ignore the article number and focus on whether the information in the statement is mentioned/supported by the corresponding summary.

Evaluation Interface The interface for coverage evaluation is shown in Figure 7.

B.4 Faithfulness Evaluation

Guidelines In this task, you will evaluate the *faithfulness* between each sentence of automatically generated summaries and a list of source articles used to generate the summaries. To correctly solve this task, follow these steps:

- Carefully read the generated summaries and the source articles.
- For each sentence, compare it against the list of source articles and decide if any of the source articles support this **sentence**.
- If there is at least one article that supports this sentence, rate the sentence as **Present**. Otherwise, select **Not Present**.

Warning: Annotations will be checked for quality against control labels, **low-quality work will be rejected**.

Faithfulness: The rating depends on if the information in the generated sentence is mentioned/supported by any source articles, not if the exact words are stated in the source articles Nonsense sentences should always be considered unfaithful, and you should select Not Present. Examples of these are shown below:

- As a language model, I cannot ...
- I am ready to summarize...
- Please provide the next set of news sentences...
- Sentence 1: 1: \n* n* 1: 1: 1: 1: 1:

Interface We display the interface for faithfulness evaluation in Figure 8.

B.5 Inter-annotator Agreement

We compute the quality of our annotations and evaluations using Krippendorff's Alpha (Krippendorff, 1970). For faithfulness and coverage evaluations, the inter-annotator agreement is 0.61 and 0.60, respectively. For reference annotations, the inter-annotator agreement is 0.69. These numbers represent a moderate to high agreement.

C LLM Prompts

In this section, we display all the prompts used in our experiments. Texts marked in boldface indicate placeholders.

C.1 LLM Prompts for Reference Annotation

Data collection pipeline consists of three components that are based on prompting ChatGPT: question generation, question answering, and post-processing. The prompt to each component is displayed in Figure 9, Figure 10, and Figure 11, respectively.

C.2 LLM Prompts for Summarization

We use different prompts for long-context and standard LLMs since the latter does not have long enough contexts to process all the input articles. The prompt template for long-context LLMs is displayed in Figure 13, while the two prompt templates for standard LLMs are shown in Figure 14 and Figure 15.

Note that the prompts displayed in the above-mentioned figures have undergone meticulous prompt engineering. We found that these prompts in general produce summaries with a higher coverage. In particular, we found that adding "Don't worry about the summary being too lengthy." in the prompt to GPT-4 is the key to generating more comprehensive summaries. As a comparison, we show our initial prompt to long-context LLMs in Figure 16, which is much shorter than the prompt in Figure 13. We use summary length to approximate coverage. As shown in Figure 12, the final prompt we used can significantly increase the length of the generated summaries.

C.3 LLM Prompts for Evaluation

In this section, we display the prompts to GPT-4 used in our evaluation.

D LLM Bias Analysis

In this section, we present the details of the bias analysis we conducted in §4.2.

