On Positional Bias of Faithfulness for Long-form Summarization

David Wan[1*] **Jesse Vig**[2] **Mohit Bansal**[1] **Shafiq Joty**[2]
[1]UNC Chapel Hill [2]Salesforce AI Research
{davidwan,mbansal}@cs.unc.edu
{jvig, sjoty}@salesforce.com

Abstract

Large Language Models (LLMs) often exhibit positional bias in long-context settings, under-attending to information in the middle of inputs. We investigate the presence of this bias in long-form summarization, its impact on faithfulness, and various techniques to mitigate this bias. To consistently evaluate faithfulness, we first compile a benchmark of eight human-annotated long-form summarization datasets and perform a meta-evaluation of faithfulness metrics. We show that LLM-based faithfulness metrics, though effective with full-context inputs, remain sensitive to document order, indicating positional bias. Analyzing LLM-generated summaries across six datasets, we find a "U-shaped" trend in faithfulness, where LLMs faithfully summarize the beginning and end of documents but neglect middle content. Perturbing document order similarly reveals models are less faithful when important documents are placed in the middle of the input. We find that this behavior is partly due to shifting focus with context length: as context increases, summaries become less faithful, but beyond a certain length, faithfulness improves as the model focuses on the end. Finally, we experiment with different generation techniques to reduce positional bias and find that prompting techniques direct model attention to specific positions, whereas more sophisticated approaches offer limited improvements. Our data and code are available in `https://github.com/meetdavidwan/longformfact`.

1 Introduction

Large language models (LLMs) have enabled high-quality summary generation. However, the use of LLMs for long-context scenarios, where either the source document(s) or the generated summary is very long, still remains challenging (Chang et al., 2024; Kim et al., 2024a). A recent line of work has identified a problem with LLMs, positional bias, where models attend less to relevant information in the middle (Liu et al., 2024a). This “lost-in-the-middle” trend has been observed beyond long-form question-answering and for summarization (Ravaut et al., 2024), where LLMs do not utilize information from the middle of the documents.

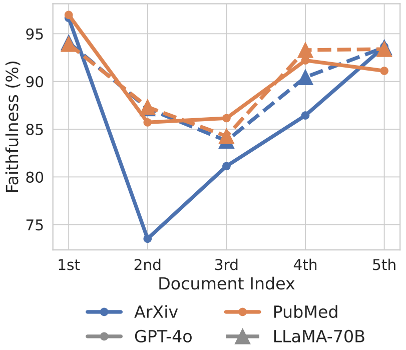

Figure 1: Positional bias in long-form summarization: On two representative models and datasets, summaries are less faithful to the documents in the middle.

One question that arises from such studies is: How does this affect faithfulness? Previous studies have shown that models hallucinate when they are uncertain in their responses (Cao et al., 2022; van der Poel et al., 2022). From the “lost-in-the-middle” finding, it can be inferred that the weak attention towards the middle context should also lead to hallucinations when generating content about that part. In this paper, we take a step further and analyze the relationship between faithfulness and positional bias for long-form summarization, as summarized in Table 1. We focus on the following research questions: (1) **What is the best configuration of LLM-based faithfulness metric for long-form summarization, and how does positional bias affect the metrics?** (2) **Are LLM-generated summaries prone to positional bias?** (3) **What methods can reduce positional bias?**

[*] Work done during an internship at Salesforce AI Research.

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
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

RQ1: How to measure faithfulness for long-form summarization?
- LLM-based metrics excel at using full context. - Taking the maximum faithfulness score over documents and average over summary sentences is the best merging strategy. - LLM-based metrics are sensitive towards different orders of documents.
RQ2: How faithful are the summaries with respect to the input documents? Are they prone to positional bias?
- Summaries show lead or lost-in-the-middle bias for faithfulness. - LLMs are sensitive towards perturbation of documents. - LLM exhibit a shift of focus with different lengths of the input source, summarizing information towards the end more faithfully when encountering long context.
RQ3: How to reduce such positional bias?
- Simply prompting the models to focus on the different locations of the documents is moderately effective. - Performing methods that changes the input structure (i.e., hierarchical merging or incremental updating) hurts faithfulness.

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

For our analysis of faithfulness in long-form document summarization, we first need to determine the best faithfulness metrics, as very few studies have performed an extensive study. Specifically, we want to verify whether current automatic metrics can evaluate summaries well given the large context. And if the metric needs to break the input context into chunks, what is the best way to merge the faithfulness scores. To do so, we collate a large, unified benchmark LONGFORMFACT for evaluating the performance of metrics on long-form summarization across 8 human-annotated benchmarks. We evaluate LLM-based metrics and find that the models handle full context well, and for the splitting case, taking the maximum over the source documents and taking the average across the summary sentences yield the highest correlation with human judgments. We further perform a perturbation experiment for the full-context setting, where we sort the documents according to their similarity with the summary. An ideal metric should not be affected by the order of the documents, but we find that the metrics are sensitive towards order perturbation and thus suffer from positional bias.

After determining the faithfulness metrics for long-form summarization, we use it to perform an extensive analysis of the faithfulness of the generated summaries across the input documents. Across 6 datasets, we generate summaries and plot the faithfulness scores to verify whether the model hallucinates more for the documents in the middle. Similar to Ravaut et al. (2024) that analyzes context utilization for summarization, we observe a U-shape and lead bias when analyzing the faithfulness of the generated summaries. Next, we perform a similar order perturbation experiment and find that LLMs are sensitive to the order, summarizing documents more faithfully for the documents at the beginning. Lastly, we analyze how length correlates with faithfulness by measuring how faithfulness changes as the input length increases. We find that models gradually introduce more hallucinations as we introduce more documents, and after a certain threshold, the model becomes more faithful as it attends to the documents at the end.

Finally, we investigate methods that attempt to mitigate positional bias. We explore methods such as prompting to focus on certain parts, hierarchical merging, incremental updating, and calibration methods. We find that prompting methods are moderately effective at improving summary's faithfulness towards certain positions, while more sophisticated methods struggle to address this issue.

2 Preliminaries

2.1 Long-form Summarization

We consider the task of summarization, where a model generates an m-sentence summary $S = \{s_1, s_2, ..., s_m\}$ from input document(s) D. For long-form summarization tasks, we follow Ravaut et al. (2024) and generally consider that the documents need to contain at least 2k tokens. For this task, such as multi-document summarization, the input D consists of n documents: $D = \{d_1, d_2, ..., d_n\}$. We refer to the boundaries of these documents as *natural document boundaries*, since there are no restrictions on how long each document may be. To unify different datasets, we can similarly split a single-document dataset, such as a scientific document, into different sections and refer to these sections as "documents." Alternatively, one may split the document into fixed-length chunks of words or tokens. To generate summaries,

Tasks	Num Ex.	Doc Split	Doc Words	Summ Split	Summ Words	ann. level	Models
MultiNews	90	3.4	767.2	7.1	175.0	summ	GPT-3.5, UniSumm, PEGASUS
QMSumm	90	-	1252.8	3.04	69.2	summ	GPT-3.5, UniSumm, PEGASUS
GovReport	147	-	2353.0	14.5	449.2	sent	PEGASUS, BART
PubMed	40	6.9	3299.2	10.4	195.0	sent	BART, BARTDPR
ArXiv	146	5.6	4805.6	6.4	164.9	sent	PEGASUS, BART
SQuALITY	40	-	5946.4	18.9	387.9	sent	BART, BARTDPR
ChemSumm	90	15.4	5974.5	7.2	197.7	summ	LongT5, PRIMERA
Diversesumm	377	10.0	7644.1	7.6	203.3	sent	GPT-4, GPT-3.5, Vicuna, LongChat

Table 2: Faithfulness meta-evaluation statistics. Ann. level indicates the granularity of the faithfulness annotation.

the entire input is truncated to fit the context window of each respective model.

2.2 Generating Long-form Summaries

In standard generation, the model M_g processes the documents D to produce the output summary S, as represented by $M_g(D) = S$. To help the model recognize document boundaries, a special indicator is usually inserted between each document; the most commonly used indicator is "====". Unless otherwise stated, we use this basic generation setup in most cases. In Section 5, we further explore other more advanced generation techniques.

