Hallucinate at the Last in Long Response Generation: A Case Study on Long Document Summarization

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have significantly advanced text generation capabilities, including tasks like summarization, often producing coherent and fluent outputs. However, faithfulness to source material remains a significant challenge due to the generation of hallucinations. While extensive research focuses on detecting and reducing these inaccuracies, less attention has been paid to the positional distribution of hallucination within generated text, particularly in long outputs. In this work, we investigate *where* hallucinations occur in LLM-based long response generation, using long document summarization as a key case study. Focusing on the challenging setting of long context-aware long response generation, we find a consistent and concerning phenomenon: hallucinations tend to concentrate disproportionately in the latter parts of the generated long response. To understand this bias, we explore potential contributing factors related to the dynamics of attention and decoding over long sequences. Furthermore, we investigate methods to mitigate this positional hallucination, aiming to improve faithfulness specifically in the concluding segments of long outputs.

1 Introduction

Recent advancements in Large Language Models (LLMs) have pushed the boundaries of human-like language generation, particularly in tasks like text summarization (Chang et al., 2024b). With their enhanced scale and sophisticated training, LLMs now produce summaries exhibiting remarkable coherence and fluency, approaching human-level quality (Roit et al., 2023; Song et al., 2025). This capability is transforming how we interact with large volumes of text, making information more accessible (Achiam et al., 2023; Grattafiori et al., 2024; Team et al., 2024; GLM et al., 2024).

As LLMs continue to evolve, a paradigm shift is emerging from short-form to long-form generation, enabled by the ability to process extended contexts (Wu et al., 2025a). Long-output generation is critical for complex reasoning tasks like long Chain-of-Thought prompting (Jaech et al., 2024) and language agents (Sumers et al., 2024; Wang et al., 2024), inherently requiring coherent, contextually grounded long responses. While recent benchmarks evaluate long-generation capabilities (Wu et al., 2025b; Ye et al., 2025), they often lack contextual grounding, overlooking a crucial aspect of long-form generation: *context-aware faithfulness*.

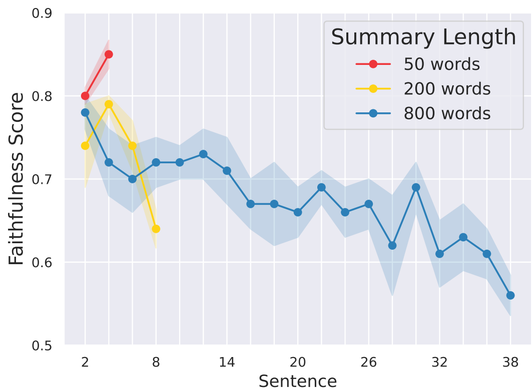


Figure 1: Comparison of the faithfulness scores of summaries generated around 50 words, 200 words, and 800 words in length for a 6K-token length Wikipedia context using GPT-4o mini.

A persistent and critical challenge faced by LLMs is the phenomenon of *hallucination*, wherein the generated content is unfaithful to or unsupported by the input context (Huang et al., 2025). While extensive prior work has focused on detecting and mitigating hallucinations, a significant limitation is the lack of research into the spatial distribution of factual errors within the generated sequence. Related research has explored positional biases in how models process *input contexts*, notably the *Lost in the Middle* phenomenon (Liu et al., 2024b; Wan et al., 2025). However, understanding the distribution of errors *within the generated output itself*, particularly for long responses from long contexts, is equally crucial for effective diagnosis and mitigation.

Motivated by this critical research gap, our work provides the first dedicated investigation into the spatial distribution of hallucinations during *long-response generation*, especially in the challenging setting of long document summarization. Unlike prior work primarily addressing long input contexts or short outputs (Tang et al., 2023; Zhang et al., 2024a; Bishop et al., 2024), we tackle *long context-aware long response generation*, which requires processing lengthy inputs and maintaining faithfulness over a significantly long output sequence. Specifically, we pose a core research question: How frequently do hallucinations occur in long-form generation tasks such as document summarization, and when they do, where in the output are they most likely to appear?

Our analysis reveals a surprising and concerning trend: hallucinations tend to concentrate towards the end of the generated text, a phenomenon we term "hallucinate at the last." As shown in Figure 1, faithfulness (See Section 3.1 for evaluation metric.) significantly decreases towards the end for long summaries (e.g., 800 words). Contrary to expectations of uniform distribution, this distinct positional bias highlights a critical vulnerability in LLMs when generating extended text. To systematically analyze this phenomenon and its implications, our research addresses three core questions:

RQ1. Where Do Hallucinations Most Frequently Occur? (§3)

RQ2. What Factors Contribute to Hallucination Concentration at the Last Part? (§4)

RQ3. How Can We Mitigate the Hallucination in the Last Part? (§5)

To answer these questions, we first empirically characterize the positional distribution of hallucinations (**RQ1**). Building on this, we explore potential causes of this phenomenon related to the generative process (**RQ2**). Finally, we propose and evaluate mitigation strategies targeting late-stage hallucinations (**RQ3**). Our findings highlight the necessity of considering the spatial dimension of errors and provide insights for developing more robust long-form generation systems.

Our contributions can be summarized as follows:

- We provide the first empirical characterization of the *Hallucinate at the Last* phenomenon in LLM-based long response generation, particularly in long document summarization.
- We offer initial insights and analysis into the potential factors contributing to this late-stage hallucination.
- We investigate mitigation strategies specifically tailored to address positional hallucination.

2 Generating Summaries

This section describes the experimental setup used to generate the summaries analyzed in our study. We prompt LLMs to generate summaries from original documents, varying both the context length and the output length.

2.1 Input Context Lengths

To analyze the impact of input length, we evaluate model performance across a range of context lengths, varying from approximately 1K tokens up to 16K tokens. The specific context lengths used in our experiments range from 1K to 8K tokens, with increments of 1K.

2.2 Output Length Categories

We define two categories for output length to study the *Hallucinate at the Last* phenomenon:

- **Short Output:** Summaries with a length between 100 and 200 words.
- **Long Output:** Summaries with a length up to 30% of the input context length.

This distinction allows us to compare hallucination patterns in standard summary lengths versus significantly longer generated responses. By including both short and long output categories, we aim to systematically analyze if and how the positional bias of hallucinations manifests and potentially becomes more severe as the length of the generated summary increases. See Appendix A for more details.

2.3 Overall Faithfulness

To provide a general understanding of the faithfulness levels observed in the generated long summaries before conducting a detailed positional analysis, we first report the overall faithfulness scores.

As shown in Table 1, we report the overall faithfulness of long summaries generated by Llama3.1-8B-Instruct (Grattafiori et al., 2024) on the Wikipedia dataset, with context lengths ranging from 1K to 8K. We employ the FineSurE (Song et al., 2024) framework with GPT-4 (Achiam et al.,

Context Length	Faithfulness(%)
1K	83.3
2K	86.7
3K	78.9
4K	82.9
5K	86.6
6K	85.9
7K	83.9
8K	86.8

Table 1: Faithfulness scores of long summaries on the Wikipedia dataset, evaluated using FineSurE.

2023) to compute the overall faithfulness scores. FineSurE evaluates summaries by comparing each sentence with the source context and identifying specific error types when inconsistencies are detected.

As shown in the results, the overall performance is consistent with the results on the CNNDM (Hermann et al., 2015) dataset reported in the original FineSurE paper, where the faithfulness of summaries generated by the Llama3-70B-Instruct model is reported to be 85.5%. Building on this, We perform a more fine-grained evaluation of summary faithfulness in Section 3, which includes a positional analysis to examine how factual consistency varies across different parts of the output.

3 Where Do Hallucinations Most Frequently Occur?

In this section, we empirically investigate where hallucinations most frequently occur in long document summarization. We analyze the positional distribution of factual errors by examining faithfulness across varying models, datasets, and output positions.

3.1 Experimental Setup

Models We conduct experiments using several state-of-the-art Large Language Models to ensure our findings are not model-specific. The models evaluated include Llama3.1-8B-Instruct (Grattafiori et al., 2024), Mistral-7B-Instruct-v0.3 (Albert Q. Jiang et al., 2023), Qwen2.5-7B-Instruct (Yang et al., 2024a), and GPT-4o mini (Achiam et al., 2023).