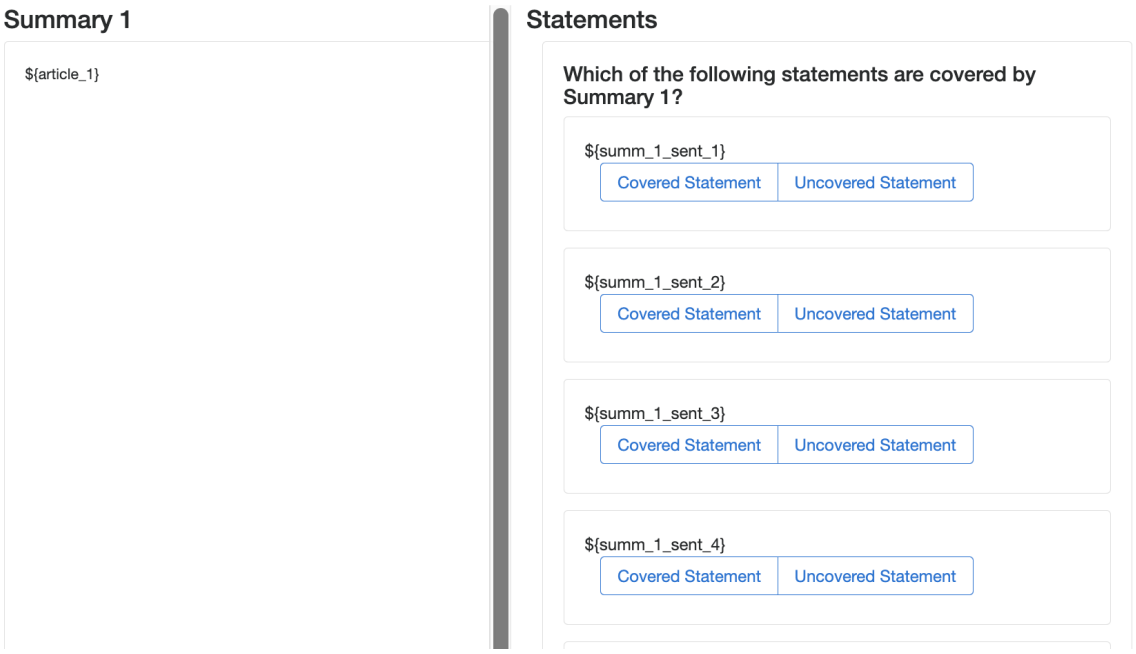

Figure 7: Interface for coverage evaluation.

D.1 Position Bias

As discussed in §4.2, position bias is most relevant to pairwise comparison. Figure 23 shows the breakdown analysis for coverage evaluation, while the faithfulness evaluation is displayed in Figure 24. In both coverage and faithfulness evaluation, the evaluator based on GPT-4 exhibits significant preference towards the second summaries placed in the inputs. In particular, we observe that position bias is most serious when the quality of two summaries is very similar (e.g. (a) in Figure 23).

D.2 Verbosity Bias

As illustrated in Table 4, pairwise comparison can significantly mitigate the verbosity bias. Hence, in the section, we only show the results for single-answer grading (see Figure 25). We see that the GPT-4-based evaluator prefers shorter summaries for all models, no matter when evaluating faithfulness or coverage. The result is surprising since we expect GPT-4 to prefer longer summaries when performing coverage evaluation.

E Topic and Question Distribution

Figure 26 and Figure 27 show the topic distribution and question distribution of our DIVERSESUMM dataset.