2.3 Faithfulness Evaluation

For faithfulness evaluation, we consider entailment-based metrics that predict a binary faithfulness label given the document and the generated summaries. In the simplest form, we can make use of the full input $M_e(D, S) \in \{0, 1\}$. However, due to the prohibitive context length, many metrics that do not have such a large context window require either truncating the input—which loses information crucial for faithfulness evaluation—or splitting the task into evaluations of different document chunks and summaries. Thus, we evaluate both D and S in its more fine-grained form, $M_e(d_i, s_j)$ for the ith document chunk and jth summary sentence. Finally, to combine the scores, we explore three aggregation methods for both documents and summary sentences: Taking the maximum, minimum, or average $AGG \in \{max, min, mean\}$. Thus, the final score is $AGG_{S_{j=0}^{n}} AGG_{D_{i=0}^{m}} M_e(d_i, s_j)$.

3 Faithfulness Metrics Meta-Evaluation

Given the limited studies in determining the best automatic faithfulness metric for long-form summarization, we first aim to comprehensively test the best strategy for applying current evaluation methods in the long-form context.

3.1 LongFormFact

To better evaluate faithfulness metrics, we collate a large, unified benchmark consisting of eight long-form summarization datasets. We extend the effort by Zhang et al. (2024a) by including additional important long-form summarization annotation (Huang et al., 2024; Krishna et al., 2023). Statistics about the datasets are reported in Table 2. Similar to prior unifying efforts (Laban et al., 2022; Tang et al., 2024a; Zhang et al., 2024a), we convert different annotation schemes into binary faithfulness judgments. For Likert-based evaluations, we consider a summary to be faithful only if it receives the highest score. We describe the datasets below and include more details in Appendix A.1:

MultiNews (Fabbri et al., 2019) is a large multi-document news summarization dataset. Chen et al. (2023) collected 90 examples with Likert faithfulness scores at summary level.

QMSUM (Zhong et al., 2021) is a query-based, multi-domain meeting summarization dataset. Chen et al. (2023) similarly collected 90 examples with summary-level Likert faithfulness scores.

ArXiv (Cohan et al., 2018) is a summarization dataset of scientific articles. Koh et al. (2022) collected 146 examples by asking whether each summary sentence contains faithfulness errors.

GovReport (Huang et al., 2021) consists of long reports from government research agencies. Koh et al. (2022) collected 147 sentence-level annotations for faithfulness.

ChemSumm (Adams et al., 2023) is a scientific long-form summarization dataset in the chemical domain. The authors collected summary-level faithfulness annotations represented by binary labels.

PubMed (Cohan et al., 2018) is a scientific long-form summarization dataset in the medical domain.

Metric	Doc Merge	MN	QM	GR	PB	AX	SQ	CS	DS	Average
MiniCheck	Original	53.8	53.9	50.0	79.7	50.0	**83.0**	50.9	**55.2**	59.6
	Min	54.7	**55.7**	**55.1**	50.0	50.0	**83.0**	54.2	49.6	56.5
	Mean	**64.5**	**55.7**	**55.1**	52.6	53.4	**83.0**	54.2	46.0	58.1
	Max	56.6	**55.7**	**55.1**	**84.8**	**62.3**	**83.0**	**59.7**	49.3	**63.3**
GPT-4o	Full	52.2	**64.3**	**61.6**	**80.1**	**57.6**	76.7	**63.4**	**57.4**	**64.1**
	Min	50.0	60.7	60.0	54.4	52.1	**80.6**	54.2	50.3	57.8
	Mean	46.3	60.7	60.0	64.3	56.2	**80.6**	54.2	56.6	59.9
	Max	**60.0**	60.7	60.0	76.0	55.6	**80.6**	60.6	53.0	63.3

Table 3: BACC on meta-evaluation benchmarks, where MN refers to MultiNews, QM to QMSumm, PB to PubMed, GR to GovReport, AX to ArXiv, SQ to SQuALITY, CS to ChemSumm, and DS to Diversesumm. Min, mean, and max represent the respective operations to merge the document-level faithfulness labels, while full predicts faithfulness with all documents. For MiniCheck, we report the original method by the authors, which internally performs document chunking. We **bold** the best merging strategy for each metric and underline the second-best.

Krishna et al. (2023) collected sentence-level binary judgments of faithfulness. We use the FINE annotations, containing faithfulness judgments for each summary sentence.

SQuALITY (Wang et al., 2022) is a question-focused, long-document summarization dataset, where the documents are short stories from Project Gutenberg. Krishna et al. (2023) similarly collected sentence-level binary judgments of faithfulness. We use the FINE annotations.

DiverseSumm (Huang et al., 2024) is a multi-document news summarization dataset that focuses on conflicting information. The authors collected sentence-level binary faithfulness judgments.

3.2 Experimental Setup

Evaluation Strategy. Recognizing that models may exhibit positional bias, we experiment with assessing the faithfulness of the summary with respect to each document individually, as well as using the full input source. We also decompose the summary into individual sentences, a technique proven effective by Huang et al. (2024). We explore applying minimum, mean, and maximum aggregation methods over the documents. For summary sentence merging, we report mean, and report the result of different merging strategies in Appendix B. We use the *natural document boundaries* to separate the documents, and explore chunking the input into fixed number of tokens in Appendix B.1.

Evaluation Models. We primarily experiment using GPT-4o as the backbone LLM for the metric. We also explore the applicability of MiniCheck[1] (Tang et al., 2024b) – an automatic metric that has demonstrated efficacy comparable to GPT-4 performance – to the long-context setting. We note that MiniCheck by default splits the input into chunks of fixed number of tokens and takes the maximum over the documents, which we include as one of the baselines. Additionally, we present results using the Llama-3.1-8B model in Appendix B.2.

Metric. To account for class imbalance, we use balanced accuracy (BACC) to calculate metrics' correlations with human judgments.

3.3 Results

The main results are shown in Table 3. For MiniCheck, we observe that taking the maximum over the document achieves the highest correlations on average. This shows that this document aggregation method achieves the best results even for long-form summarization. Interestingly, the original strategy of taking the maximum over fixed input context performs on average 3.7% lower than using natural document boundaries, suggesting that only evaluating the relevant context is important. We also note that MiniCheck trails the strongest GPT-4o-based metric by only 0.8%, while matching the performance of GPT-4o-based metric with the same aggregation method.

When looking at the GPT-4o-based metric, using the full context performs the best, achieving the highest accuracy in 5 out of 8 cases and second-best accuracy in 2 of the remaining 3 cases. This suggests that LLMs can utilize long context effectively for evaluation. The second-highest ranking evaluation strategy is still merging the documents by taking the maximum over the documents, performing on average only 0.3% lower than using the full context, aligning with MiniCheck results and previous findings on the best merging strategy.

[1] We use *Bespoke-MiniCheck-7B*.

Dataset	Random?	Document order				Sensitivity
		Original	Top	Middle	Bottom	
ArXiv	×	**57.6**	56.6	55.1	57.0	2.5
ChemSumm	×	**63.4**	61.6	57.4	59.7	6.0
PubMed	×	80.1	83.2	83.9	**85.5**	5.4
MultiNews	✓	52.2	**60.0**	56.7	56.7	7.8
DiverseSumm	✓	**57.4**	55.0	55.4	56.5	2.4
Avg. Sensitivity	-	-	3.2	3.8	3.0	-

Table 4: BACC using GPT-4o-based metric when order of the documents are perturbed. 'Random' indicates whether the initial document orders are random.

Perturbed Document Order. In addition to standard meta-evaluation, we also perform an analysis by ordering the documents in terms of importance. Specifically, for datasets where document boundaries exist, we calculate the importance of each document relative to the model-generated summary using sentence similarity.[2] We then order the documents into top (beginning), middle, and bottom (end), corresponding to the placement of the most important documents. To illustrate, assume there are five documents with importance ranks of 1, 3, 2, 5, and 4, where rank 1 denotes the most important and rank 5 the least. The "top" ordering would sort documents by importance (e.g., 1-2-3-4-5), the "bottom" ordering would prioritize the least important documents (e.g., 5-3-4-2-1), and the "middle" ordering would reflect a mid-tier arrangement (e.g., 4-2-1-3-5). Note that the only change from the regular case is the document order, which should have no effect for MultiNews and DiverseSumm, where the documents are in random order, but may have an effect for the other datasets where it breaks the natural flow of the document.The results are presented in Table 4. In addition to BACC, we also include sensitivity, defined as the maximum difference between scores computed using the original ordering and those with different orderings.