Datasets To examine whether the phenomenon is domain-specific, we analyze the spatial distribution of faithfulness within generated summaries across various domains. In addition to the Wikipedia dataset used in the main experiment, we include the arXiv (Cohan et al., 2018), PubMed (Cohan et al., 2018), and GovReport (Huang et al., 2021) datasets for further evaluation.

3.2 Evaluation Metric

Faithfulness We adopt an evaluation metric based on atomic facts, which has shown a significant effectiveness in evaluating summary factuality (Min et al., 2023; Tang et al., 2024; Scirè et al., 2024; Jing et al., 2024; Wei et al., 2024; Yu et al., 2024).

Let s_i be the i-th sentence in the generated summary S. Each sentence in the generated summary is decomposed into a set of atomic facts A_i by utilizing LLM, followed by a filtering process to remove unnecessary atomic facts (Yang et al., 2024b), which makes:

$$A_i = \{a_{i1}, a_{i2}, ..., a_{iNi}\} \quad (1)$$

Using an NLI model, each filtered atomic fact is then compared with the source document $D = \{d_1, ..., d_M\}$ in a pair-wise manner. We measure only the `entailment` score, and each atomic fact is assigned the highest entailment score among its comparisons:

$$\text{score}(a_{ij}) = \max_{d_m \in D} \texttt{Entail}(a_{ij}, d_m) \quad (2)$$

Finally, the average of the entailment scores for a set of filtered atomic facts in a sentence is used as the sentence's **faithfulness score**:

$$\text{Faithfulness}(s_i) = \frac{1}{N_i} \sum_{j=1}^{N_i} \text{score}(a_{ij}) \quad (3)$$

See Appendix B for more experimental details.

Sensitivity In addition to the faithfulness score, we propose a simple yet effective metric for analyzing positional discrepancies in generated outputs. Specifically, each summary is divided into five equal length bins, and the average faithfulness score of the atomic facts within each bin is computed. We define **sensitivity** as the difference between the faithfulness score of the last bin and the average of the first four bins. A positive sensitivity value indicates a tendency to hallucinate at the last, whereas a negative value suggests otherwise. The larger the sensitivity, the more pronounced the hallucination at the end of the output.

3.3 Hallucinate at the Last

The latter part of the summary exhibits the lowest level of faithfulness across nearly all models, and across all contexts and summary lengths.

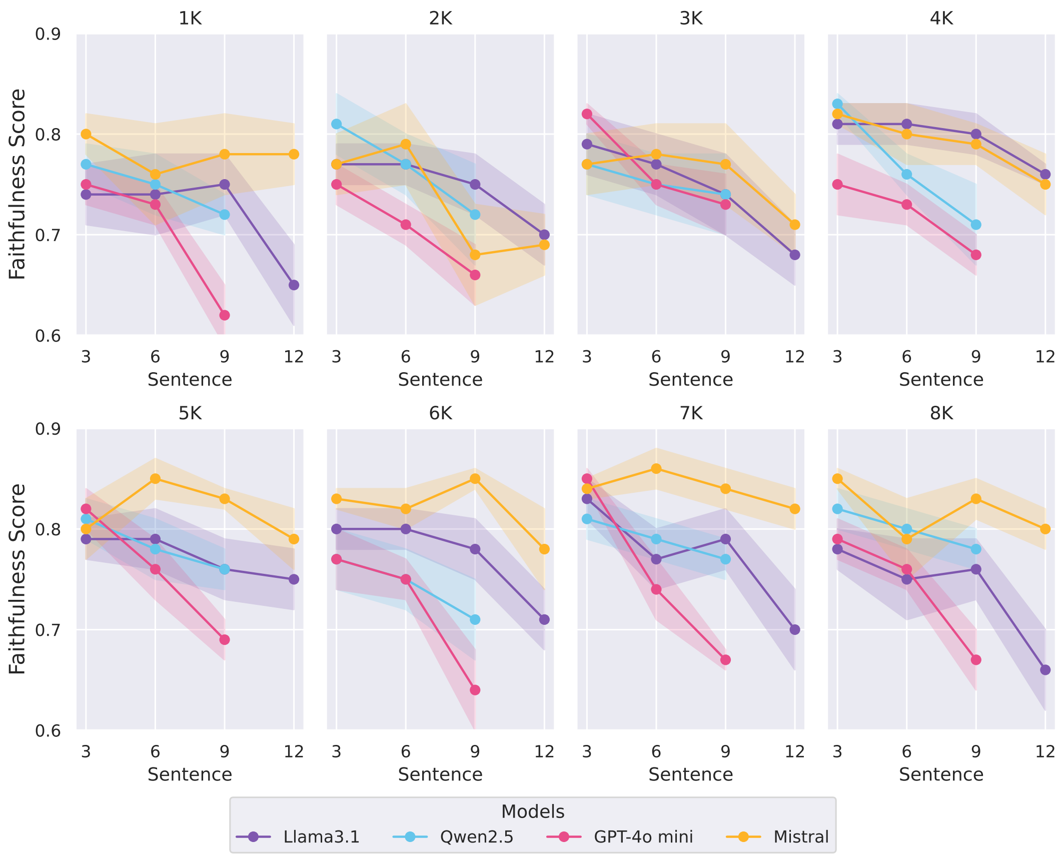


Figure 2: Comparison of faithfulness scores for **short** summaries generated by different models across increasing output lengths on the Wikipedia dataset, which context lengths ranging from 1K to 8K.

We present the positional faithfulness discrepancies of both short and long summaries generated by different models on the Wikipedia dataset, with respect to varying input context lengths, in Figure 2 and Figure 3, respectively. As shown in Figure 2, short summaries generated by the three models—excluding Mistral—consistently exhibit a decline in faithfulness towards the latter positions of the summary, regardless of the input context length, with the lowest faithfulness scores mostly observed at the end. Interestingly, we found that the positional discrepancies in generated summaries are not significantly affected by the input context length. Specifically, the decline in faithfulness at the end of summaries remains similar between the 1K and 8K context length settings.

However, the *Hallucinate at the Last* phenomenon becomes even more severe in long summaries, as shown in Figure 3. As the length of the generated summary increases, the faithfulness scores toward the end continue to decline, eventually dropping below 0.6 when 45 sentences are generated from an 8K context.

We report the sensitivity results for long summaries generated by the Llama3.1 model in Table 2, and those for the Mistral model in Table 3. As shown in the results, summaries generated by Llama3.1 tend to exhibit the lowest faithfulness scores in the final bin, resulting in high sensitivity. In contrast, the Mistral model often shows the lowest faithfulness scores in the initial bins, with sensitivity frequently falling below zero. We analyze the underlying causes of these contrasting patterns in Section 4. See Appendix C for more results.

Model	Context Length	Generated Summary Bins 1	2	3	4	5	Sensitivity
Llama3.1-8B-Instruct	1K	**0.77**	0.74	0.69	0.69	0.69	3.3
	2K	**0.74**	0.72	0.70	0.66	0.63	7.5
	3K	**0.75**	0.70	0.69	0.65	0.58	11.8
	4K	**0.74**	0.71	0.68	0.69	0.64	6.5
	5K	**0.73**	0.72	0.72	0.68	0.60	11.3
	6K	**0.74**	0.72	0.66	0.65	0.61	8.3
	7K	**0.73**	0.70	0.69	0.66	0.62	7.5
	8K	**0.76**	0.70	0.72	0.69	0.61	10.8

Table 2: Comparison of faithfulness scores and Sensitivity for **long** summaries generated by Llama3.1-8B-Instruct across output bins on the Wikipedia dataset, which context lengths ranging from 1K to 8K.