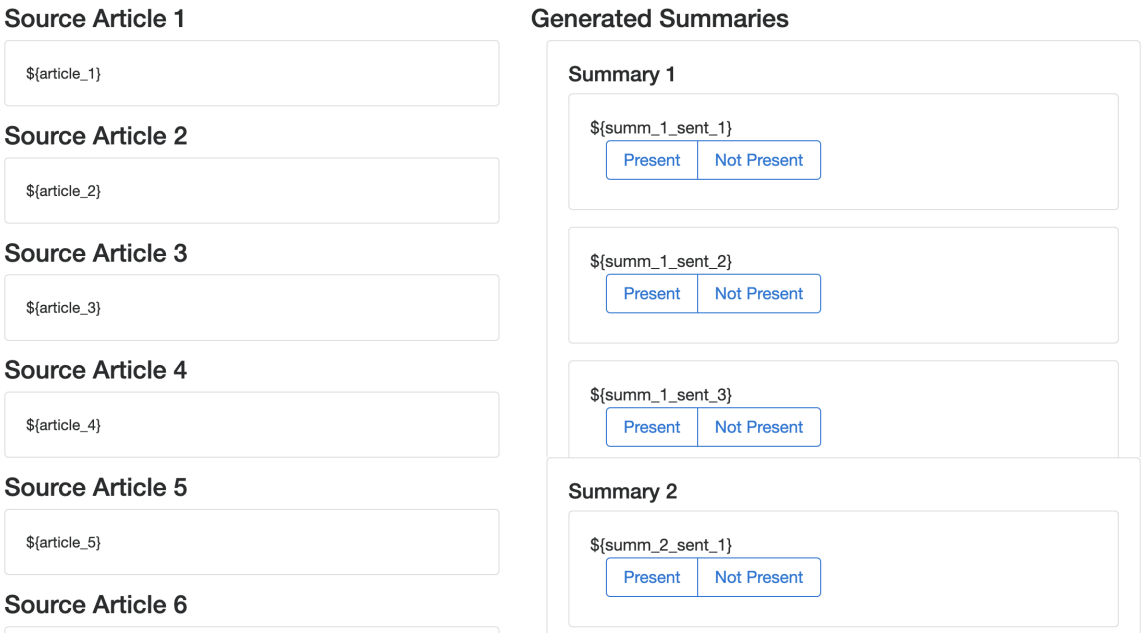

Figure 8: Interface for faithfulness evaluation.

[NEWS ARTICLES]
Given the above news articles. Complete the below two tasks:
Task 1: Write down 5 central factual questions for the news event that most sources will have likely answered. These questions, and their answer should relate the most important facts of the event. For example, for the US Presidential Election, the questions might be: Who won the election? What is the electoral college vote? What is the popular vote? What is the margin of victory? (each question should be up to 14 words)
Task 2: Write down 15 opinion or prediction questions for the news event that most sources will have likely answered in a unique way. These questions, and their answer should surface important points that news sources might analyze or present differently. For example, the questions might be: Who is more likely to win an election? Will there be a recession in 2023? What are the causes to the recession? (each question should be up to 14 words)
In your answer, specify the task number explicitly (Task 1, Task 2), and use line breaks between tasks, so that your report is structured.

Figure 9: The prompt for question generation.

Read the following news article and answer only the question '{question}'. Extract the exact sentence from the article changing up to 5 words. You should include ALL the answers that can be found in the article and must give your answers in a structured format: 'Answer 1: [extracted answer 1] \n Answer 2: [extracted answer 2] ...'. If the article contains no information to the given question, write: 'No Answer'.
==========
[ARTICLE]

Figure 10: The prompt for question answering.

[ARTICLES]
Read the above articles as well as the question and extracted answers below.
Task 1: Identify ALL the invalid answers that does NOT make sense or cannot be used to answer the question. You should specify the answer with their corresponding number: "Answer x: [answer x], Answer y: [answer y],...", where x and y are the number of the answer. If no such answer, then write down "Task 1: No invalid answers.".
Task 2: Identify ALL the answers that contradict with each other or form diverse information/opinion. These answers should not be invalid (i.e. should not be included in your responses for Task 1). You should specify the answer with their corresponding number: "Answer x: [answer x], Answer y: [answer y],...", where x and y are the number of the answer. If no such answer, then write down "Task 2: No diverse/conflicting answers.".
In your response, specify the task number explicitly (Task 1, Task 2), and use line breaks between tasks, so that your report is structured. The answer numbering in your response "Answer x: [answer x]" should correspond to the exact answer numbering and answer as shown below. Do not provide explanation for your response.
=======
Question: **[QUESTION]**
=======
Answers: **[ANSWERS]**

Figure 11: The prompt for post-processing.

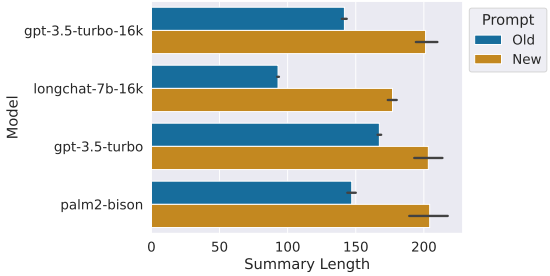

Figure 12: Lengths of summaries (token counts) produced by different models and different prompts. **New** indicates the final prompt we used, while **Old** denotes the initial prompt we tried.

Read the following news articles. Produce a summary that only covers the diverse and conflicting information across the following articles, without discussing the information all articles agree upon. Elaborate when you summarize diverse or conflicting information by stating what information different sources cover and how is the information diverse or conflicting. You must give your in a structured format: ```Summary: [your summary]```, where [your summary] is your generated summary.
==========
[ARTICLES]
==========
Remember, your output should be a summary that discusses and elaborates the diverse and conflicting information presented across the articles. You need to elaborate on the differences rather than only mentioning which topic they differ. Don't worry about the summary being too lengthy.

