# *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. Despite these advances, hallucinations remain a critical challenge. While prior studies primarily focus on detecting or mitigating hallucinations, little attention has been paid to where hallucinations occur within long generated outputs. In this work, we investigate *where* hallucinations frequently 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 positional 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, modern LLMs now produce summaries exhibiting remarkable coherence and fluency, approaching human-level quality (Song et al., 2025). This capability is transforming how we interact with large volumes of text, making information more accessible (Achiam et al., 2023; Team et al., 2024).

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), requiring coherent and contextually grounded long responses. Such applications demand outputs that are not only fluent but also faithful to the given context, as unfaithful generation can lead to critical errors in downstream tasks. While recent benchmarks evaluate long-generation capabilities (Wu et al., 2025; Ye et al., 2025), they often lack contextual grounding, overlooking a crucial aspect of long-form generation: *context-aware faithfulness*, which we define as the degree to which each part of a long generated output remains supported by the input context throughout the entire generation process.

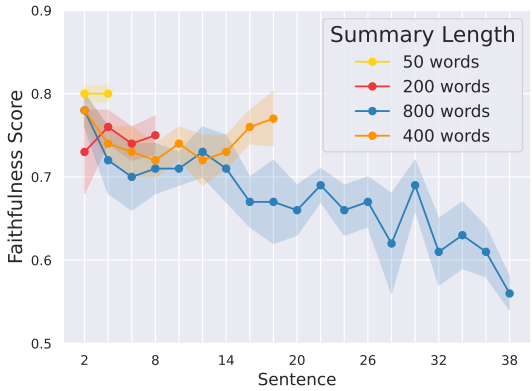



Figure 1: Comparison of the faithfulness scores of summaries generated around 50, 200, and 800 words using GPT-4o mini.

A key obstacle to context-aware faithfulness is the phenomenon of *hallucination* (Huang et al., 2025), wherein models generate content that is unfaithful to or unsupported by the input context. While extensive prior work has focused on detecting and mitigating hallucinations, little attention has been paid to the positional 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 where errors arise

*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, we present a systematic investigation into the positional 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 (Zhang et al., 2024a; Bishop et al., 2024), we focus on long document summarization as a representative setting of *long context-aware long response generation*, where models must process lengthy inputs and maintain 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 consistent 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 significantly decreases towards the final portion of long summaries (e.g., 800 words). To systematically analyze this phenomenon, we investigate the following research 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 characterize the positional distribution of hallucinations (**RQ1**), analyze potential causes (**RQ2**), and explore mitigation strategies (**RQ3**).

Our contributions can be summarized as follows:

- We identify and quantify the *Hallucinate at the Last* phenomenon, showing that hallucinations disproportionately occur near the end of long outputs.
- We analyze potential causes of this phenomenon, revealing that attention shifts toward previously generated tokens as generation progresses.
- 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. The prompt is provided in Appendix G.

### 2.1 Input Context Lengths

We evaluate model performance with input length of over 4K tokens. For the Llama model specifically, we further extend the evaluation to context lengths of 12K and 16K tokens.

### 2.2 Output Length Categories

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

- **Standard Summary:** Summaries with a length between 100 and 200 words. [1]
- **Long Summary:** 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 standard 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 G for more details.

### 2.3 Summary-Level Faithfulness

As a preliminary step, we measure the overall faithfulness of the generated long summaries with FineSurE (Song et al., 2024) to ensure consistent quality over various context lengths, prior to our main positional analysis. Table 1 shows results for Llama3.1-8B-Instruct (Grattafiori et al., 2024) on the Wikipedia dataset, evaluated using FineSurE with GPT-4 (Achiam et al., 2023). FineSurE assesses each sentence against the source, identifying error types when inconsistencies occur. The results align with those on CNNDM (Hermann et al., 2015), where Llama3-70B-Instruct achieved 85.5% faithfulness. We next conduct a finer-grained positional analysis in Section 3 for various setups.

[1] The length of the standard summaries in our study was determined based on the average length of the reference summaries in the dataset, which is employed in Section 4.

| CONTEXT LENGTH | FAITHFULNESS(%) |
|---|---|
| 4K | 82.9 |
| 5K | 86.6 |
| 6K | 85.9 |
| 7K | 83.9 |
| 8K | 86.8 |

Table 1: Summary-level faithfulness of **long** summaries on the Wikipedia dataset, evaluated using FineSurE.