The latter part of the summary consistently demonstrates the lowest level of faithfulness

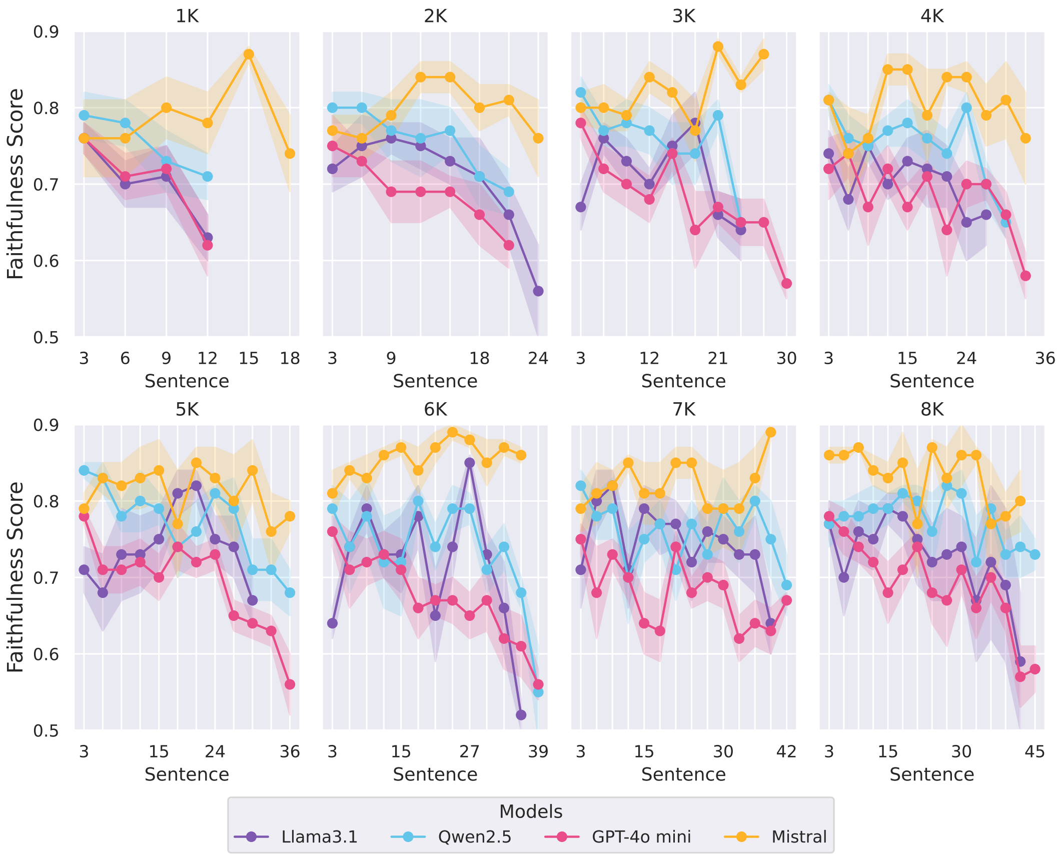


Figure 3: Comparison of faithfulness scores for **long** summaries generated by different models across increasing output lengths on the Wikipedia dataset, which context lengths ranging from 1K to 8K.

Model	Context Length	Generated Summary Bins 1	2	3	4	5	Sensitivity
Mistral-7B-Instruct-v0.3	1K	0.78	0.80	**0.84**	0.82	0.81	0.0
	2K	0.75	0.80	**0.84**	0.82	0.81	-0.8
	3K	0.81	0.80	0.82	**0.84**	0.83	-1.3
	4K	0.79	0.81	**0.83**	0.81	0.81	0.0
	5K	0.80	0.84	**0.86**	0.85	0.79	4.8
	6K	0.82	**0.87**	**0.87**	**0.87**	0.86	-0.3
	7K	0.82	0.84	0.82	0.83	**0.85**	-2.3
	8K	0.87	0.83	0.87	**0.89**	0.81	5.5

Table 3: Comparison of faithfulness scores and Sensitivity for **long** summaries generated by Mistral-7B-Instruct-v0.3 across output bins on the Wikipedia dataset, which context lengths ranging from 1K to 8K.

across nearly all datasets, contexts, and summary lengths. We present the positional faithfulness discrepancy of long summaries generated by the Llama3.1-8B-Instruct model across different datasets in Figure 4. As the results indicate, the *Hallucinate at the Last* tendency is consistently observed across multiple domains. Faithfulness scores tend to decrease toward the final segments as the output length increases. In particular, as the length of the generated summary increases, the faithfulness of its later parts continues to decline. Notably, for the Pubmed dataset, the faithfulness score falls below 0.5 when the model generates 45 sentences at a context length of 8K. Additional sensitivity results across diverse datasets are reported in Appendix D. Notably, across all datasets, none of the computed sensitivity scores fall below zero.

4 Why Do Hallucinations Frequently Occur at the Last?

This section investigates the underlying factors that contribute to the *Hallucinate at the Last* phenomenon, exploring two main hypotheses. The first attributes the phenomenon to the inherent nature of summarization: key information is typically concentrated at the beginning of the summary, with less important content appearing toward the end. The second posits that as LLMs generate longer outputs, they increasingly attend to previously generated tokens rather than the original input context, resulting in a shift in attention distribution.

4.1 Is it Intrinsic to Summarization?

This hypothesis is based on the intrinsic structure of human-written and model-generated summaries. In most summarization tasks—especially in news, scientific, and Wikipedia-style documents—the main ideas and most salient information are typically

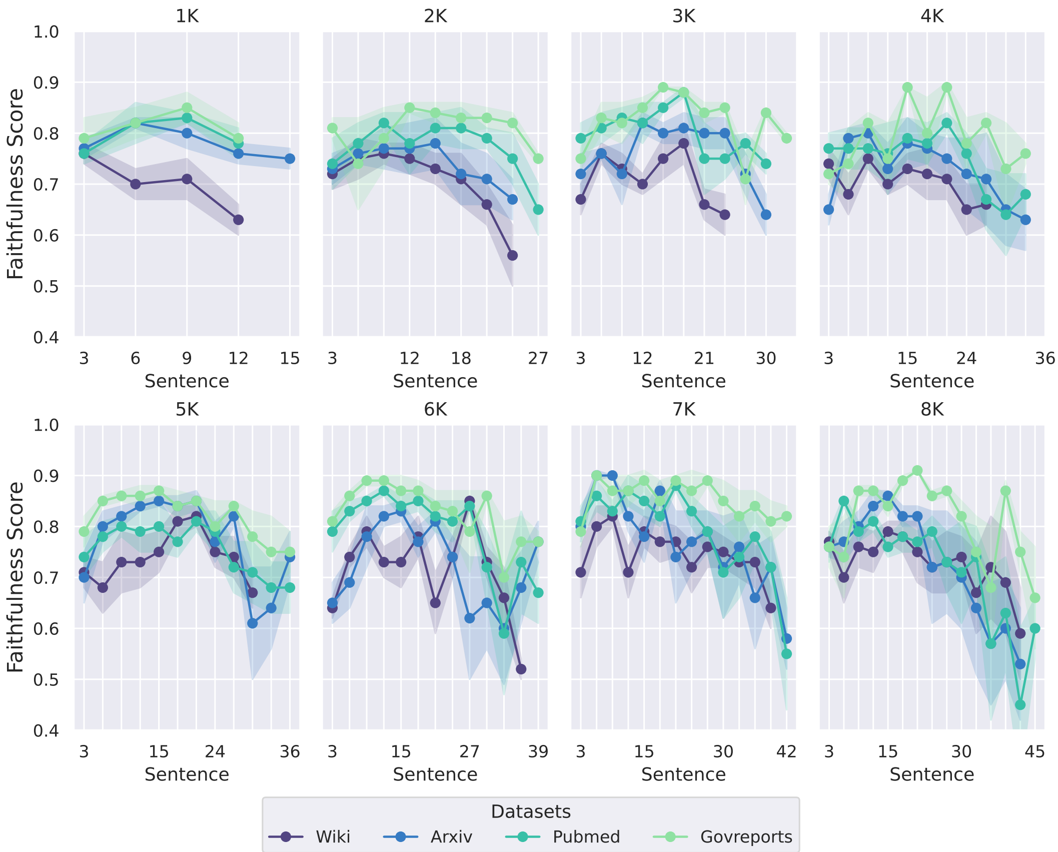


Figure 4: Comparison of faithfulness scores in **long** summaries generated by the Llama3.1-8B-Instruct model across increasing output lengths in multiple domains, with context lengths ranging from 1K to 8K. We report results in **short** summaries in Appendix D.