Figure 13: The prompt to long-context LLMs for direct summarization from all input articles.

Read the following news article. Extract the most important 10 sentences from the article and do not change words in the sentences. Your extracted sentence must be in a structured format: 'Sentence 1: [sentence 1] \n Sentence 2: [sentence 2] \n Sentence 3: [sentence 3] ...' where [sentence 1] should be the most important sentence.
==========
[ARTICLE]
==========

Figure 14: The prompt to standard LLMs for extracting important sentences from a given article.

Read the following sentences from different articles. Produce a summary that only covers the diverse and conflicting information across the following articles, without discussing the information all articles agree upon. Elaborate when you summarize diverse or conflicting information by stating what information different sources cover and how is the information diverse or conflicting. You must give your in a structured format: ```Summary: [your summary]```, where [your summary] is your generated summary.
==========
[EXTRACTED_SENTENCES]
==========
Remember, your output should be a summary that discusses and elaborates the diverse and conflicting information presented across the articles. You need to elaborate on the differences rather than only mentioning which topic they differ. Don't worry about the summary being too lengthy.

Figure 15: The prompt to standard LLMs for summarizing the extracted sentences.

Read the following news articles. Produce a summary that only covers the diverse and conflicting information across the following articles, without discussing the information all articles agree upon. Elaborate when you summarize diverse or conflicting information. You must give your in a structured format: ```Summary: [your summary]```, where [your summary] is your generated summary.
==========
[ARTICLES]
==========

Figure 16: The prompt to standard LLMs for summarizing the extracted sentences.

Please act as an impartial judge and evaluate the quality of the response provided by an AI assistant. Your evaluation should consider faithfulness of the summary with regard to the given article (i.e. whether the summary is factually consistent with the article).
Begin your evaluation by providing a short explanation. Be as objective as possible. After providing your explanation, please rate the response on as either 0 or 1 by strictly following this format: “[[rating]]”, for example: “Rating: [[0]]”. "[[1]]" indicates faithful, whereas "[[0]]" indicates unfaithful.

[Article]
[ARTICLE]

[The Start of Assistant Answer]
[SUMMARY]
[The End of Assistant Answer]

Figure 17: The prompt to GPT-4 for the binary single-answer grading faithfulness evaluation protocol.

Please act as an impartial judge and evaluate the quality of the response provided by an AI assistant. Your evaluation should consider faithfulness of the summary with regard to the given article (i.e. whether the summary is factually consistent with the article).
Begin your evaluation by providing a short explanation. Be as objective as possible. After providing your explanation, please rate the response on a scale of 1 to 10 by strictly following this format: “[[rating]]”, for example: “Rating: [[5]]”. "[[1]]" indicates lowest faithfulness, whereas "[[10]]" indicates highest faithfulness.

[Article]
[ARTICLE]

[The Start of Assistant Answer]
[SUMMARY]
[The End of Assistant Answer]

Figure 18: The prompt to GPT-4 for the Likert-scale single-answer grading faithfulness evaluation protocol.

Please act as an impartial judge and evaluate the quality of the summaries generated by two AI assistants to the user question displayed below. You should choose the assistant that follows the user's instructions and answers the user's question better.
Your evaluation should consider faithfulness of the summary with regard to the given article (i.e. whether the summary is factually consistent with the article). Begin your evaluation by comparing the two summaries and provide a short explanation. Avoid any position biases and ensure that the order in which the summaries were presented does not influence your decision. Do not allow the length of the summaries to influence your evaluation. Do not favor certain names of the assistants. Be as objective as possible. After providing your explanation, output your final verdict by strictly following this format: "[[A]]" if assistant A is better, "[[B]]" if assistant B is better, and "[[C]]" for a tie.