## 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 evaluate our approach on multiple state-of-the-art LLMs, including Llama3.1-8B-Instruct, Gemma3-12B-Instruct (Team et al., 2025), Qwen2.5-7B-Instruct (Yang et al., 2024a), and GPT-4o mini, to ensure model-agnostic findings. Throughout all experiments, summaries are generated using greedy decoding.

**Datasets** To examine whether the phenomenon is domain-specific, we analyze the positional faithfulness of generated summaries across four datasets: Wikipedia, arXiv (Cohan et al., 2018), PubMed (Cohan et al., 2018), and GovReport (Huang et al., 2021). For each dataset, we randomly sampled 20 examples for analysis.

### 3.2 Evaluation Metric

**Sentence-Level Faithfulness** In addition to FineSurE, which evaluates sentence-level faithfulness by assessing the presence of errors using an LLM, we further adopt a complementary metric to enable more fine-grained and thus more accurate assessment of faithfulness for positional evaluation (Min et al., 2023; Jing et al., 2024; Yu et al., 2024). Therefore, we adopt MiniCheck-Flan-T5-Large (Tang et al., 2024), which decomposes each sentence into atomic facts and assigns a continuous faithfulness score between 0 and 1 to each fact against the source document, enabling precise measurement of faithfulness at each output position.

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

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

We then compare each filtered atomic fact with the source document $D = d_1, ..., d_M$ using the MiniCheck model $\mathcal{M}$ in a pairwise fashion. For each atomic fact, we take the maximum score obtained across all source sentences:

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

Finally, the **faithfulness score for each sentence** is computed as the average score of all filtered atomic facts within that sentence:

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

Each sentence is assigned a faithfulness score in the range $[0, 1]$, where 1 indicates that all atomic facts in the sentence are fully supported by the source document, and 0 indicates a complete lack of support.

**Sensitivity** In addition to the faithfulness score, we propose a simple yet effective metric for analyzing positional discrepancies in generated outputs. Following Singh and Goshtasbpour (2022), each summary is divided into five equal length bins, and the average faithfulness score of the atomic facts within each bin is computed. The bins are defined using the `cut` function in the pandas package, and the scores of the atomic facts are assigned to bins using open-interval boundaries, meaning that boundary values are not included. Let $f_k$ denote the average faithfulness score of the $k$-th bin ($k = 1, \ldots, 5$). We define **sensitivity** as:

$$\text{Sensitivity} = 100 \times \left( \frac{1}{4} \sum_{k=1}^{4} f_k - f_5 \right).$$

A positive value indicates that the faithfulness in the final bin is lower than in the earlier bins (i.e., a tendency to hallucinate at the end), whereas a negative value suggests the opposite trend. Larger sensitivity values correspond to stronger hallucination effects near the end of the generated output.

Additional experimental details and an example of the evaluation procedure are provided in Appendix C. Furthermore, the human evaluation of the quality and completeness of the generated atomic facts and experimental details is presented in Appendix B.

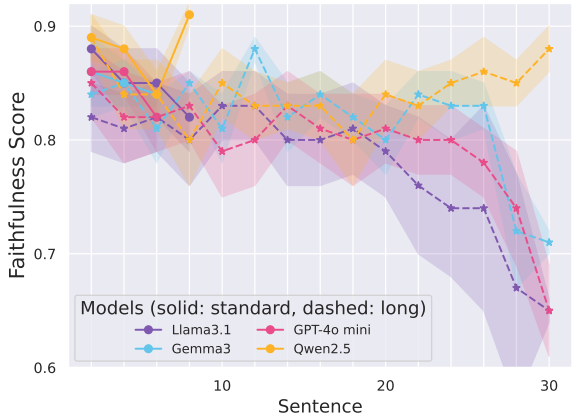



Figure 2: Comparison of faithfulness scores (evaluated by MiniCheck) for summaries generated by different models on the Wikipedia dataset.

### 3.3 Hallucinate at the Last

**The latter part of the summary exhibits the lowest level of faithfulness across multiple models. The decline becomes most pronounced when the summaries are long.** We present the positional faithfulness discrepancies of both standard and long summaries generated by different models on the Wikipedia dataset in Figure 2. As shown in the figure, standard summaries remain relatively flat in faithfulness across positions, whereas long summaries produced by the three models (excluding Qwen) exhibit a clear decline toward the later portions, with the lowest scores typically appearing at the end. This *Hallucinate at the Last* phenomenon becomes even more pronounced in long summaries: as the summary length increases, the faithfulness near the end continues to decrease, eventually dropping below 0.65 for these models. In contrast, Qwen displays a different pattern. Its faithfulness tends to drop most sharply in the middle of the summary, after which the scores recover toward the end.