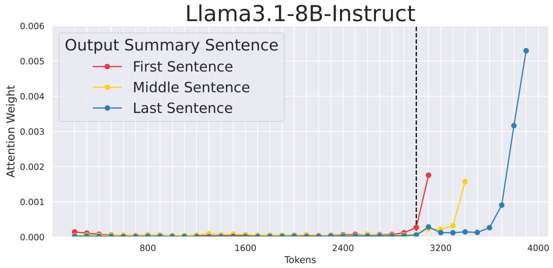
presented early in the summary (Hermann et al., 2015; Narayan et al., 2018; Cohan et al., 2018). This structure mirrors the **lead bias** often observed in source documents, where the beginning contains the most crucial content (Kim et al., 2019; Zhao et al., 2022; Ravaut et al., 2024). Motivated by this, our study investigates whether such structural bias also manifests in the form of decreasing factual consistency in generated outputs. Specifically, we compare the positional faithfulness scores of human-written reference summaries and model-generated summaries to assess whether the decline in faithfulness is an artifact of summarization's inherent structure or a model-specific behavior.

Experimental Setup We use human-written reference summaries from the CNNDM dataset, selecting examples with long summaries and input contexts of around 2K tokens. We compare these references to summaries generated by the Llama3.1-8B-Instruct model, evaluated using the metric described in Section 3.1.

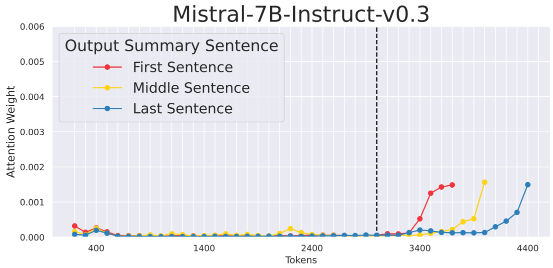
Results & Analysis Figure 5 presents a positional faithfulness comparison between human-

Figure 5: Comparison of faithfulness scores between the reference summary and summaries generated by the Llama3.1-8B-Instruct model across increasing output lengths on the CNNDM dataset.

written reference summaries and model-generated summaries from the CNNDM dataset, analyzed across increasing output lengths. As shown in the results, the reference summaries exhibit a different pattern compared to the generated summaries. While the faithfulness scores of the generated summaries drop below 0.6 in the final segment, the reference summaries show a temporary dip around



(a) Average attention weights on the first, middle, and last sentences of summaries generated by the **Llama3.1-8B-Instruct** model.



(b) Average attention weights on the first, middle, and last sentences of summaries generated by the **Mistral-7B-Instruct-v0.3** model.

Figure 6: Average attention weights of two models exhibiting contrasting trends in Figure 3. The dashed line separates the input context from the generated output.

the middle but recover toward the end. This observation suggests that the *Hallucination at the Last* phenomenon **cannot be solely attributed to the inherent characteristics of summarization**.

4.2 Is it Intrinsic to Attention?

Prior research on LLMs has demonstrated that attention weight distributions are closely correlated with the generation process, influencing output coherence (Dong et al., 2021; Zhang et al., 2024b). Recent studies on Large Vision-Language Models (LVLMs) further suggest that hallucinations frequently occur in the later parts of generated text (Liu et al., 2024a; Lee et al., 2024; Min et al., 2025). This phenomenon has been attributed to a shift in attention: as text generation progresses, attention weights increasingly favor previously generated text tokens over image tokens. Inspired by this observation, we investigate whether a similar trend exists in LLMs by analyzing how attention weights on generated tokens evolve as output length increases.

Experimental Setup Unlike previous studies that focus on attention to the first generated token (Hsieh et al., 2024), we compute attention weights at the sentence level. Moreover, we segment the full sequence into chunks of 100 tokens. Our analysis specifically focuses on three positions within the generated output: the first sentence, the middle sentence, and the final sentence. See Appendix E for more details.

Results & Analysis Figure 6 presents a visualization of average attention weights across sentences. As shown in the results, the Llama3.1-8B-Instruct model—which exhibited the *Hallucinate at the Last* pattern in Figure 3—assigns nearly four times more attention to the final sentence of the generated summary compared to the first and middle sentences. In contrast, the Mistral-7B-Instruct-v0.3 model assigns similar levels of attention to all three sentence positions. These findings suggest that **increased attention to previously generated text correlates with a higher likelihood of hallucination**. Moreover, we observe that for most models—excluding Mistral—as output length increases, attention becomes increasingly concentrated on generated tokens, which in turn amplifies hallucination.

5 How to Mitigate the *Hallucination at the Last*?

We investigate how we can resolve the *Hallucinate at the Last* phenomenon. To this end, we apply four methods.

Experimental Setup In this experiment, we generate summaries for the Wikipedia dataset using four different methods, with a context length of 7K tokens. As a baseline, we use summaries generated by the Llama3.1-8B-Instruct model. For comparison, we evaluate the following four methods:

- **BOOOOKSCORE** (Chang et al., 2024a) segments the input context into chunks, generates summaries for each chunk individually, and then merges the partial summaries.
- **MINFERENCE** (Jiang et al., 2024) employs sparse attention mechanisms to efficiently process long input sequences.
- **LONGWRITER-LLAMA3.1-8B** (Bai et al., 2025) is a model fine-tuned on a long-output dataset and further enhanced using DPO (Rafailov et al., 2023).
- **ADACAD** (Wang et al., 2025) enhances factual consistency during generation by context-aware decoding.

See Appendix F for more experimental details.

Methods	Generated Summary Bins 1	2	3	4	5	Sensitivity
Llama3.1-8B	0.74	**0.75**	**0.75**	0.73	<u>0.64</u>	10.3
+ BooookScore	0.73	0.77	0.78	**0.80**	0.75	2.0
+ MInference	0.77	**0.78**	**0.78**	0.71	<u>0.69</u>	7.0
+ LongWriter	**0.85**	0.77	0.81	<u>0.63</u>	0.70	6.5
+ AdaCAD	0.79	**0.82**	0.77	0.81	<u>0.64</u>	15.8

Table 4: Comparison of faithfulness scores and Sensitivity for **long** summaries generated by different mitigation methods across output bins on the Wikipedia dataset with a 7K context length.

Figure 7: Comparison of faithfulness scores for **long** summaries generated by different mitigation methods across increasing output lengths, using the Wikipedia dataset with a 7K context length.

Results & Analysis We report the faithfulness scores across output bins and the corresponding Sensitivity results in Table 4, and the faithfulness scores across increasing output lengths for each method in Figure 7. As shown in Table 4, BooookScore achieves the lowest sensitivity (2.0), making it the closest to zero among the four methods. Moreover, as illustrated in Figure 7, BooookScore maintains the highest level of faithfulness, particularly around the 30th sentence, and notably avoids the sharp decline in faithfulness observed in other methods at the end of the summary. These results suggest that **generating summaries independently for each chunk and subsequently merging them can be an effective strategy** for mitigating the *Hallucinate at the Last* phenomenon. However, while BooookScore mitigates the phenomenon, it doesn't fundamentally solve the problem of generating long responses directly from raw long inputs, relying instead on decomposing the problem via chunking and merging. This highlights the necessity of future work to develop methods that enable models to maintain faithfulness when generating long text directly from long contexts.

6 Related Work

The problem of hallucination in LLMs has been a significant area of research. Prior work has largely focused on developing methods for detecting factual inconsistencies in generated text (Chuang et al., 2024; Hu et al., 2024; Kim et al., 2024; Zhong and Litman, 2025) and proposing strategies to mitigate their occurrence, often through improved training (Zhang et al., 2023; Wan et al., 2023), decoding (Shi et al., 2024; Wang et al., 2025), or prompting (Zhou et al., 2023). While effective in reducing overall hallucination rates, these studies typically do not analyze the within-sequence distribution of errors.