[Article]
[ARTICLE]

[The Start of Assistant A's Answer]
[SUMMARY1]
[The End of Assistant A's Answer]
[The Start of Assistant B's Answer]
[SUMMARY2]
[The End of Assistant B's Answer]

Figure 19: The prompt to GPT-4 for the pairwise comparison faithfulness evaluation protocol.

Please act as an impartial judge and evaluate the quality of the response provided by an AI assistant. Your evaluation should consider coverage of the summary with regard to the question and answers (i.e. how much information in the question and answers is covered by the summary). Begin your evaluation by providing a short explanation. Be as objective as possible. After providing your explanation, please rate the response on a scale of 1 to 10 by strictly following this format: "[[rating]]", for example: "Rating: [[0]]". "[[0]]" indicates insufficient coverage, whereas "[[1]]" indicates sufficient coverage.

[Questions and Answers]
[QAs]

[The Start of Assistant Answer]
[SUMMARY]
[The End of Assistant Answer]

Figure 20: The prompt to GPT-4 for the binary single-answer grading coverage evaluation protocol.

Please act as an impartial judge and evaluate the quality of the response provided by an AI assistant. Your evaluation should consider coverage of the summary with regard to the question and answers (i.e. how much information in the question and answers is covered by the summary). Begin your evaluation by providing a short explanation. Be as objective as possible. After providing your explanation, please rate the response on a scale of 1 to 10 by strictly following this format: "[[rating]]", for example: "Rating: [[5]]". "[[1]]" indicates lowest coverage, whereas "[[10]]" indicates highest coverage.

[Questions and Answers]
[QAs]

[The Start of Assistant Answer]
[SUMMARY]
[The End of Assistant Answer]

Figure 21: The prompt to GPT-4 for the Likert-scale single-answer grading coverage evaluation protocol.

Please act as an impartial judge and evaluate the quality of the summaries generated by two AI assistants. You should choose the assistant that follows the user's instructions and answers the user's question better. Your evaluation should consider coverage of the summary with regard to the question and answers (i.e. how much information in the question and answers is covered by the summary). Begin your evaluation by comparing the two summaries and provide a short explanation. Avoid any position biases and ensure that the order in which the summaries were presented does not influence your decision. Do not allow the length of the summaries to influence your evaluation. Do not favor certain names of the assistants. Be as objective as possible. After providing your explanation, output your final verdict by strictly following this format: "[[A]]" if assistant A is better, "[[B]]" if assistant B is better, and "[[C]]" for a tie.

[Questions and Answers]
[QAs]

[The Start of Assistant A's Answer]
[SUMMARY1]
[The End of Assistant A's Answer]
[The Start of Assistant B's Answer]
[SUMMARY2]
[The End of Assistant B's Answer]

Figure 22: The prompt to GPT-4 for the pairwise comparison coverage evaluation protocol.