Tables 2 and 3 report the faithfulness and sensitivity values for the different models, based on MiniCheck and FineSurE, respectively. All models except Qwen show markedly low faithfulness scores in the final bin, indicating high sensitivity. Consistent with the human evaluation in Table 5, Llama demonstrates the highest sensitivity regardless of model size. In contrast, Qwen is the only model that records negative sensitivity values. Unlike the consistent decline observed in the other models, Qwen maintains relatively stable faithfulness across bins, resulting in negative sensitivity. We provide a detailed analysis of these contrasting patterns in Section 4.

Table 4 presents the faithfulness and sensitivity results for both the standard and long summaries generated by the LLaMA model with context lengths of 12K and 16K tokens. Interestingly, we find that positional discrepancies in the generated summaries are not significantly affected by the input context length. As shown in the table, even with longer contexts of approximately 12K and 16K tokens, sensitivity remains high, consistently demonstrating the *Hallucinate at the Last* phenomenon. Supplementary results across different context lengths and methods are presented in Appendix D.

| Models | Generated Summary Bins 1 | 2 | 3 | 4 | 5 | Sensitivity |
|---|---|---|---|---|---|---|
| Llama3.1-8B-Instruct | 0.81 | **0.82** | 0.79 | 0.80 | 0.75 | 5.50 |
| Llama3.1-70B-Instruct | **0.83** | 0.82 | **0.83** | 0.81 | 0.77 | 5.25 |
| Gemma3-12B-Instruct | **0.85** | 0.83 | 0.84 | 0.83 | 0.79 | 4.75 |
| GPT-4o mini | **0.83** | 0.81 | 0.81 | 0.78 | 0.76 | 4.75 |
| Qwen2.5-7B-Instruct | **0.86** | 0.83 | 0.83 | 0.82 | 0.84 | -0.50 |

Table 2: Faithfulness scores (evaluated by MiniCheck) and sensitivity for **long** summaries generated by various models on the Wikipedia dataset.

| Models | Generated Summary Bins 1 | 2 | 3 | 4 | 5 | Sensitivity |
|---|---|---|---|---|---|---|
| Llama3.1-8B-Instruct | 0.24 | **0.41** | 0.30 | 0.29 | 0.03 | 28.41 |
| Gemma3-12B-Instruct | **0.87** | 0.85 | 0.84 | 0.83 | 0.63 | 24.28 |
| GPT-4o mini | 0.59 | 0.54 | 0.57 | **0.67** | 0.48 | 15.97 |
| Qwen2.5-7B-Instruct | 0.69 | 0.59 | 0.57 | 0.58 | **0.77** | -15.16 |

Table 3: Faithfulness scores (evaluated by FineSurE) and sensitivity for **long** summaries generated by various models on the Wikipedia dataset.

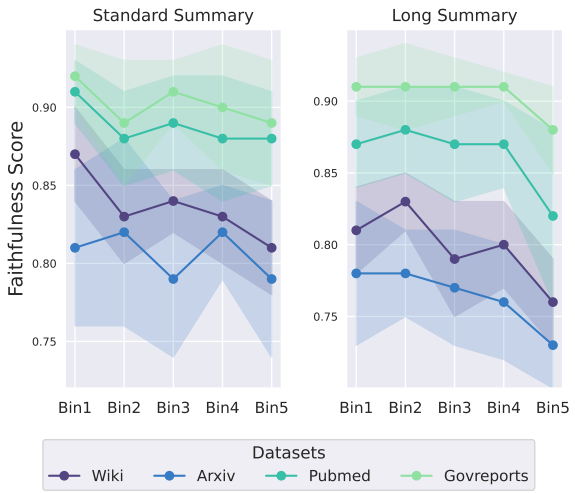



Figure 3: Comparison of faithfulness scores (evaluated by MiniCheck) for summaries generated by the Llama3.1-8B-Instruct model in multiple domains.

| Models | Context Length | Generated Summary Bins 1 | 2 | 3 | 4 | 5 | Sensitivity |
|---|---|---|---|---|---|---|---|
| **Standard Summary** | | | | | | | |
| Llama3.1-8B-Instruct | 12K | **0.91** | 0.87 | 0.85 | 0.76 | 0.67 | 17.65 |
| | 16K | **0.84** | 0.83 | 0.76 | 0.71 | 0.69 | 3.00 |
| **Long Summary** | | | | | | | |
| Llama3.1-8B-Instruct | 12K | **0.80** | 0.79 | 0.66 | 0.63 | 0.69 | 9.50 |
| | 16K | **0.86** | 0.83 | 0.85 | 0.83 | 0.79 | 5.25 |