Overall, we find that the metric is sensitive to document order, with high sensitivity observed across each dataset and reaching a 7.8% difference for MultiNews. When comparing only the top, middle, and bottom orderings, we find no clear trend across datasets; however, on average, the sensitivity is highest when the important document is placed in the middle. This indicates that the metric achieves the lowest BACC when the important document is in the middle, whereas placing it at the top results in the smallest difference, suggesting that the LLM has a stronger lead bias, i.e. performing better when the important documents are at the beginning of the input. Therefore, it may still be beneficial to use the metric that evaluates each document individually and aggregates the results via the maximum operation to reduce positional bias. This approach achieves a similar BACC compared to the full-context setting while inherently not being sensitive to input order.

Tasks	Doc Split	Doc Words	Summ Split	Summ Words
MultiXScience	5	804.9	6.9	186.0
PubMed	5	2850.3	7.4	190.1
MultiNews	5	4925.5	8.5	215.3
ArXiv	5	5825.5	7.1	181.6
DiverseSumm	10	7561.5	16.0	452.5
SummHay	100	87913.1	7.5	52.2

Table 5: Statistics for the generated summaries.

Takeaway. We demonstrate that while the GPT-4o based metric is able to utilize the full context, the model exhibits a "lost-in-the-middle" behavior when using an LLM as the metric. Therefore, we recommend evaluating each document individually and taking the maximum faithfulness score.

4 Faithfulness of Long-form Summaries

Next, we evaluate the faithfulness of summaries generated from different datasets, and perform detailed faithfulness analysis, including assessing faithfulness across each document, performing a perturbed analysis in which we sort documents by importance, and understanding how faithfulness changes as the number of documents increases.

4.1 Experimental Setup

Datasets. We include two representative multi-document summarization datasets, MultiNews (Fabbri et al., 2019) and MultiXScience (Lu et al., 2020); two long-form summarization datasets, ArXiv and PubMed (Cohan et al., 2018); and two recent summarization datasets with extremely long contexts, DiverseSumm (Huang et al., 2024) and SummHay (Laban et al., 2024). For ArXiv, PubMed, MultiNews, and MultiXScience, we randomly sample 100 examples from the validation set, each consisting of five documents or sections. For DiverseSumm, we use all original 10 documents and randomly sample 100 examples. The dataset statistics are shown in Table 5.

Models. To comprehensively evaluate positional bias across a range of models, we run GPT-3.5, GPT-4o (OpenAI, 2024), Llama-3.1-8B and

[2] We use SentenceTransformer (Reimers and Gurevych, 2019) with the *all-mpnet-base-v2* model.

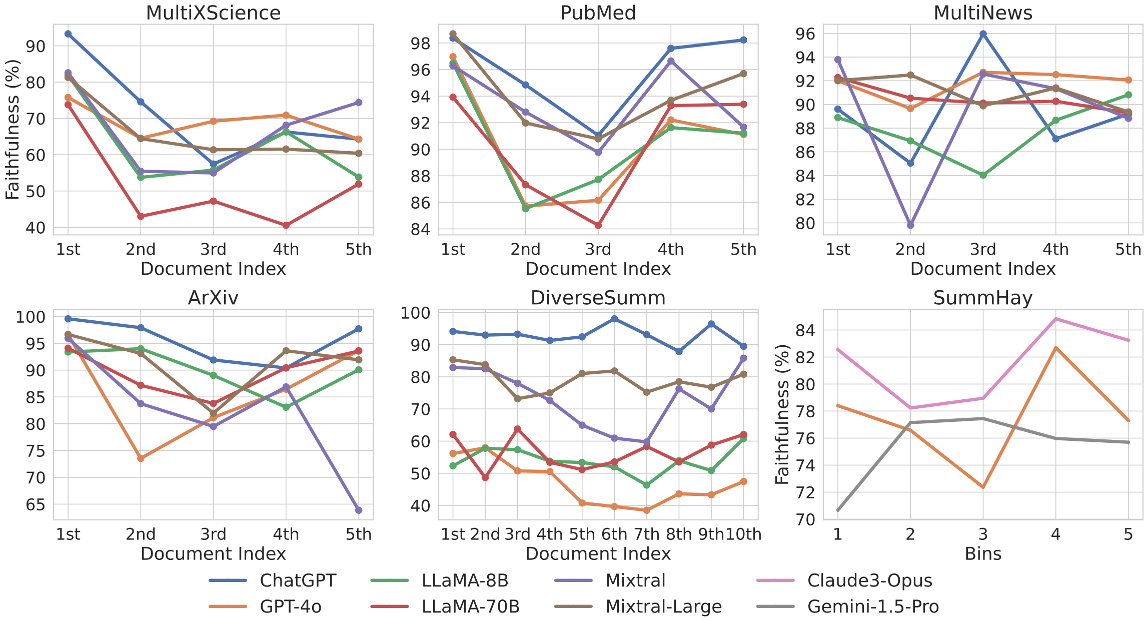

Figure 2: Faithfulness analysis across different positions of documents.

70B (Dubey et al., 2024), and Mixtral-7×8B and Mixtral-8×22B (Jiang et al., 2024). For SummHay, due to the high computational cost, we reuse the provided generated summaries from GPT-4o, Claude-Opus, and Gemini-1.5-pro.

Evaluation Metrics. Although GPT-4o demonstrates better faithfulness performance in Section 3.3, we choose MiniCheck for its nearly comparable performance, as well as its greater efficiency and lower cost. In Appendix C.6, we show evaluating with GPT-4o-based metric on a small subset, which exihibits similar trends.

4.2 Faithfulness Analysis

Our main analysis evaluates whether the summary is more faithful to documents in certain positions. To do so, we calculate the Minicheck faithfulness score with respect to all documents individually. We note that if we directly report the summary's faithfulness of each document, it may not accurately reflect faithfulness, as it is also confounded by *coverage*, i.e., how much the summary draws on content from each document. In fact, Huang et al. (2024) use the faithfulness score per document to measure coverage bias, and we provide an analysis of this coverage in Appendix C.5 as well as an analysis of document content overlap in Appendix C.1. To remove the effect of coverage, we consider *attribution*; that is, determining which document each sentence of the summary is discussing. For each summary sentence, we take the maximum faithfulness score over all documents. This is, in fact, the same process as one of the best document merging strategies reported in Section 4.4.

To illustrate, assume we have as input five documents with no overlap, and a summary consisting of five sentences, each sentence perfectly faithful to one of the documents and unfaithful with respect to the others. If we were to calculate the faithfulness score for each document separately, this summary would appear to be only 20% faithful towards each document (since only one sentence is faithful and the other four are not), as the score is misconstrued by coverage. However, if we only take the maximum faithfulness score over the documents, we remove the coverage effect and show that the summary is indeed faithful towards all documents.

In Appendix C.4, we demonstrate that taking the maximum over the faithfulness scores as attribution also exhibits the same trend as using SuperPAL (Ernst et al., 2021), a document-summary sentence alignment method that achieves the best alignment for long-form summarization (Krishna et al., 2023).