The concept of positional bias in generation errors has been observed in related domains. Notably, studies on Large Vision-Language Models (LVLMs) have indicated that hallucinations tend to increase towards the end of generated textual descriptions of images (Liu et al., 2024a; Lee et al., 2024; Min et al., 2025).

Recently, several studies have begun to explore long-output generation in LLMs (Wu et al., 2025b; Ye et al., 2025). However, these works focus on procedural generation tasks that are not grounded in contextual input, limiting their applicability to real-world summarization scenarios. A notable study on multi-document summarization aligns with our findings (Belém et al., 2025). Nevertheless, it does not address positional context-aware faithfulness over extended summary lengths.

Our work presents the first dedicated study of where hallucinations occur in LLM-based long document summarization, moving beyond detection to uncover positional patterns, underlying causes, and mitigation strategies.

7 Conclusion

We identified and characterized the *Hallucinate at the Last* phenomenon in LLM-based long response generation, specifically in long document summarization. Our findings show that hallucinations disproportionately increase towards the end of long outputs, a bias amplified in longer summaries. We investigated the contributing factors and explored targeted mitigation strategies. Our work highlighted the importance of the output's temporal dimension in LLM faithfulness and motivates future research into spatially-aware generation techniques.

Limitations

In this study, we explored four domains, primarily because the corresponding datasets provided input contexts of the desired length. However, we were not able to investigate domains such as books, dialogues, movie scripts, or meeting transcripts. Books were excluded due to their excessive length, while the other domains lacked datasets with suitable input lengths or sufficient sample sizes. Additionally, we also did not explore models of varying sizes.

Despite these limitations, we believe this work offers an important foundation for future research in the relatively underexplored domain of long-form output generation, particularly in summarization.

Ethics Statement

This study leverages publicly available datasets, including Wikipedia, Arxiv, Pubmed, Govreport, and CNNDM, to analyze long-form text generation in LLMs. All experiments were conducted in a consistent and reproducible manner across models and datasets, without any manipulation or omission of data or results. We investigate the phenomenon of *Hallucinate at the Last*, a tendency observed in long context-aware summarization models to generate hallucinated content toward the final portion of the summary. By drawing attention to this issue, our study contributes to ongoing efforts to enhance the reliability and factual consistency of LLM-generated summaries.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Alexandre Sablayrolles Albert Q. Jiang, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Yushi Bai, Jiajie Zhang, Xin Lv, Linzhi Zheng, Siqi Zhu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2025. Longwriter: Unleashing 10,000+ word generation from long context LLMs. In *The Thirteenth International Conference on Learning Representations*.

Forrest Bao, Miaoran Li, Rogger Luo, and Ofer Mendelevitch. 2024. HHEM-2.1-Open.

Catarina G Belém, Pouya Pezeshkpour, Hayate Iso, Seiji Maekawa, Nikita Bhutani, and Estevam Hruschka. 2025. From single to multi: How LLMs hallucinate in multi-document summarization. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5276–5309, Albuquerque, New Mexico. Association for Computational Linguistics.

Jennifer A. Bishop, Sophia Ananiadou, and Qianqian Xie. 2024. LongDocFACTScore: Evaluating the factuality of long document abstractive summarisation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10777–10789, Torino, Italia. ELRA and ICCL.

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

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024b. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.

Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James R. Glass. 2024. Lookback lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1419–1436, Miami, Florida, USA. Association for Computational Linguistics.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Yue Dong, Chandra Bhagavatula, Ximing Lu, Jena D. Hwang, Antoine Bosselut, Jackie Chi Kit Cheung, and Yejin Choi. 2021. On-the-fly attention modulation for neural generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1261–1274, Online. Association for Computational Linguistics.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 1693–1701, Cambridge, MA, USA. MIT Press.

Cheng-Yu Hsieh, Yung-Sung Chuang, Chun-Liang Li, Zifeng Wang, Long Le, Abhishek Kumar, James Glass, Alexander Ratner, Chen-Yu Lee, Ranjay Krishna, and Tomas Pfister. 2024. Found in the middle: Calibrating positional attention bias improves long context utilization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14982–14995, Bangkok, Thailand. Association for Computational Linguistics.

Xiangkun Hu, Dongyu Ru, Lin Qiu, Qipeng Guo, Tianhang Zhang, Yang Xu, Yun Luo, Pengfei Liu, Yue Zhang, and Zheng Zhang. 2024. Refchecker: Reference-based fine-grained hallucination checker and benchmark for large language models.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

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.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Huiqiang Jiang, Yucheng Li, Chengruidong Zhang, Qianhui Wu, Xufang Luo, Surin Ahn, Zhenhua Han, Amir H. Abdi, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2024. MInference 1.0: Accelerating pre-filling for long-context LLMs via dynamic sparse attention. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. 2024. FaithScore: Fine-grained evaluations of hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5042–5063, Miami, Florida, USA. Association for Computational Linguistics.

Byeongchang Kim, Hyunwoo Kim, and Gunhee Kim. 2019. Abstractive summarization of Reddit posts with multi-level memory networks. 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 2519–2531, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Saehyung Lee, Seunghyun Yoon, Trung Bui, Jing Shi, and Sungroh Yoon. 2024. Toward robust hyper-detailed image captioning: A multiagent approach and dual evaluation metrics for factuality and coverage. *arXiv preprint arXiv:2412.15484*.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2024a. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*.

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

Kyungmin Min, Minbeom Kim, Kang-il Lee, Dongryeol Lee, and Kyomin Jung. 2025. Mitigating hallucinations in large vision-language models via summary-guided decoding. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4183–4198, Albuquerque, New Mexico. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

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.

Paul Roit, Johan Ferret, Lior Shani, Roee Aharoni, Geoffrey Cideron, Robert Dadashi, Matthieu Geist, Sertan Girgin, Leonard Hussenot, Orgad Keller, Nikola Momchev, Sabela Ramos Garea, Piotr Stanczyk, Nino Vieillard, Olivier Bachem, Gal Elidan, Avinatan Hassidim, Olivier Pietquin, and Idan Szpektor. 2023. Factually consistent summarization via reinforcement learning with textual entailment feedback. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6252–6272, Toronto, Canada. Association for Computational Linguistics.

Alessandro Scirè, Karim Ghonim, and Roberto Navigli. 2024. FENICE: Factuality evaluation of summarization based on natural language inference and claim extraction. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14148–14161, Bangkok, Thailand. Association for Computational Linguistics.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 783–791, Mexico City, Mexico. Association for Computational Linguistics.

Hwanjun Song, Hang Su, Igor Shalyminov, Jason Cai, and Saab Mansour. 2024. FineSurE: Fine-grained summarization evaluation using LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 906–922, Bangkok, Thailand. Association for Computational Linguistics.

Hwanjun Song, Taewon Yun, Yuho Lee, Jihwan Oh, Gihun Lee, Jason Cai, and Hang Su. 2025. Learning to summarize from LLM-generated feedback. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 835–857, Albuquerque, New Mexico. Association for Computational Linguistics.

Theodore R. Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L. Griffiths. 2024. Cognitive architectures for language agents.

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

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

David Wan, Jesse Vig, Mohit Bansal, and Shafiq Joty. 2025. On positional bias of faithfulness for long-form summarization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8791–8810, Albuquerque, New Mexico. Association for Computational Linguistics.

David Wan, Shiyue Zhang, and Mohit Bansal. 2023. HistAlign: Improving context dependency in language generation by aligning with history. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2941–2960, Singapore. Association for Computational Linguistics.

Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. 2025. AdaCAD: Adaptively decoding to balance conflicts between contextual and parametric knowledge. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11636–11652, Albuquerque, New Mexico. Association for Computational Linguistics.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6).

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Zixia Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V Le.

2024. Long-form factuality in large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Yuhao Wu, Yushi Bai, Zhiqing Hu, Shangqing Tu, Ming Shan Hee, Juanzi Li, and Roy Ka-Wei Lee. 2025a. Shifting long-context llms research from input to output.