Table 4: Faithfulness scores and sensitivity (evaluated by MiniCheck) for **standard** and **long** summaries generated by Llama on the Wikipedia dataset, with 12K and 16K context length.

| Model | Generated Summary Bins 1 | 2 | 3 | 4 | 5 | Sensitivity |
|---|---|---|---|---|---|---|
| Llama3.1-8B-Instruct | **0.90** | 0.87 | 0.84 | 0.84 | 0.70 | 16.25 |

Table 5: Human evaluation of faithfulness scores and sensitivity for **long** summaries generated by Llama on the Wikipedia dataset.

**The latter part of the summary consistently demonstrates the lowest level of faithfulness across all datasets, with a marked decline on the long summaries.** We present the positional faithfulness discrepancy of long summaries generated by the Llama model across different datasets in Figure 3. As the results indicate, the *Hallucinate at the Last* tendency is consistently observed across multiple domains. Faithfulness scores gradually decrease toward the final segments as the output length increases, and this decline becomes most severe in the concluding portions of longer summaries. Additional sensitivity results across diverse datasets are provided in Appendix E.

**The *Hallucinate at the Last* phenomenon persists across a range of decoding strategies for several models.** To demonstrate that the *Hallucinate at the Last* phenomenon is not merely a consequence of error propagation, we first conduct experiments using various decoding strategies. Following Shi et al. (2024a), we adopt the following decoding strategies to generate summaries.

- **Temperature & Top-$p$ Sampling** samples from tokens whose cumulative probability exceeds $p$. We conduct experiments with temperature values $\tau \in \{0.5, 0.7\}$ and with top-$p$ thresholds of 0.7 and 0.9.
- **Top-$k$ Sampling** samples tokens from the $k$ most probable candidates. We experiment with $k$ values of 20 and 100.
- **$\eta$-Sampling** discards tokens whose probabilities fall below an entropy dependent threshold. The hyperparameter $\eta$ is explored with 6e-4 and 4e-3.
- **Beam Search** keeps the top candidate sequences at each step based on cumulative log-probabilities. We experiment with beam sizes of 4 and 8.

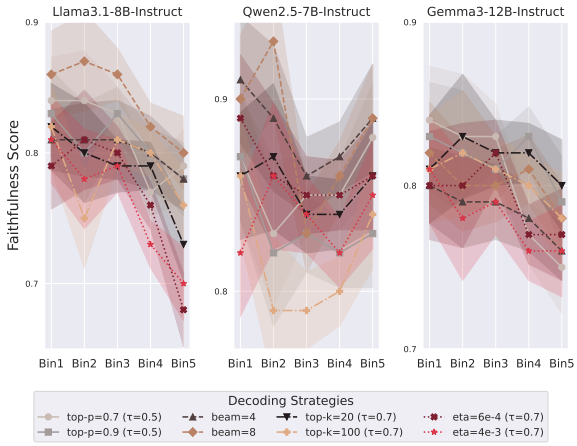



Figure 4: Comparison of faithfulness scores (evaluated by MiniCheck) for **long** summaries generated using different decoding strategies on the Wikipedia dataset.

We present the positional faithfulness discrepancies in long summaries generated using different decoding strategies in Figure 4. For the Llama and Gemma models, the latter portions of the summaries exhibit the lowest levels of faithfulness across all decoding strategies. This effect is particularly pronounced under $\eta$-sampling for Llama, where the faithfulness score remains above 0.8 in the earlier segments but drops below 0.7 in the final bin. In contrast, the Qwen model shows a different trend. With the exception of two decoding strategies, the faithfulness of the generated summaries does not decline toward the end but instead increases. Notably, the faithfulness scores for Qwen are lowest in the middle portions of the summaries rather than at the end. Motivated by these observations, we further investigate the factors associated with these positional faithfulness patterns.