Results. We present the results in Figure 2. Generally, we observe that the models exhibit a dominant U-shaped curve, particularly on ArXiv, PubMed, MultiNews, and MultiXScience, where the middle documents are less faithful than the first or last documents. While most models exhibit this trend, Mixtral shows a more pronounced lead bias, especially on ArXiv. The trend for DiverseSumm is interesting, as it is more of a linear trend. This

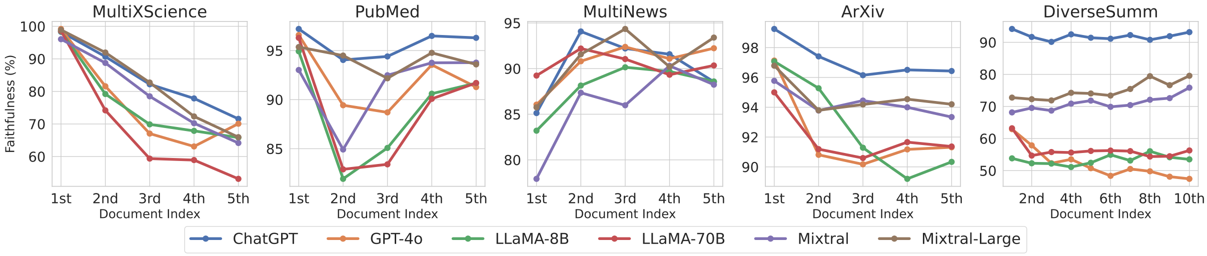

Figure 3: Faithfulness analysis by increasing the context length by adding documents.

Dataset	random?	Document order				Sensitivity
		Original	Top	Middle	Bottom	
ArXiv	✗	91.3	91.5	**92.2**	91.7	0.9
PubMed	✗	91.3	92.6	**93.4**	92.2	2.1
MultiNews	✓	**92.2**	90.0	89.0	90.6	3.2
MultiXScience	✓	**70.1**	67.6	62.1	68.2	8.0
SummHay	✓	73.5	62.7	66.5	**84.6**	11.1
Avg. Sensitivity	-	-	3.4	4.3	3.2	-

Table 6: Faithfulness score when perturbing the document order. ‘Random’ indicates whether the initial document orders are random.

is partly due to the nature of the task, where we explicitly ask to perform synthesis across the documents, which helps the model to focus on different parts. On SummHay, although the three models behave differently, we observe that GPT-4o and Claude yield lower faithfulness scores for the middle document. Interestingly, Gemini is the only model that exhibits the reverse trend, performing better on the middle documents; however, its average faithfulness score is not as high as that of the other two models. Appendix C.3 includes further discussions on variance across tasks.

4.3 Perturbed Input Analysis

Next, we conduct a similar perturbation analysis with GPT-4o as described in Section 3.3, where we reorder the documents according to importance. Here, importance is determined using the similarity between each document and the reference summary. We exclude DiverseSumm, as it does not contain reference summaries. The results are reported in Table 6. We observe a similar trend when we analyze the perturbation for the metric: The models generally generate the most faithful summaries either with the original order or when the important document is placed at the front. When looking at the sensitivity across each ordering, the middle case has the highest sensitivity. Interestingly, we observe the bottom ordering achieves a score 21.9 points over the top case for SummHay. We posit that this is because of how models handle long contexts, i.e., focusing towards the end when the context increases, which we confirm in the subsequent analysis on the correlation between increasing context length and faithfulness.

4.4 Faithfulness and Length Correlation

So far, our analyses have been post-hoc, where we attempt to analyze the faithfulness scores when the input is fixed. Here, we try to analyze how faithfulness correlates with length by incrementally increasing the number of documents. For all datasets, we start with summarizing one document, and then we add one more document and summarize again. We then calculate the faithfulness scores of the generated summaries using the corresponding set of documents. We exclude SummHay here, as it is computationally expensive to run the subsets for all 100 documents incrementally.

Results are in Figure 3. In MultiXScience, which contains relatively few words per document, we observe a strong lead bias. However, when moving to long-form summarization on datasets such as Pubmed and ArXiv, we observe a U-shaped trend: summary faithfulness decreases as the number of documents grows, then begins to improve beyond a certain threshold. This suggests that the model may “switch modes” at a particular context length and focus more on documents introduced later. On MultiNews, a similar pattern appears, as faithfulness steadily increases for most models with additional input documents. These observations are consistent with prior research on LLMs’ behavior with extremely long inputs, where models concentrate much of their attention on the later sections (Kim et al., 2024b; Laban et al., 2024).

5 Methods for Reducing Positional Bias

Finally, we explore different generation methods and verify whether they can reduce positional bias. We use the same sets of examples used in Section 4.

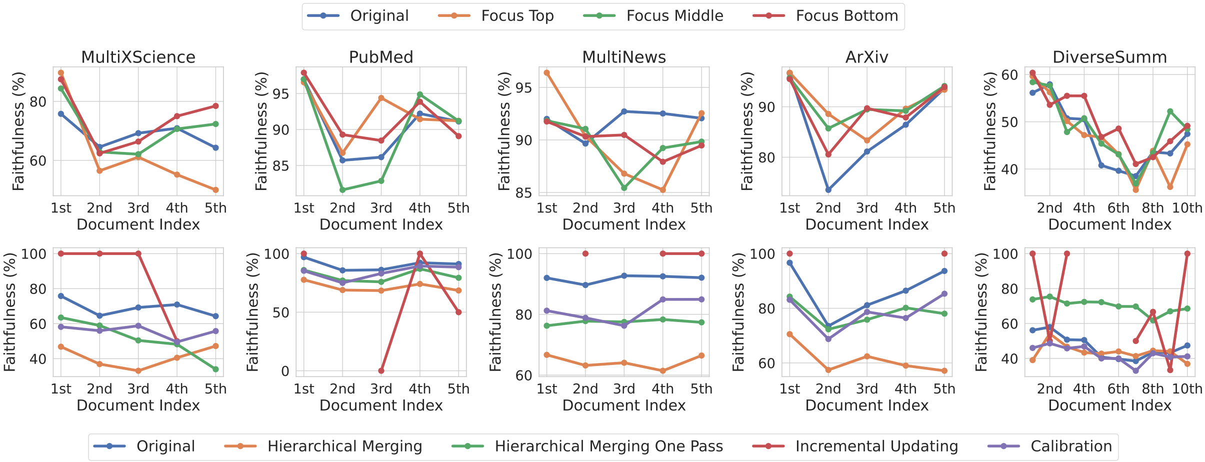

Figure 4: Faithfulness analysis for different summarization techniques.

5.1 Methods

Focus Prompt. We ask the model to focus on different parts of the input. We append the following instruction to the original prompt: "Focus on the top documents", "focus on the bottom documents", and "focus on the documents in the middle".

Hierarchical Merging. Explored in Chang et al. (2024), this method generates one summary for each document, and then iteratively merge the summaries. We explore both merging two summaries at a one until one final summary is produced, as well as one pass over all the individual summaries.

Incremental Updating. Similarly explored in Chang et al. (2024), starting with summarizing the first document, the model updates its output given the current summary and the next document.

Calibration. Past studies have found that calibrating the model specifically for the positional bias can reduce such bias. For both open-source and propietary models, we generate summary from all permutations of the input order, and then ask the model to combine the generated summaries.[3]

5.2 Results

We present the results of running the different methods with GPT-4o in Figure 4. The top portion of the figure shows the prompting-based methods. Instructing the model to focus on either the top or bottom documents proves effective for improving faithfulness at those specific positions. For instance, the “focus top” prompt yields higher faithfulness scores for the last document than the original prompt in 4 out of 5 datasets. The “focus middle” prompt achieves higher faithfulness than the original case for the middle document on ArXiv and DiverseSumm. However, we observe that the model often fails to follow the instructions adequately to focus on the middle documents: “focus top” achieves the best faithfulness for the middle documents on PubMed, while “focus bottom” performs best on DiverseSumm. This suggests that prompting can partially alleviate positional bias.

Interestingly, the more sophisticated methods illustrated at the bottom of Figure 4 perform worse than the original prompt across all datasets. It is noteworthy, as we generally observe a very different trend for methods that change the summarization protocol. For example, on ArXiv, PubMed, and DiverseSumm, incremental updating contains gaps in the corresponding lines on the plot. This issue arises because the method maintains only a working summary cache and updates it solely based on the next document, leading to a pronounced recency bias: the model either retains its current summary and thus remains faithful to the first document or focuses on the last document. While this approach may appear to improve faithfulness, it ultimately replaces one positional bias, i.e., lost-in-the-middle, with another, i.e., recency bias.