Yuhao Wu, Ming Shan Hee, Zhiqiang Hu, and Roy Ka-Wei Lee. 2025b. Longgenbench: Benchmarking long-form generation in long context LLMs. In *The Thirteenth International Conference on Learning Representations*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024a. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Joonho Yang, Seunghyun Yoon, ByeongJeong Kim, and Hwanhee Lee. 2024b. FIZZ: Factual inconsistency detection by zoom-in summary and zoom-out document. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, Miami, Florida, USA. Association for Computational Linguistics.

Xi Ye, Fangcong Yin, Yinghui He, Joie Zhang, Howard Yen, Tianyu Gao, Greg Durrett, and Danqi Chen. 2025. Longproc: Benchmarking long-context language models on long procedural generation.

Tianyu Yu, Haoye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganqu Cui, Yunkai Dang, Taiwen He, Xiaocheng Feng, Jun Song, Bo Zheng, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2024. Rlaif-v: Open-source ai feedback leads to super gpt-4v trustworthiness.

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.

Qingru Zhang, Chandan Singh, Liyuan Liu, Xiaodong Liu, Bin Yu, Jianfeng Gao, and Tuo Zhao. 2024b. Tell your model where to attend: Post-hoc attention steering for LLMs. In *The Twelfth International Conference on Learning Representations*.

Yunxiang Zhang, Muhammad Khalifa, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, and Lu Wang. 2023. Merging generated and retrieved knowledge for open-domain QA. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4710–4728, Singapore. Association for Computational Linguistics.

Chao Zhao, Faeze Brahman, Kaiqiang Song, Wenlin Yao, Dian Yu, and Snigdha Chaturvedi. 2022. NarraSum: A large-scale dataset for abstractive narrative summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 182–197, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yang Zhong and Diane Litman. 2025. Discourse-driven evaluation: Unveiling factual inconsistency in long document summarization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2050–2073, Albuquerque, New Mexico. Association for Computational Linguistics.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. Context-faithful prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14544–14556, Singapore. Association for Computational Linguistics.

Context Length	Words Range
1K	200 to 250
2K	400 to 500
3K	600 to 750
4K	800 to 1000
5K	1000 to 1250
6K	1200 to 1500
7K	1400 to 1750
8K	1600 to 2000

Table 5: Words range used in the prompt for long output generation.

A Details for Generating Summaries

Building upon Section 2, this section details our approach to generating long summaries and presents an analysis of the generated outputs.

Prompt Short outputs were standardized to a length of 100 to 200 words, while long outputs were approximated by converting 30% of the total context token length into a corresponding word range. The specific word ranges used for prompting long-summary generation are provided in Table 5, and the prompt template employed for summary generation is shown in Table 7.

Generated Summary Length Our experiments revealed that controlling the output length of LLMs remains a significant challenge. This difficulty becomes more pronounced as the input context length increases and as the target word range specified in the prompt becomes larger. In particular, with longer contexts, LLMs frequently generate long summaries that contain repeated sentences or even entire paragraphs. To ensure a fair and reliable evaluation, we excluded such faulty outputs and sampled only summaries with comparable lengths for use in our analysis.

Table 6 presents the average word counts of both short and long summaries generated by each model. As the results indicate, LongWriter-Llama3.1-8B frequently produces considerably longer outputs, while Qwen2.5-7B-Instruct tends to generate relatively shorter summaries.

B Details for Evaluation Metric

In this section, we present the experimental details of the evaluation metric used to compute sentence-level faithfulness scores in Section 3.1.

For atomic fact decomposition, we utilize the Llama3.1-8B-Instruct model with vLLM (Kwon et al., 2023)[1]. For the NLI model, we use the state-of-the-art hallucination evaluation model (Bao et al., 2024)[2].

[1] https://github.com/vllm-project/vllm

Models	Avg. Words short	Avg. Words long
Llama3.1-8B-Instruct	279	834
Qwen2.5-7B-Instruct	153	751
Mistral-7B-Instruct-v0.3	285	794
GPT-4o mini	172	925
LongWriter-Llama3.1-8B	398	3291

Table 6: The average word counts of both **short** and **long** summaries generated by each model on the Wikipedia 7K dataset.

C More Results on Varying Models

Usage of LLMs We utilized publicly available instruction-tuned models, including Llama3.1-8B-Instruct[3], Mistral-7B-Instruct-v0.3[4], and Qwen2.5-7B-Instruct[5] from HuggingFace. For the GPT-4o mini model, we used `gpt-4o-mini-2024-07-18`. All summaries were generated using greedy decoding with `float16` precision.

More Results We report the faithfulness scores and Sensitivity for **short** summaries generated by different models across output bins on the Wikipedia dataset in Table 8, and those of **long** summaries in Table 9.

D More Results on Varying Datasets

We report comparison of faithfulness scores in short summaries generated by the Llama 3.1 model at various output positions across multiple domains, with context lengths ranging from 1K to 8K in Table 8.

We report the faithfulness scores and Sensitivity of **short** summaries generated by the Llama3.1-8B-Instruct model across output bins in multiple domains in Table 10, and those of **long** summaries in Table 11.

As demonstrated by the experimental results, none of the summaries generated by the Llama3.1-8B-Instruct model across various domains exhibited a sensitivity value below zero.

[2] https://huggingface.co/vectara/hallucination_evaluation_model

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

[4] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

[5] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

E Attention Weight Calculation

This section provides a more detailed and formal explanation our attention weight computation method introduced in Section 4.2.

Let the full sequence, consisting of the input context and the generated output summary, be defined as:

$$x_{full} = x_{prompt} + x_{output} = [x_1, x_2, ..., x_L] \quad (4)$$

where L is the number of tokens.

We partition x_{full} into K non-overlapping blocks of 100 tokens each, such that each block is represented as:

$$x^k = \{x_i^k\}_{i=1}^{100} \quad (5)$$

for $k \in \{1, 2, ..., K\}$.

Let the output summary contain three target sentences: the **first**, **middle**, and **last** sentence, denoted respectively as:

$$\begin{aligned} x^{first} &= \{x_j^{first}\}_{j=1}^{T_1} \\ x^{second} &= \{x_j^{second}\}_{j=1}^{T_2} \\ x^{third} &= \{x_j^{third}\}_{j=1}^{T_3} \end{aligned} \quad (6)$$

where T_1, T_2, T_3 are the respective sentence lengths.

We define the average attention weight from block x^k to each of the three sentences as follows:

$$\begin{aligned} \mathtt{Attn}_{first}(k) &= \frac{1}{100 \cdot T_1} \sum_{i=1}^{100} \sum_{j=1}^{T_1} \mathtt{attn}(x_i^k \to x_j^{first}) \\ \mathtt{Attn}_{middle}(k) &= \frac{1}{100 \cdot T_2} \sum_{i=1}^{100} \sum_{j=1}^{T_2} \mathtt{attn}(x_i^k \to x_j^{middle}) \\ \mathtt{Attn}_{last}(k) &= \frac{1}{100 \cdot T_3} \sum_{i=1}^{100} \sum_{j=1}^{T_3} \mathtt{attn}(x_i^k \to x_j^{last}) \end{aligned} \quad (7)$$

where $\mathtt{attn}(x_i \to x_j)$ denotes the attention weight from token x_i to token x_j when generating the output.

These measures quantify the extent to which each 100-token block in the full sequence contributes to the generation of the respective sentences in the output summary.

F Details for Mitigation Methods

In this section, we provide the experimental details of the mitigation methods described in Section 5.

For the BOOOOKSCORE method, the input is divided into 2048-token chunks, summaries are generated for each chunk using the Llama3.1-8B-Instruct model, and the partial summaries are then hierarchically merged into a final summary. For LONGWRITER, we employed the LONGWRITER-LLAMA3.1-8B model[6] with greedy decoding and `bfloat16` precision.

[6] `https://huggingface.co/THUDM/LongWriter-llama3.1-8b`

Input Prompt
Write an accurate and engaging summary for the given text in range of {{words_range}} words using only the provided passage (might be irrelevant). Use an unbiased and journalistic tone. Text: {{Text}}

Table 7: Prompt template used to generate summaries from the original document in Section 2.