**Human Evaluation** To validate the reliability of our MiniCheck-based automatic evaluation in capturing positional faithfulness and to ensure the robustness of our findings, we conduct a human evaluation to serve as a ground truth comparison. Each atomic fact was labeled as true or false based on whether it was supported by the source document. True labels were scored as 1 and false as 0, and these scores were averaged within the same five positional bins as in the automatic evaluation. The resulting human-judged faithfulness scores are reported in Table 5. Furthermore, to assess the reliability of the human evaluation, we measured inter-annotator agreement among annotators who

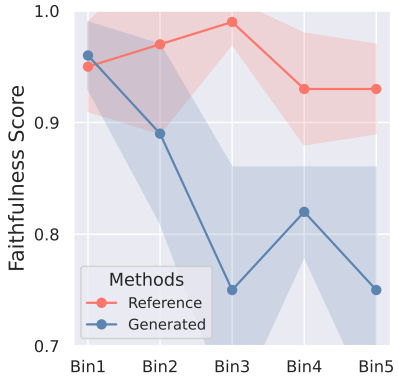



Figure 5: Faithfulness scores (evaluated by FineSurE) of reference summaries and those generated by Llama on the CNNDM dataset.

evaluated the same subset. In total, 543 atomic facts were evaluated, and the annotators agreed on 515 of them, corresponding to a 94.8% raw agreement.

## 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 motivated by the structure of human- and model-generated summaries. In domains such as news, scientific articles, and Wikipedia, key information is typically introduced early (Hermann et al., 2015; Narayan et al., 2018; Cohan et al., 2018), reflecting the **lead bias** in source texts (Kim et al., 2019; Zhao et al., 2022; Ravaut et al., 2024). We examine whether this structural bias is reflected in positional faithfulness by comparing human-written and model-generated summaries.

**Experimental Setup** We use human-written reference summaries from the CNNDM dataset, selecting examples with long reference summaries. We compare these references to summaries generated by the Llama3.1-8B-Instruct model.

**Results & Analysis** Figure 5 compares the positional faithfulness of human-written and model-generated CNNDM summaries, as evaluated by FineSurE. We observe that the faithfulness of the generated summaries declines steadily, falling below 0.75 in the final segment, whereas the reference summaries remain relatively stable throughout. This suggests that *Hallucination at the Last* **is not solely due to summarization's inherent structure**.

### 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** Following Hsieh et al. (2024), we compute attention weights by averaging the decoder self-attention matrices across all layers and heads. Unlike the prior study, however, we compute sentence-level attention by aligning token spans with sentence boundaries and averaging the token-level attention weights within each sentence. We segment the sequence into 100-token chunks and focus our analysis on three positions: the first, middle, and final sentences. Further details are provided in Appendix A.

**Results & Analysis** Figure 6 presents a visualization of average attention weights across sentences. As shown in Figure 6a, the Llama model, which exhibited the *Hallucinate at the Last* pattern in Figure 2, assigns nearly three times more attention to the final sentence of the generated summary compared to the first and middle sentences. In contrast, the Qwen model (Figure 6b) 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**.

| Models | Generated Summary Bins 1 | 2 | 3 | 4 | 5 | Sensitivity |
|---|---|---|---|---|---|---|
| **Standard Summary** | | | | | | |
| Qwen2.5-7B-Instruct | **0.89** | 0.88 | 0.85 | 0.81 | 0.87 | -1.25 |
| *w/o sliding window attention* | **0.90** | 0.88 | 0.85 | 0.83 | 0.84 | 2.5(+3.75) |
| **Long Summary** | | | | | | |
| Qwen2.5-7B-Instruct | **0.86** | 0.83 | 0.83 | 0.82 | 0.84 | -0.50 |
| *w/o sliding window attention* | **0.86** | 0.84 | 0.83 | 0.82 | 0.81 | 2.75(+3.25) |

Table 6: Faithfulness scores and sensitivity for **standard** and **long** summaries generated by Qwen2.5-7B-Instruct on the Wikipedia dataset, with and without sliding window attention. Note that higher sensitivity indicates stronger hallucination toward the end of the output.

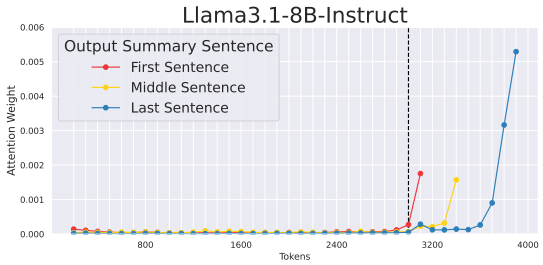
This consistent distribution in Qwen can be attributed to its sliding window attention mechanism (Child et al., 2019), which maintains a balanced focus across the context and mitigates the downstream attention concentration that triggers hallucinations. Table 6 compares Qwen with and without sliding window attention, where the sliding-window context length is fixed at 16,000 tokens. The results clearly indicate that removing the sliding window attention mechanism increases the sensitivity to 3.75 for standard summaries and 3.25 for long summaries. Furthermore, as shown in Figure 6c, the attention level becomes higher than that observed with sliding window attention. While autoregressive decoders generally allocate more attention to recent tokens, this trend is not uniform across models. Taken together, these results demonstrate that sliding window attention not only enhances decoding efficiency but also **reduces attention concentration on later tokens as the generated sequence length increases**.