6 Related Work

Summarization. As the capabilities of LLMs steadily improve, they exhibit strong performance on traditional summarization tasks (Zhang et al., 2024c), such as XSum (Narayan et al., 2018) and CNN/Daily Mail (See et al., 2017). Consequently, recent studies have focused on harder tasks to bet-

[3] We also explore logit-level calibration similar to Tang et al. (2024d), but this exceeds memory constraints.

ter understand the limitations of LLMs, such as instruction-controllable summarization (Liu et al., 2024b), query-focused summarization (Tang et al., 2024c), summarization of diverse information (Huang et al., 2024; Zhang et al., 2024d), and, more recently, long-form summarization (Laban et al., 2024; Kim et al., 2024a). Our work focuses on long-form summarization because generating and evaluating summaries with long context lengths is particularly challenging. Therefore, to our knowledge, we present the first study to analyze the faithfulness of generated summaries across a wide range of long-form summarization datasets, employing various generation techniques, and to examine the effect of positional bias on them.

Faithfulness Evaluation Metrics. To improve the benchmarking of summarization systems, many studies collect human annotations of faithfulness judgments to create meta-evaluation benchmarks that measure the effectiveness of faithfulness metrics (Fabbri et al., 2021; Pu et al., 2023; Tang et al., 2023; Goyal et al., 2023; Zhang et al., 2024c; Liu et al., 2023b, 2024b). This effort has led to the development of strong faithfulness metrics (Laban et al., 2022; Fabbri et al., 2022; Zha et al., 2023). More recently, LLM-based metrics, which prompt powerful LLMs to perform evaluations (Liu et al., 2023a; Wang et al., 2023; Fu et al., 2024), have shown high correlations with human judgments. Subsequently, MiniCheck (Tang et al., 2024b) has focused on distilling such knowledge into smaller NLI models, combining both strong performance and efficiency. While most of these studies concentrate on standard summarization tasks, few address long-form summarization (Krishna et al., 2023; Zhang et al., 2024b; Huang et al., 2024). Specifically, Zhang et al. (2024b) collate a meta-evaluation benchmark and develop an automatic metric. We further extend their work by incorporating additional benchmarks, analyzing the performance of the latest faithfulness metrics, and, more importantly, investigating sensitivity to positional bias.

Positional Bias. Many works have found that current LLMs exhibit different positional biases. For example, Sun et al. (2021) find that models exhibit a recency bias, where the most recent tokens play a stronger role, and that the order of in-context examples significantly affects performance (Liu et al., 2022; Lu et al., 2022; Li et al., 2024). Similarly, such biases also affect LLM performance in arithmetic tasks (Shen et al., 2023), multiple-choice questions (Zheng et al., 2024; Pezeshkpour and Hruschka, 2023), ranking (Alzahrani et al., 2024; Tang et al., 2024d), and evaluation (Wang et al., 2024). Specifically in summarization, Huang et al. (2024) find that evaluators of faithfulness and coverage highly prefer one choice over another in pairwise settings, and that generated summaries tend to focus on the first and last sections of documents. Additionally, Laban et al. (2024) find that different LLMs exhibit different positional preferences. To analyze this effect more rigorously, Ravaut et al. (2024) systematically analyze how positional bias affects context utilization. Our work instead focuses on faithfulness, another crucial aspect of summarization. As discussed in Section 4.2, faithfulness is harder to evaluate as it is confounded by coverage and thus requires attributions.

7 Conclusion

In this work, we present an extensive analysis of the relationship between positional bias and faithfulness for long-context summarization from three perspectives. We first evaluate the best strategy for assessing faithfulness in long-context summarization tasks, as well as the metrics' sensitivity to positional changes. We find that, although current LLM-based metrics achieve the highest correlation when using the full context, they are sensitive to changes in the order of the input documents.

We then analyze faithfulness of model generations with the best faithfulness metrics. We generate summaries using both open-source and proprietary models and find that the faithfulness of the middle documents tends to dip compared to those at the beginning and end, and the summaries also exhibit high sensitivity when the order of inputs is perturbed. One of the possible explanations, as we find, is that the models change behavior after a certain context length and focus on documents toward the end, improving faithfulness for the documents at the back after the initial dip.

Finally, we investigate several generation methods to test whether they can alleviate positional bias. We find that prompting methods can partially alleviate the middle curse, while more extensive methods provide overall less faithful summaries.

Limitations

This work extensively studies the relationship between position and faithfulness in long-context summarization. We acknowledge that there are additional LLMs, such as Claude or CommandR+, and more datasets that could be included in our evaluation. However, due to practical limitations, we have chosen to evaluate a representative and diverse set of LLMs and datasets. Although different models may exhibit varying trends, our analysis reveals that all models exhibit a similar trend regarding positional bias.

Furthermore, although our analysis relies on automatic metrics – and despite our extensive efforts in Section 3.1 to identify the most effective ones – it may not accurately reflect the trends that would emerge if human annotators evaluated all generated summaries. We do note, however, that human annotations for long-form summarization are both expensive and unreliable due to the extensive context involved (Krishna et al., 2023). Nevertheless, we hope that our work provides some initial insights into this problem.

In our experiments, we limit the number of documents to five for all datasets (except DiverseSumm) to control for the effect of varying context lengths. Exploring settings with different numbers of documents would be an interesting direction for future work. Nevertheless, we hope that our analysis of faithfulness with different input context lengths sheds light on what we would expect to observe with varying input lengths.

Lastly, we did not rigorously tune all the prompts in Section 5, which may lead to further improvements in mitigating the middle curse.

We do not forsee any particular risks beyond those inherent to any text generation task. In fact, our work actually focuses on understanding and improving faithfulness for long-form summarization.

Acknowledgement

We thank the anonymous reviewers for their valuable feedback. We also would like to thank Philippe Laban and Kung-Hsiang Huang for the helpful discussions and for providing help with the datasets SummHay and DiverseSumm, respectively.

References

Griffin Adams, Bichlien Nguyen, Jake Smith, Yingce Xia, Shufang Xie, Anna Ostropolets, Budhaditya Deb, Yuan-Jyue Chen, Tristan Naumann, and Noémie Elhadad. 2023. What are the desired characteristics of calibration sets? identifying correlates on long form scientific summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10520–10542, Toronto, Canada. Association for Computational Linguistics.

Norah Alzahrani, Hisham Alyahya, Yazeed Alnumay, Sultan AlRashed, Shaykhah Alsubaie, Yousef Almushayqih, Faisal Mirza, Nouf Alotaibi, Nora Al-Twairesh, Areeb Alowisheq, M Saiful Bari, and Haidar Khan. 2024. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13787–13805, Bangkok, Thailand. Association for Computational Linguistics.

Meng Cao, Yue Dong, and Jackie Cheung. 2022. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3340–3354, Dublin, Ireland. Association for Computational Linguistics.

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

Yulong Chen, Yang Liu, Ruochen Xu, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Yue Zhang. 2023. UniSumm and SummZoo: Unified model and diverse benchmark for few-shot summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12833–12855, Toronto, Canada. 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.

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 Allonsius, 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-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, 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 Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, 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 Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, 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 Tsimpoukelli, 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 Hermoso, Mo Metanat, Mohammad Rastegari, Munish 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, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, 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 Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin,

Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, 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, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian 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. *Preprint*, arXiv:2407.21783.

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.

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.

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.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. GPTScore: Evaluate as you desire. 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 6556–6576, Mexico City, Mexico. Association for Computational Linguistics.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2023. News summarization and evaluation in the era of gpt-3. *Preprint*, arXiv:2209.12356.

Kung-Hsiang Huang, Philippe Laban, Alexander Fabbri, Prafulla Kumar Choubey, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. 2024. Embrace divergence for richer insights: A multi-document summarization benchmark and a case study on summarizing diverse information from news articles. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 570–593, Mexico City, Mexico. Association for Computational Linguistics.

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021. Efficient attentions for long document summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1419–1436, Online. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Yekyung Kim, Yapei Chang, Marzena Karpinska, Aparna Garimella, Varun Manjunatha, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024a. Fables: Evaluating faithfulness and content selection in book-length summarization. *Preprint*, arXiv:2404.01261.

Yekyung Kim, Yapei Chang, Marzena Karpinska, Aparna Garimella, Varun Manjunatha, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024b. FABLES: Evaluating faithfulness and content selection in book-length summarization. In *First Conference on Language Modeling*.