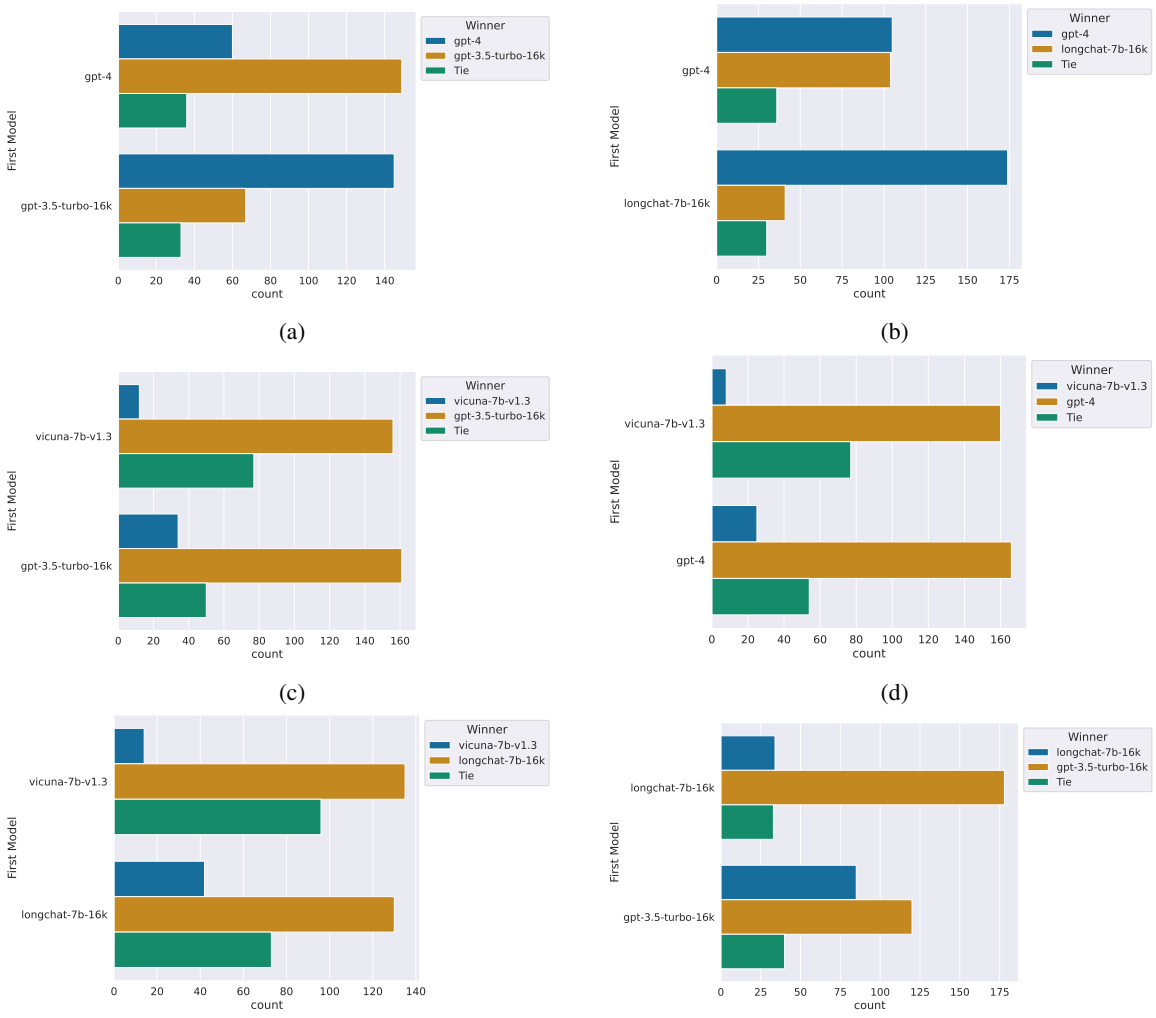

Figure 23: Position bias analysis on pairwise comparison protocols for coverage evaluation.