## 5 How to Mitigate the *Hallucinate at the Last*?

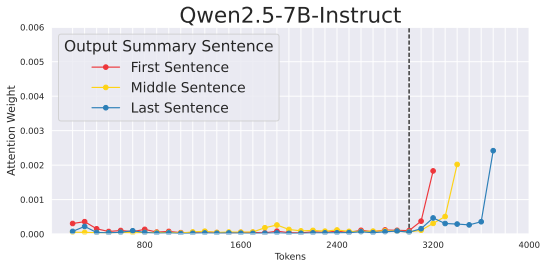
In the preceding section, we demonstrated that the *Hallucinate at the Last* phenomenon could be partially mitigated through the use of sliding window attention. In the following analysis, we investigate strategies to address this issue in models that are not trained with sliding window attention. To this end, we apply four methods that may mitigate the phenomenon.

**Experimental Setup** In this experiment, we generate Wikipedia summaries using four methods, taking Llama3.1-8B-Instruct as the baseline. For comparison, we evaluate the following four methods: (See Appendix F for more experimental details.)

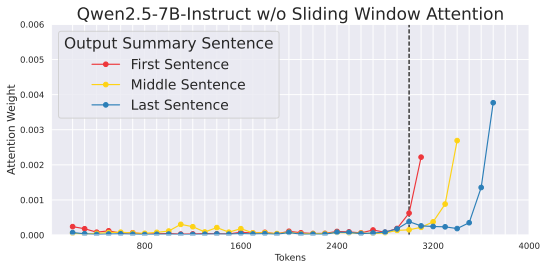
- **BOOOOKSCORE** (Chang et al., 2024a) segments the input context into chunks, generates summaries for each chunk individually, and then merges the partial summaries.
- **MINFERENCE** (Jiang et al., 2024) employs sparse attention mechanisms to efficiently process long input sequences.
- **LONGWRITER-LLAMA3.1-8B** (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 (Shi et al., 2024b).




(a) Average attention weights of summaries generated by **Llama**.




(b) Average attention weights of summaries generated by **Qwen** with sliding window attention.




(c) Average attention weights of summaries generated by **Qwen** without sliding window attention.

Figure 6: Average attention weights of two models exhibiting contrasting trends in Figure 2. The dashed line separates the input context from the generated output.

**Results & Analysis** We report the faithfulness scores across output bins and the corresponding sensitivity values in Table 7, and present the bin-wise faithfulness trends for each method in Figure 7. As shown in Table 7, BOOOOKSCORE achieves the lowest sensitivity (0.0), making it the closest to

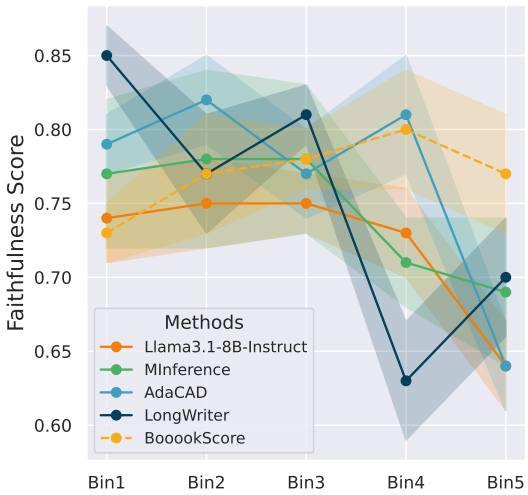



Figure 7: Faithfulness scores (evaluated by MiniCheck) for **long** summaries generated by different mitigation methods on the Wikipedia dataset across increasing output lengths.

zero among the four methods. Moreover, as illustrated in Figure 7, BOOOOKSCORE maintains the highest level of faithfulness, particularly around the final bin, 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. In addition, we report results for other evaluation dimensions on the summaries generated using BOOOOKSCORE in Table 8. The results show that **BOOOOKSCORE not only mitigates the *Hallucinate at the Last* phenomenon, but also improves the overall quality of the generated summaries.**

Nevertheless, this approach relies on chunk-wise generation, highlighting the need for future studies that maintain faithfulness in long responses.