Huan Yee Koh, Jiaxin Ju, He Zhang, Ming Liu, and Shirui Pan. 2022. How far are we from robust long abstractive summarization? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2682–2698, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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, Alexander R Fabbri, Caiming Xiong, and Chien-Sheng Wu. 2024. Summary of a haystack: A challenge to long-context llms and rag systems. *arXiv preprint arXiv:https://arxiv.org/pdf/2407.01370*.

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.

Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. 2024. Long-context llms struggle with long in-context learning. *Preprint*, arXiv:2404.02060.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024a. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

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

Yixin Liu, Alex Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023b. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4140–4170, Toronto, Canada. Association for Computational Linguistics.

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.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. 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.

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*, Brussels, Belgium.

OpenAI. 2024. Hello gpt-4o.

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.

Mathieu Ravaut, Aixin Sun, Nancy Chen, and Shafiq Joty. 2024. On context utilization in summarization with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2764–2781, Bangkok, Thailand. Association for Computational Linguistics.

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.

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, Sébastien Bubeck, Ronen Eldan, Yin Tat Lee, Yuanzhi Li, and Yi Zhang. 2023. Positional description matters for transformers arithmetic. *Preprint*, arXiv:2311.14737.

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.

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*. Association for Computational Linguistics.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024b. Minicheck: Efficient fact-checking of llms on grounding documents. *Preprint*, arXiv:2404.10774.

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. 2024c. 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.

Raphael Tang, Crystina Zhang, Xueguang Ma, Jimmy Lin, and Ferhan Ture. 2024d. Found in the middle: Permutation self-consistency improves listwise ranking in large language models. 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 2327–2340, Mexico City, Mexico. Association for Computational Linguistics.

Liam van der Poel, Ryan Cotterell, and Clara Meister. 2022. Mutual information alleviates hallucinations in abstractive summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5956–5965, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alex Wang, Richard Yuanzhe Pang, Angelica Chen, Jason Phang, and Samuel R. Bowman. 2022. SQuALITY: Building a long-document summarization dataset the hard way. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1139–1156, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is ChatGPT a good NLG evaluator? a preliminary study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11, Singapore. Association for Computational Linguistics.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. 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.

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.

Huajian Zhang, Yumo Xu, and Laura Perez-Beltrachini. 2024a. Fine-grained natural language inference based faithfulness evaluation for diverse summarisation tasks. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1701–1722, St. Julian's, Malta. Association for Computational Linguistics.

Huajian Zhang, Yumo Xu, and Laura Perez-Beltrachini. 2024b. Fine-grained natural language inference based faithfulness evaluation for diverse summarisation tasks. *Preprint*, arXiv:2402.17630.

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.

Yusen Zhang, Nan Zhang, Yixin Liu, Alexander Fabbri, Junru Liu, Ryo Kamoi, Xiaoxin Lu, Caiming Xiong, Jieyu Zhao, Dragomir Radev, Kathleen McKeown, and Rui Zhang. 2024d. Fair abstractive summarization of diverse perspectives. 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 3404–3426, Mexico City, Mexico. Association for Computational Linguistics.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 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, Online. Association for Computational Linguistics.

Method	MN	QM	GR	AX	CS	DS	Avg.
Full	52.2	64.3	61.6	57.6	63.4	57.4	59.4
Natural	51.9	62.5	51.5	50.0	50.0	50.6	52.8
Chunk 1024	58.1	57.1	68.5	54.7	64.8	54.3	**59.6**
Chunk 2048	50.0	50.3	47.0	53.2	66.7	55.1	53.7
Chunk 4096	50.0	50.3	47.3	43.7	53.7	52.8	49.6
Chunk 8192	50.0	50.0	50.0	50.7	53.2	52.3	51.0

Table 7: BACC of different chunking methods with GPT-4o. We use the best strategy of taking the maximum over documents and average over the summary sentences. As baselines, we report the full input setting and running the same metric with natural document boundaries.

A Additional Experimental Setup Details

A.1 Dataset Details

The licenses for the datasets are as follows. QMSUM and MultiNews are under MIT License. Diversesumm, SummHay, ArXiv, and Pubmed are released under the Appache 2.0 license. SQuality is under the CC BY 4.0 license. GovReport and ChemSumm do not specify any license. We use the authors' original repository and instructions to prepare and process the dataset. The authors of the respective datasets have filtered any harmful content. For annotations, we similarly follow author's instructions to download and process the data.

A.2 Model Details

For all models, we use the default generation methods. For open-source models, we use the available Huggingface repository for Llama-3.1-8B[4] and 70B[5], and also for Mixtral[6] and Mixtral Large.[7] For GPT-4o, we use *gpt-4o* as of October 13th, 2024. We run with *bfloat16* and use 8 A100s to run all generations and evaluations. The approximate costs for GPT-4o are as follows: metric full setting costs $37.1, metric splitting the document setting costs $79.9, generation costs $ 15.2, and generation with different methods costs $97.2.

B Additional Metric Results

B.1 Results on Chunking

Instead of using natural document boundaries, we explore chunking the full input into fixed numbers of tokens. Using the same setup as in the meta-evaluation but excluding PubMed and SQuality, we split the input documents into chunks of 1,024, 2,048, 4,096, and 8,192 tokens. We employ the GPT-4-based metric and the best merging strategy (i.e., maximum over documents and average over the summary sentences). The result is shown in Table 7. We observe that chunking into smaller chunks (1,024 tokens) leads to stronger performance than using the full context, while other chunk sizes result in worse correlations compared to the full context. This may indicate that smaller context windows help the models, since LLMs still prefer shorter contexts. The fact that natural boundaries are only better than chunking with 4,096 tokens suggests that limiting based on size may be preferable to keeping the original documents as they are, since there is no control over how long each document is.

B.2 Full Metric Results

We present the full results, including an exploration of summary sentence-level merging strategies, in Table 8. For MiniCheck, the original chunking method using the minimum function over summary sentences yields the highest correlation—though this is largely driven by high accuracy on ArXiv. Meanwhile, the maximum strategy with mean aggregation, which previous studies have identified as most effective, falls short by only 0.2 points. For GPT-4o, we observe that, across all document merging strategies, averaging over the summary sentences achieves the highest correlations overall. This finding is consistent with prior work.

We also run the meta-evaluation with Llama-3.1-8B model in the bottom of Table 8. Similar to running with GPT-4o, we observe that using the full context achieves the highest correlation on average. Nevertheless, the best summary sentence merging strategy is taking the minimum, and the second-best document merging strategy is mean.

C Additional Faithfulness Analysis Results

C.1 Content Overlap Analysis

Taking the maximum faithfulness implicitly assumes that the conent do not overlap. To verify the variability in similarity, we calculate ROUGE-1/2/L between all document pairs and aggregate the results to show the variablity for each task . We exclude SummHay as it is prohibitively expen-