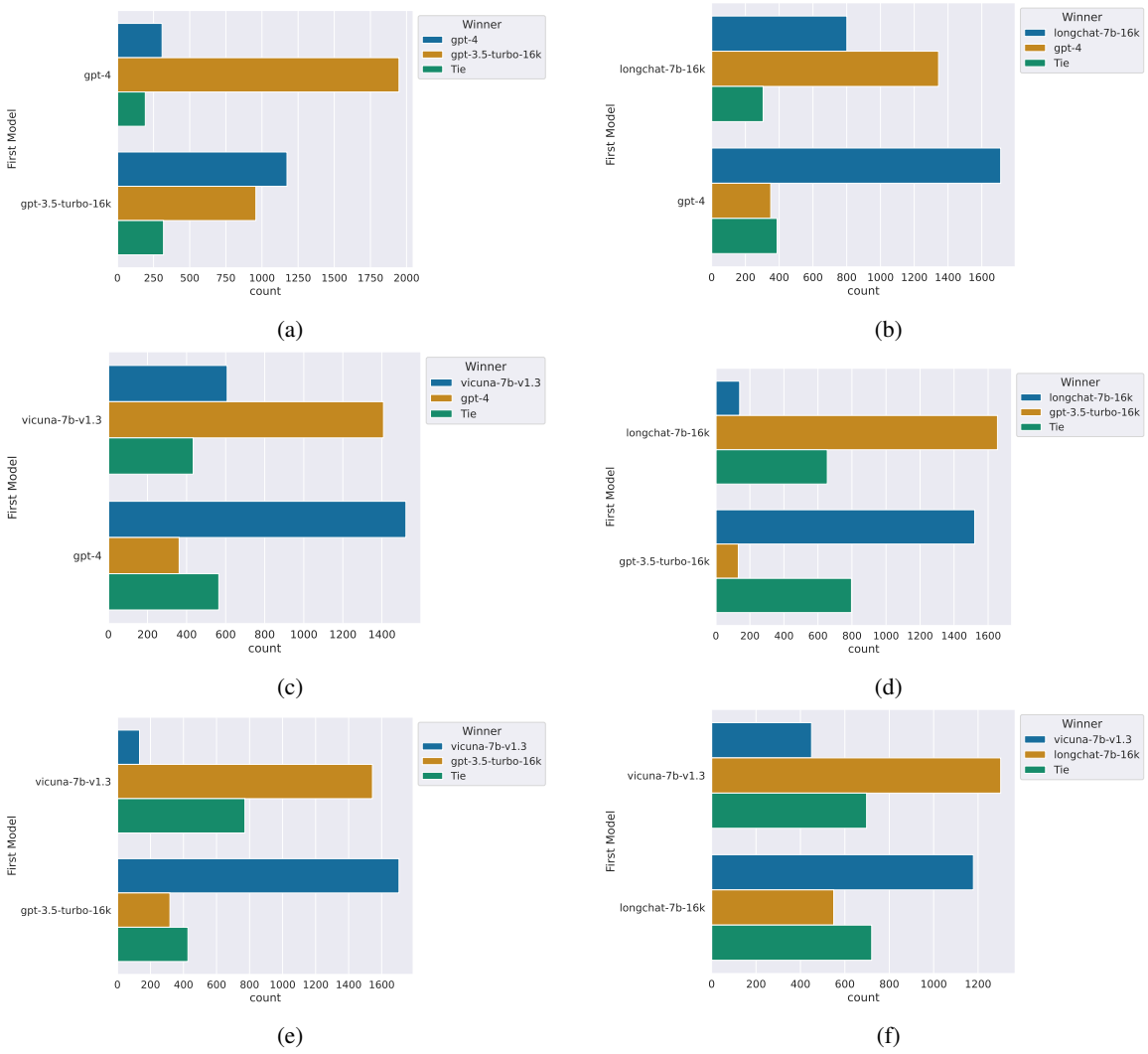

Figure 24: Position bias analysis on pairwise comparison protocols for faithfulness evaluation.

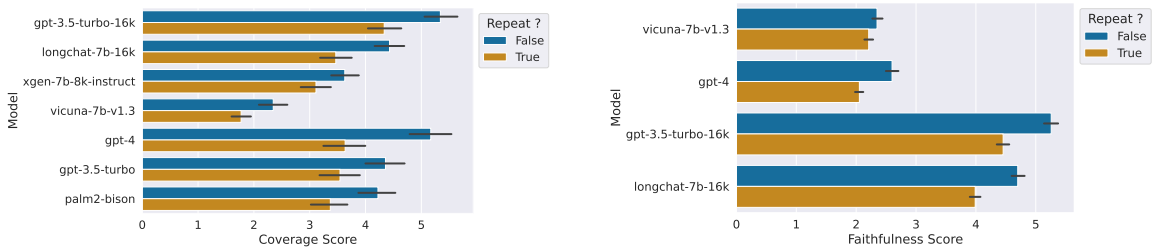

Figure 25: Verbosity analysis using the single-answer grading evaluation protocol. Repeat=False indicates the original summary, while Repeat=True denotes the summary is extended by repeating itself one time.

Figure 26: Word cloud representations of the topic distributions over our DIVERSESUMM dataset.

Figure 27: Question distribution of our DIVERSESUMM dataset.