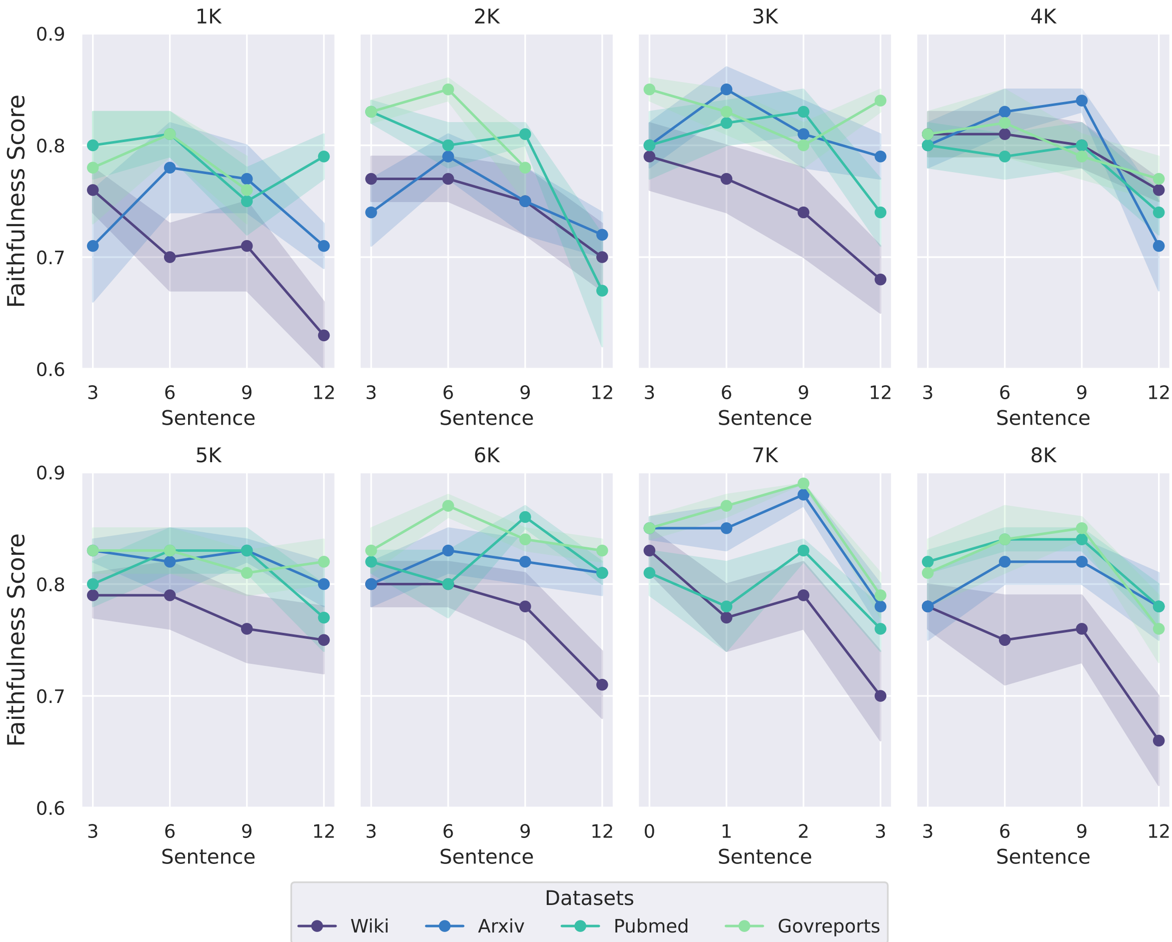


Figure 8: Comparison of faithfulness scores in **short** summaries generated by the Llama3.1-8B-Instruct model across increasing output lengths in multiple domains, with context lengths ranging from 1K to 8K.

Models	Context Length	Generated Summary Bins 1	2	3	4	5	Sensitivity
Llama3.1-8B-Instruct	1K	0.73	0.73	**0.75**	0.74	<u>0.71</u>	3.0
	2K	**0.77**	0.76	0.75	<u>0.73</u>	<u>0.73</u>	2.25
	3K	0.79	**0.80**	0.76	0.76	<u>0.69</u>	8.75
	4K	0.81	**0.82**	0.81	0.79	<u>0.76</u>	4.75
	5K	**0.80**	0.79	0.79	0.76	<u>0.75</u>	3.5
	6K	**0.81**	0.80	0.80	0.76	<u>0.71</u>	8.25
	7K	**0.85**	0.76	0.78	0.78	<u>0.70</u>	9.25
	8K	**0.80**	0.76	0.73	0.74	<u>0.63</u>	12.75
Qwen2.5-7B-Instruct	1K	**0.78**	0.75	0.77	0.72	<u>0.69</u>	6.5
	2K	**0.83**	0.82	0.74	<u>0.72</u>	0.74	3.75
	3K	**0.82**	0.78	<u>0.69</u>	0.81	0.72	5.5
	4K	**0.85**	0.77	0.78	<u>0.72</u>	0.74	4.0
	5K	**0.83**	0.82	0.79	0.80	<u>0.74</u>	7.0
	6K	**0.79**	0.75	0.77	0.75	<u>0.70</u>	6.5
	7K	0.80	**0.82**	0.77	0.80	<u>0.75</u>	4.75
	8K	**0.84**	0.81	0.78	0.82	<u>0.77</u>	4.25
GPT-4o mini	1K	**0.79**	0.69	0.76	0.73	<u>0.65</u>	9.25
	2K	**0.79**	0.75	0.73	<u>0.66</u>	<u>0.66</u>	7.25
	3K	**0.84**	0.81	0.80	0.73	<u>0.71</u>	8.5
	4K	0.74	**0.78**	0.75	0.72	<u>0.68</u>	6.75
	5K	**0.86**	0.81	0.79	0.75	<u>0.68</u>	12.25
	6K	0.77	**0.80**	0.76	0.77	<u>0.67</u>	10.5
	7K	**0.87**	0.85	0.75	0.77	<u>0.68</u>	13.0
	8K	**0.80**	**0.80**	0.76	0.77	<u>0.68</u>	10.25
Mistral-7B-Instruct-v0.3	1K	**0.83**	0.77	0.77	<u>0.75</u>	0.80	-2.0
	2K	0.74	**0.80**	**0.80**	0.73	<u>0.67</u>	9.75
	3K	0.76	0.75	**0.78**	**0.78**	<u>0.70</u>	6.75
	4K	0.79	**0.82**	0.75	0.80	<u>0.74</u>	5.0
	5K	0.83	0.83	**0.84**	0.81	<u>0.78</u>	4.75
	6K	0.82	0.81	0.82	**0.84**	<u>0.79</u>	3.25
	7K	0.83	0.82	**0.88**	0.87	<u>0.81</u>	4.0
	8K	**0.85**	0.84	0.83	0.82	<u>0.77</u>	6.5

Table 8: Comparison of faithfulness scores and Sensitivity for **short** summaries generated by different models across output bins on the Wikipedia dataset, with context lengths ranging from 1K to 8K. The highest faithfulness score in each bin is marked in **bold**, while the lowest is <u>underlined</u>.