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

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

[6] https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1

[7] https://huggingface.co/mistralai/Mixtral-8x22B-Instruct-v0.1

Metric	Doc Merge	Summ Merge	MN	QM	GR	PB	AX	SQ	CS	DS	Avg.
MiniCheck	Original	Min	47.8	56.6	64.7	79.7	76.0	83.0	46.8	53.5	**63.5**
	Original	Mean	56.1	55.7	55.1	79.7	57.0	83.0	57.9	54.1	62.3
	Original	Max	53.8	53.9	50.0	79.7	50.0	83.0	50.9	**55.2**	59.6
	Min	Min	49.4	56.6	64.7	50.0	50.0	83.0	51.4	49.8	56.9
	Min	Mean	54.7	55.7	55.1	50.0	50.0	83.0	54.2	49.6	56.5
	Min	Max	60.3	53.9	50.0	50.0	50.8	83.0	54.2	53.8	57.0
	Mean	Min	49.4	56.6	64.7	52.6	50.0	83.0	51.4	49.3	57.1
	Mean	Mean	64.5	55.7	55.1	52.6	53.4	83.0	54.2	46.0	58.1
	Mean	Max	60.4	53.9	50.0	52.6	46.2	83.0	66.7	51.2	58.0
	Max	Min	39.4	56.6	64.7	84.8	69.7	83.0	53.2	47.1	62.3
	Max	Mean	56.6	55.7	55.1	84.8	62.3	83.0	59.7	49.3	63.3
	Max	Max	53.2	53.9	50.0	84.8	50.0	83.0	50.9	54.2	60.0
GPT-4o	Full	Min	38.1	60.5	**64.7**	**80.1**	**73.1**	76.7	51.4	**60.6**	63.1
	Full	Mean	52.2	64.3	61.6	**80.1**	57.6	76.7	63.4	57.4	**64.1**
	Full	Max	52.2	64.3	61.6	**80.1**	57.6	76.7	63.4	57.4	**64.1**
	Min	Min	50.0	**65.4**	58.8	54.4	50.4	**80.6**	50.0	50.2	57.5
	Min	Mean	50.0	60.7	60.0	54.4	52.1	**80.6**	54.2	50.3	57.8
	Min	Max	50.0	60.7	60.0	54.4	52.1	**80.6**	54.2	50.3	57.8
	Mean	Min	46.8	**65.4**	58.8	64.3	51.7	**80.6**	50.0	52.8	58.8
	Mean	Mean	46.3	60.7	60.0	64.3	56.2	**80.6**	54.2	56.6	59.9
	Mean	Max	55.2	59.2	50.0	64.3	49.8	**80.6**	**70.8**	54.4	60.5
	Max	Min	44.6	**65.4**	58.8	76.0	68.8	**80.6**	47.7	58.2	62.5
	Max	Mean	**60.0**	60.7	60.0	76.0	55.6	**80.6**	60.6	53.0	63.3
	Max	Max	51.3	59.2	50.0	76.0	50.0	**80.6**	50.0	51.5	58.6
Llama-3.1-8B	Full	Min	52.0	56.8	**64.9**	**69.9**	56.3	**73.4**	**67.6**	52.6	**61.7**
	Full	Mean	53.8	55.0	51.2	**69.9**	50.0	**73.4**	50.9	52.1	57.0
	Full	Max	53.8	55.0	51.2	**69.9**	50.0	**73.4**	50.9	52.1	57.0
	Min	Min	49.4	**63.9**	63.4	63.7	**59.5**	69.9	53.2	49.1	59.0
	Min	Mean	50.1	62.5	51.5	63.7	59.0	69.9	55.6	50.8	57.9
	Min	Max	50.1	62.5	51.5	63.7	59.0	69.9	55.6	50.8	57.9
	Mean	Min	56.7	**63.9**	63.4	63.5	51.3	69.9	61.6	51.4	60.2
	Mean	Mean	**59.3**	62.5	51.5	63.5	50.0	69.9	61.6	50.9	58.7
	Mean	Max	49.3	52.6	50.0	63.5	50.0	69.9	50.0	49.7	54.4
	Max	Min	50.9	**63.9**	63.4	53.7	51.4	69.9	48.6	**53.3**	56.9
	Max	Mean	51.9	62.5	51.5	53.7	50.0	69.9	50.0	50.6	55.0
	Max	Max	50.0	52.6	50.0	53.7	50.0	69.9	50.0	50.0	53.3

Table 8: Full BACC on meta-evaluation benchmarks. CS=ChemSumm, AX=ArXiv, GR=GovReport, QM=QMSumm, MN=MultiNews, DS=Diversesumm. For each metric, best and second-best are bolded and underlined, respectively.BACC on meta-evaluation benchmarks, where MN refers to MultiNews, QM to QMSumm, PB to PubMed, GR to GovReport, AX to ArXiv, SQ to SQuALITY, CS to ChemSumm, and DS to Diversesumm. We **bold** the best merging strategy for each metric and underline the second-best.

sive to calculate this for 100 documents. Table 9 shows the results. Examining unigrams (R1), we observe a high degree of overlap, particularly for DiverseSumm. This is expected, as all 10 documents focus on the same news event. We note that while such overlap may not be covered by only considering the highest faithfulness for one of the documents, we evaluate faithfulness relative to a single document to minimize the influence of coverage, as mentioned in Section 4.2. Determining how many documents a summary can be attributed to is complex and would require defining a thresh-

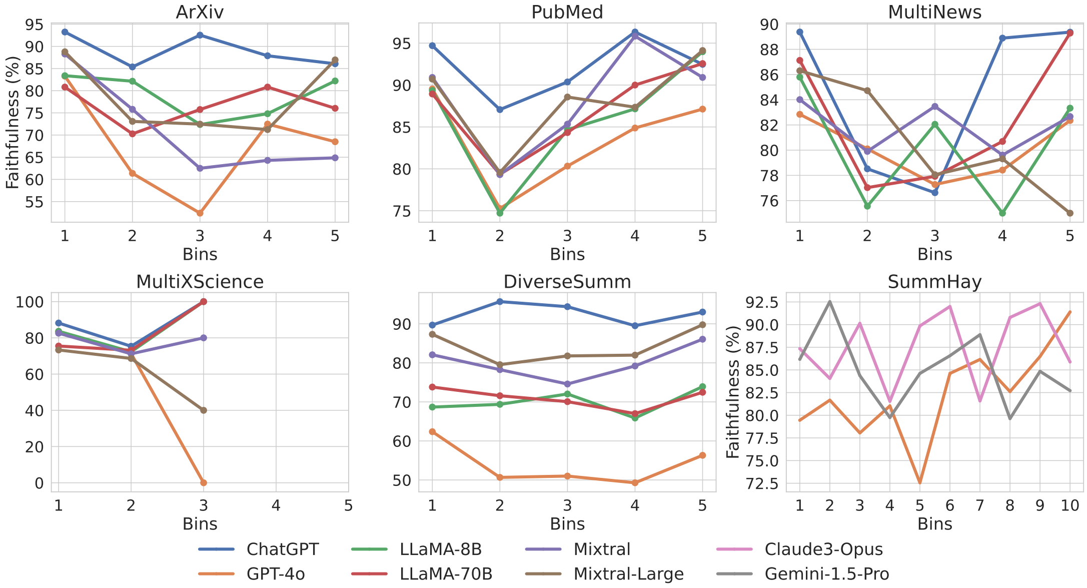

Figure 5: Faithfulness analysis across different positions of documents with chunking.

Task	R1	R2	RL
MultiXScience	24	24	12
PubMed	26	26	13
MultiNews	18	28	13
ArXiv	32	32	13
DiverseSumm	41	41	21

Table 9: Similarity of the documents using ROUGE.

old, which we leave for future work. Nevertheless, the coverage analysis shown in Figure 7 can be considered the case where the summary is related to all documents.

C.2 Faithfulness Analysis with Chunking

We perform the analysis with calculating the faithfulness with chunking, instead of natural boundaries. We show it in Figure 5.

C.3 Discussions on Variance Across Tasks

In Figure 2, we observe a large variance across the different tasks. We provide a discussion on possible reasons.

Context Length. Based on the dataset statistics provided earlier, we analyzed the tasks from those with the shortest document lengths to those with the longest. This reveals an intriguing trend: Multi-XScience, shown in Figure 2, exhibits primarily a lead bias due to its short input length. As we move to datasets with longer input lengths, the U-shaped trend becomes more apparent. MultiXScience can thus be interpreted as representing the early stage of this trend, which evolves as context length increases. We demonstrate this similar progression in Section 4.4, where increasing contexts shift the observed behavior. The downward trend of Multi-XScience in Figure 3 can similarly be attributed to insufficient context length, leaving it in the early lead bias phase.

Task Type. Another observable trend is the variation across different types of summarization tasks. For example, ArXiv and PubMed involve single long-document summarization, whereas MultiNews, MultiXScience, DiverseSumm, and SummHay are classified as MDS. Our findings indicate that single-document summarization tasks exhibit more consistent model behavior and lower variance, while MDS tasks show much higher variance. We hypothesize that this increased variability stems from the need for synthesis across multiple documents, which impacts the faithfulness of generated summaries.