## 6 Related Work

The problem of hallucination in LLMs has been a significant area of research. Prior work has largely

| METHODS | GENERATED SUMMARY BINS 1 | 2 | 3 | 4 | 5 | SENSITIVITY |
|---|---|---|---|---|---|---|
| Llama3.1-8B | 0.74 | **0.75** | **0.75** | 0.73 | 0.64 | 10.3 |
| + BOOOOKSCORE | 0.73 | 0.77 | 0.78 | **0.80** | 0.77 | 0.0 |
| + MINFERENCE | 0.77 | **0.78** | **0.78** | 0.71 | 0.69 | 7.0 |
| + LONGWRITER | **0.85** | 0.77 | 0.81 | 0.63 | 0.70 | 6.5 |
| + ADACAD | 0.79 | **0.82** | 0.77 | 0.81 | 0.64 | 15.8 |

Table 7: Faithfulness scores and sensitivity (evaluated by MiniCheck) for **long** summaries generated by different mitigation methods on the Wikipedia dataset.

| METHODS | FAITHFULNESS(%) | COMPLETENESS(%) | CONCISENESS(%) |
|---|---|---|---|
| Llama3.1-8B | 90.33 | 100.0 | 41.35 |
| BOOOOKSCORE | 94.33 | 100.0 | 47.83 |

Table 8: Summary-level faithfulness, completeness and conciseness scores of **long** summaries on the Wikipedia dataset, evaluated using FineSurE.

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 (Wan et al., 2023), decoding (Shi et al., 2024b), or prompting (Zhou et al., 2023). In addition, Wan et al. (2025) analyzed positional hallucinations in long-context summarization, focusing on settings with long input contexts rather than long-form output generation.

Positional bias in generation errors has also been observed in related domains. In particular, studies on Large Vision–Language Models report that hallucinations tend to increase toward the end of generated image descriptions (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., 2025; 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; Zhao et al., 2025). Nevertheless, it does not address positional context-aware faithfulness over extended summary lengths.

To the best of our knowledge, this work is the first to study where hallucinations occur in LLM-based long document summarization, revealing positional patterns, 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 faithfulness and motivates future research into positionally-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.

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, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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, and 1 others. 2024b. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers. *Preprint*, arXiv:1904.10509.

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.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 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. *Preprint*, arXiv:2405.14486.

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, and 1 others. 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*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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.

Chufan Shi, Haoran Yang, Deng Cai, Zhisong Zhang, Yifan Wang, Yujiu Yang, and Wai Lam. 2024a. A thorough examination of decoding methods in the era of LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8601–8629, Miami, Florida, USA. Association for Computational Linguistics.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024b. 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.

Rishabh Singh and Shirin Goshtasbpour. 2022. Plattbin: Efficient posterior calibrated training for NLP classifiers. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3673–3684, Dublin, Ireland. 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. *Preprint*, arXiv:2309.02427.

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, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Louis Rouillard, Thomas Mesnard, and 1 others. 2025. Gemma 3 technical report. *arXiv e-prints*, pages arXiv–2503.

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).

Yuhao Wu, Ming Shan Hee, Zhiqiang Hu, and Roy Ka-Wei Lee. 2025. 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, and 1 others. 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. *Preprint*, arXiv:2501.05414.

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. *Preprint*, arXiv:2405.17220.

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*.

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.

James Xu Zhao, Jimmy Z. J. Liu, Bryan Hooi, and See-Kiong Ng. 2025. How does response length affect long-form factuality. *Preprint*, arXiv:2505.23295.

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.

## A Attention Weight Calculation

This section provides a more detailed and formal explanation of our attention weight computation method introduced in Section 4.

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:

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

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:

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

## B Details for Atomic Fact Generation

This section details our experimental details and human evaluation results from Section 2.

**Prompt** The prompt template employed for atomic fact generation is shown in Table 16.

**Human Evaluation** Furthermore, we assess the quality of the atomic facts generated by the LLM through human evaluation, as reported in Table 9, focusing on atomic facts extracted from long summaries generated by Llama. For each context length, we randomly sampled 20 summaries, and the authors manually conducted the evaluation. Each atomic fact was independently annotated by the authors as either true or false, depending on whether it was supported by the corresponding summary. In total, 1,569 atomic facts were evaluated, and the two annotators agreed on 1,541 of them, corresponding to a 98.2% raw agreement. Because the task involves binary labels with a skewed class distribution, chance-corrected agreement coefficients are known to become unstable or difficult to interpret. We therefore report raw agreement as a clearer and more reliable indicator of annotation consistency in this setting.

| Context Length | Validation(%) |
|---|---|
| 4K | 95.92 |
| 5K | 99.01 |
| 6K | 96.18 |
| 7K | 99.12 |
| 8K | 99.64 |

Table 9: Human evaluation of the quality of atomic facts.