Models	Context Length	Generated Summary Bins 1	2	3	4	5	Sensitivity
Llama3.1-8B-Instruct	1K	**0.76**	0.75	0.69	0.68	0.68	4.0
	2K	0.73	**0.76**	0.73	0.70	0.63	10.0
	3K	0.71	0.72	0.72	**0.74**	0.62	10.25
	4K	0.71	**0.75**	0.70	0.70	0.67	4.5
	5K	0.71	0.75	**0.79**	**0.79**	0.70	6.0
	6K	0.69	**0.76**	0.73	**0.76**	0.61	12.5
	7K	0.78	**0.79**	0.75	0.76	0.67	10.0
	8K	0.74	**0.79**	0.74	0.77	0.67	9.0
Qwen2.5-7B-Instruct	1K	**0.83**	0.76	0.79	0.73	0.70	7.75
	2K	**0.81**	0.79	0.78	0.78	0.70	9.0
	3K	**0.81**	0.78	0.80	0.73	0.70	8.0
	4K	**0.77**	0.76	0.74	0.76	0.69	6.75
	5K	**0.82**	0.81	0.77	0.76	0.71	8.0
	6K	**0.81**	0.78	0.78	0.75	0.67	11.0
	7K	**0.81**	0.74	0.75	0.76	0.75	1.5
	8K	**0.79**	0.78	**0.79**	0.78	0.73	5.5
GPT-4o mini	1K	**0.77**	0.74	0.69	0.69	0.69	3.3
	2K	**0.74**	0.72	0.70	0.66	0.63	7.5
	3K	**0.75**	0.70	0.69	0.65	0.58	11.8
	4K	**0.74**	0.71	0.68	0.69	0.64	6.5
	5K	**0.73**	0.72	0.72	0.68	0.60	11.3
	6K	**0.74**	0.72	0.66	0.65	0.61	8.3
	7K	**0.73**	0.70	0.69	0.66	0.62	7.5
	8K	**0.76**	0.70	0.72	0.69	0.61	10.8
Mistral-7B-Instruct-v0.3	1K	0.78	0.80	**0.84**	0.82	0.81	0.0
	2K	0.75	0.80	**0.84**	0.82	0.81	-0.8
	3K	0.81	0.80	0.82	**0.84**	0.83	-1.3
	4K	0.79	0.81	**0.83**	0.81	0.81	0.0
	5K	0.80	0.84	**0.86**	0.85	0.79	4.8
	6K	0.82	**0.87**	**0.87**	**0.87**	0.86	-0.3
	7K	0.82	0.84	0.82	0.83	**0.85**	-2.3
	8K	0.87	0.83	0.87	**0.89**	0.81	5.5

Table 9: Comparison of faithfulness scores and Sensitivity for **long** summaries generated by different models across output bins on the Wikipedia dataset, with context lengths ranging from 1K to 8K. The highest faithfulness score in each bin is marked in **bold**, while the lowest is underlined.

Datasets	Context Length	Generated Summary Bins 1	2	3	4	5	Sensitivity
Wikipedia	1K	0.73	0.73	**0.75**	0.74	0.71	3.0
	2K	**0.77**	0.76	0.75	0.73	0.73	2.25
	3K	0.79	**0.80**	0.76	0.76	0.69	8.75
	4K	0.81	**0.82**	0.81	0.79	0.76	4.75
	5K	**0.80**	0.79	0.79	0.76	0.75	3.5
	6K	**0.81**	0.80	0.80	0.76	0.71	8.25
	7K	**0.85**	0.76	0.78	0.78	0.70	9.25
	8K	**0.80**	0.76	0.73	0.74	0.63	12.75
Arxiv	1K	0.69	0.76	0.79	**0.81**	0.71	5.25
	2K	0.75	0.76	**0.80**	0.76	0.70	6.75
	3K	0.79	**0.86**	0.84	0.83	0.70	13.0
	4K	0.79	**0.83**	**0.83**	**0.83**	0.75	7.0
	5K	0.82	**0.84**	**0.84**	0.83	0.79	4.25
	6K	0.81	**0.83**	0.81	0.81	0.81	0.5
	7K	0.85	**0.86**	**0.86**	**0.86**	0.76	9.75
	8K	0.78	**0.83**	**0.83**	0.81	0.77	4.25
Pubmed	1K	0.81	**0.83**	0.81	0.73	0.79	0.5
	2K	0.82	**0.84**	0.78	0.81	0.74	7.25
	3K	0.83	**0.84**	0.81	0.83	0.76	6.75
	4K	0.81	**0.82**	0.78	0.80	0.76	4.25
	5K	0.81	0.80	**0.83**	0.82	0.73	8.5
	6K	0.79	0.80	**0.86**	0.83	0.78	4.0
	7K	0.80	0.81	**0.82**	0.79	0.78	2.5
	8K	0.83	**0.84**	0.81	**0.84**	0.75	8.0
Govreport	1K	0.72	0.79	0.80	**0.82**	0.76	2.25
	2K	0.83	0.82	**0.86**	0.83	0.77	6.5
	3K	0.83	**0.84**	0.83	0.82	0.81	2.0
	4K	0.82	0.76	**0.83**	0.78	0.78	1.75
	5K	0.83	0.81	**0.86**	0.82	0.82	1.0
	6K	0.84	0.83	**0.87**	0.86	0.82	3.0
	7K	0.83	0.86	**0.89**	**0.89**	0.79	7.75
	8K	0.84	0.82	**0.87**	0.85	0.81	3.5

Table 10: Comparison of faithfulness scores and Sensitivity for **short** summaries generated by the Llama3.1-8B-Instruct model across output bins in multiple domains, with context lengths ranging from 1K to 8K. The highest faithfulness score in each bin is marked in **bold**, while the lowest is underlined.

Datasets	Context Length	Generated Summary Bins 1	2	3	4	5	Sensitivity
Wikipedia	1K	**0.76**	0.75	0.69	<u>0.68</u>	<u>0.68</u>	4.0
	2K	0.73	**0.76**	0.73	0.70	<u>0.63</u>	10.0
	3K	0.71	0.72	0.72	**0.74**	<u>0.62</u>	10.25
	4K	0.71	**0.75**	0.70	0.70	<u>0.67</u>	4.5
	5K	0.71	0.75	**0.79**	**0.79**	<u>0.70</u>	6.0
	6K	0.69	**0.76**	0.73	**0.76**	<u>0.61</u>	12.5
	7K	0.78	**0.79**	0.75	0.76	<u>0.67</u>	10.0
	8K	0.74	**0.79**	0.74	0.77	<u>0.67</u>	9.0
Arxiv	1K	0.77	**0.82**	0.81	0.80	<u>0.74</u>	6.0
	2K	0.76	0.78	**0.80**	0.72	<u>0.71</u>	5.5
	3K	0.75	0.79	**0.81**	0.78	<u>0.68</u>	10.25
	4K	<u>0.71</u>	**0.81**	0.74	0.77	0.73	2.75
	5K	0.77	0.83	**0.85**	0.81	<u>0.74</u>	7.5
	6K	<u>0.69</u>	**0.82**	0.79	0.75	0.75	1.25
	7K	**0.86**	**0.86**	0.80	0.80	<u>0.77</u>	6.0
	8K	0.79	0.81	**0.85**	0.82	<u>0.73</u>	8.75
Pubmed	1K	<u>0.75</u>	**0.85**	0.84	0.80	0.78	3.0
	2K	0.76	**0.83**	0.82	**0.83**	<u>0.72</u>	9.0
	3K	0.81	0.85	**0.86**	0.78	<u>0.76</u>	6.5
	4K	0.77	**0.82**	0.79	0.79	<u>0.68</u>	11.25
	5K	0.78	**0.81**	**0.81**	0.78	<u>0.70</u>	9.5
	6K	0.80	**0.87**	0.83	0.77	<u>0.74</u>	7.75
	7K	**0.85**	**0.85**	0.84	0.79	<u>0.77</u>	6.25
	8K	0.80	0.79	**0.81**	0.77	<u>0.67</u>	12.25
Govreport	1K	0.80	0.80	0.85	**0.87**	<u>0.78</u>	5.0
	2K	0.82	**0.85**	0.83	**0.85**	<u>0.78</u>	5.75
	3K	<u>0.81</u>	0.86	**0.89**	<u>0.81</u>	0.82	2.25
	4K	<u>0.76</u>	**0.86**	**0.86**	0.82	0.77	5.5
	5K	0.82	0.85	**0.88**	0.82	<u>0.78</u>	6.25
	6K	0.86	**0.88**	**0.88**	0.87	<u>0.80</u>	7.25
	7K	0.86	**0.91**	0.89	0.88	<u>0.83</u>	5.5
	8K	0.84	0.87	**0.89**	0.84	<u>0.80</u>	6.0

Table 11: Comparison of faithfulness scores and Sensitivity for **long** summaries generated by the Llama3.1-8B-Instruct model across output bins in multiple domains, with context lengths ranging from 1K to 8K. The highest faithfulness score in each bin is marked in **bold**, while the lowest is <u>underlined</u>.