C.4 Faithfulness Metric as Alignment and Attribution

We evaluate whether the faithfulness score can be used for alignment. Specifically, we use SuperPal (Ernst et al., 2021), the state-of-the-art document-summary sentence alignment model, to test whether it reaches similar conclusions in terms of faithfulness. SuperPal operates at the sentence level, aligning summary sentences to document sentences. We use the indices of the aligned document

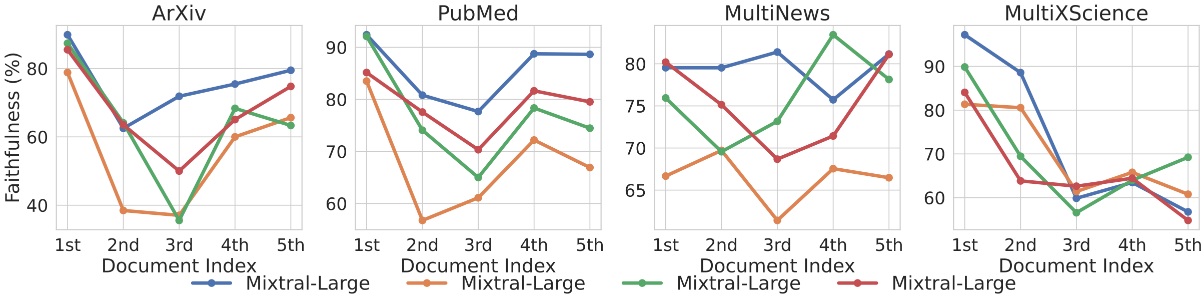

Figure 6: Faithfulness analysis across different positions with SuperPal as alignment.

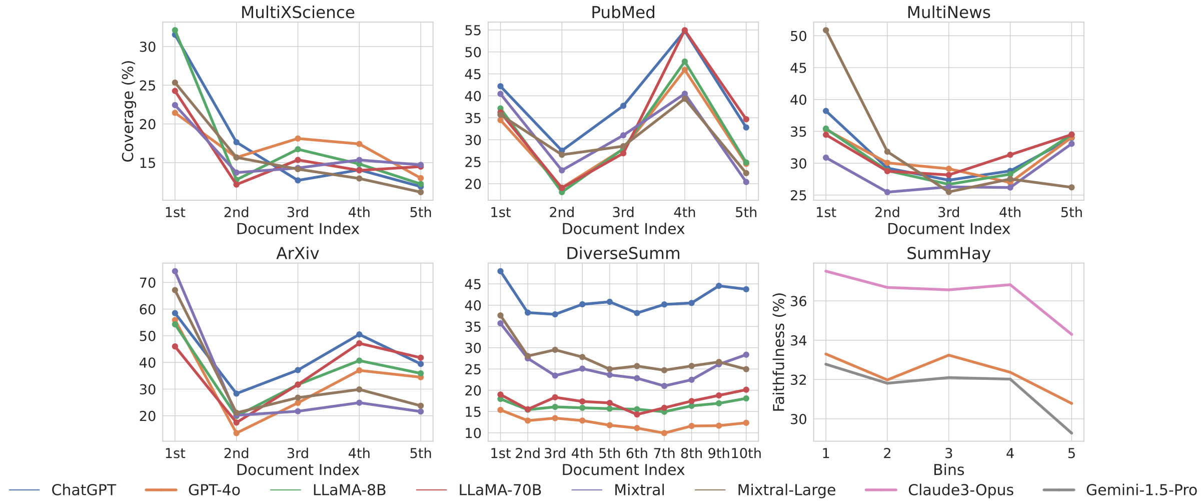

Figure 7: Coverage analysis across different positions.

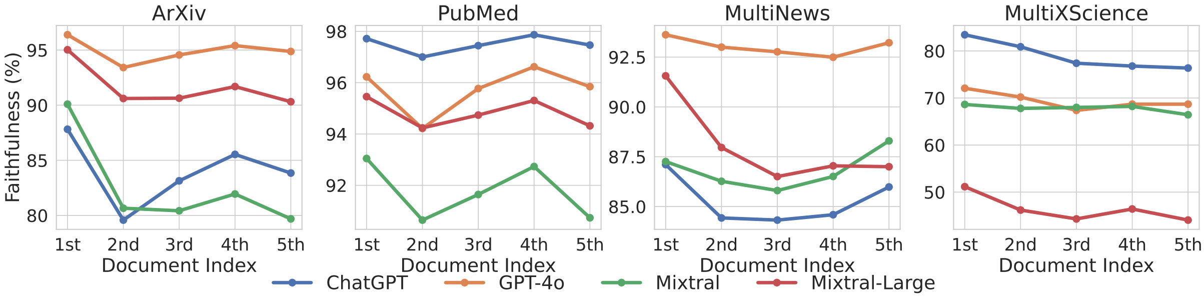

Figure 8: Faithfulness analysis across different positions using GPT-4o faithfulness metric.

sentences to compare alignment positions. The correlation between the two metrics is high, with a Spearman correlation of 0.74. We also verify the alignment visually, as shown in Figure 6. We observe the same trends as shown in Figure 2 and discussed in Section 4.2: A dominant U-shaped curve for ArXiv, PubMed, and MultiNews, and a strong lead bias for MultiXScience. This demonstrates that using the maximum faithfulness score as an attribution method provides similar conclusions to those obtained from a trained document-summary sentence alignment model.

C.5 Coverage Analysis

We also provide the coverage analysis, by using the faithfulness score of all documents. We show the figure in Figure 7. The trend generally follows the observation of Ravaut et al. (2024), who uses bigram matching between the summary and documents, observing either a U-shape or a lead bias.

C.6 Faithfulness Analysis using GPT-4o

Here, we now evaluate faithfulness using our best faithfulness metrics on a subset that excludes Llama-based models and includes only ArXiv, PubMed, MultiNews, and MultiXScience. Compared to Figure 8, which uses MiniCheck, we observe a smoother graph, but the key findings remain the same. For example, we observe the same U-shape occurring for ArXiv, PubMed, and MultiNews, while MultiXScience exhibits a strong lead bias. As mentioned in Section 4.1, since the computation is expensive, we still use MiniCheck for the remaining analyses.

D Prompts

We present the prompts used for evaluation and generation in Figure Table 10. The faithfulness metric prompt is adapted from (Tang et al., 2024c). For standard generation and focus prompt, we adapted the prompt from (Ravaut et al., 2024). Finally, we adapt the iterative update and hierarchical merging prompts from (Chang et al., 2024).

Method	Prompt
Faithfulness evaluation	Document: [ARTICLE] Sentence: [SUMMARY] Determine if the sentence is factually consistent with the document provided above. A sentence is factually consistent if it can be entailed (either stated or implied) by the document. Please start your answer with “Yes.” or “No.” Please briefly explain the reason within 50 words."""
ArXiv generation	Read the following scientific paper. Produce a summary in 6 sentences. You must give your in a structured format: ”’Summary: [your summary]”’, where [your summary] is your generated summary. ========== [ARTICLES] ==========
PubMed generation	Read the following scientific paper. Produce a summary in 7 sentences. You must give your in a structured format: ”’Summary: [your summary]”’, where [your summary] is your generated summary. ========== [ARTICLES] ==========
MultiNews generation	Read the following news articles. Produce a summary in 10 sentences. You must give your in a structured format: ”’Summary: [your summary]”’, where [your summary] is your generated summary. ========== [ARTICLES] ==========
MultiXScience generation	Read the following abstracts. Produce a summary in 5 sentences. You must give your in a structured format: ”’Summary: [your summary]”’, where [your summary] is your generated summary. ========== [ARTICLES] ==========
Focus prompt	[Generation Prompt] Pay special attention to the [top articles/articles in the middle/bottom articles].
Iterative prompt	Read the following section of a scientific paper. ========== [NEXT DOCUMENT] ========== Below is a summary up until this point: ========== [SUMMARY] ========== We are going over the articles sequentially to gradually update one comprehensive summary. Produce an updated summary in 6 sentences. You must give your in a structured format: ”’Summary: [your summary]”’, where [your summary] is your generated summary.
Hierarchical merging	Below are several summaries: — [SUMMARIES] — Create one comprehensive summary by recursively merging summaries of its chunks. Despite this recursive merging process, you need to create a summary that seems as though it is written in one go. The summary must be within 6 sentences. You must give your in a structured format: ”’Summary: [your summary]”’, where [your summary] is your generated summary.

Table 10: Prompts for evaluation (top), standard generations (middle), and advanced generation techniques (bottom).