## C Details for Model Usage

In this section, we present the experimental details for the model usage in Section 3.

**Details for Evaluation Method** For atomic fact decomposition, we utilize the Llama3.1-8B-Instruct model with vLLM (Kwon et al., 2023)[2]. To further illustrate the evaluation process, Table 10 and Table 11 present concrete examples demonstrating how a given sentence is decomposed into atomic facts and subsequently assessed for faithfulness using MiniCheck.

As shown in the last sentence examples, atomic facts that receive low faithfulness scores are not necessarily grammatically incorrect or implausible. Rather, they reflect plausible-sounding but unsupported claims that the LLM introduces during generation. For instance, statements such as "Sara has a unique blend of skill" or "Sara inspires current players" may appear reasonable in context, yet they are absent from the source document and thus cannot be verified. This illustrates a key challenge in long-form generation: as the output grows longer, the model tends to generate fluent but *hallucinated* content that goes beyond what the source supports.

**Details for LLMs Usage** We utilized publicly available instruction-tuned models, including

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

| First Sentence | |
|---|---|
| Sara Sorribes Tormo is a talented Spanish professional tennis player born on October 8, 1996, in Castellón de la Plana, Spain. | |
| **Atomic Facts** | **Score** |
| Sara Sorribes Tormo is a Spanish professional tennis player. | 0.92 |
| Sara Sorribes Tormo was born in Spain. | 0.90 |
| Sara Sorribes Tormo was born on October 8. | 0.87 |
| Sara Sorribes Tormo was born in 1996. | 0.96 |
| Sara Sorribes Tormo was born in Castellón de la Plana. | 0.93 |
| **Avg. Score** | **0.92** |
| **Last Sentence** | |
| As Sara navigates her career, her unique blend of skill, perseverance, and spirit positions her as a promising athlete, inspiring both current and aspiring players. | |
| **Atomic Facts** | **Score** |
| Sara navigates her career. | 0.87 |
| Sara's career is that of an athlete. | 0.91 |
| Sara has a unique blend of skill. | 0.14 |
| Sara has perseverance. | 0.81 |
| Sara is a promising athlete. | 0.30 |
| Sara inspires current players. | 0.11 |
| Sara inspires aspiring players. | 0.15 |
| **Avg. Score** | **0.47** |

Table 10: Example of the evaluation procedure for a long summary generated by Llama using the MiniCheck-Flan-T5-Large model.

| First Sentence | |
|---|---|
| Shimoga district, officially designated as Shivamogga district, is a significant region in the Karnataka state of India, marked by its rich history, diverse geography, and vibrant culture. | |
| **Atomic Facts** | **Score** |
| Shimoga district is officially designated as Shivamogga district. | 0.97 |
| Shivamogga district is a region in Karnataka state. | 0.97 |
| Karnataka is a state in India. | 0.98 |
| Shimoga district has a rich history. | 0.86 |
| Shimoga district has diverse geography. | 0.91 |
| **Avg. Score** | **0.94** |
| **Last Sentence** | |
| As Shimoga district progresses into the future, the preservation of its cultural identity alongside economic development will be pivotal in ensuring Shimoga remains a vital part of Karnatakaś landscape. | |
| **Atomic Facts** | **Score** |
| Shimoga is a location. | 0.98 |
| The preservation of Shimoga's cultural identity is important. | 0.15 |
| Economic development is happening in Shimoga. | 0.57 |
| Shimoga's cultural identity and economic development are connected. | 0.05 |
| Shimoga is part of Karnataka. | 0.98 |
| **Avg. Score** | **0.55** |

Table 11: Example of the evaluation procedure for a long summary generated by GPT-4o mini using the MiniCheck-Flan-T5-Large model.

Llama3.1-8B-Instruct[3], Gemma3-12B-Instruct[4], and Qwen2.5-7B-Instruct[5] from HuggingFace. For Gemma, we set the sliding window attention context length to 1 token (as it cannot be disabled entirely), whereas for Qwen, we modify the default setting of over 131K tokens to 16K tokens. For the GPT-4o mini model, we used `gpt-4o-mini-2024-07-18`. All summaries were generated using greedy decoding with `float16` precision.

**Details for Bins** Throughout all experiments, we used five bins. The bins were computed using the `cut` function in the pandas package, and the scores of the atomic facts were assigned to bins using open interval boundaries, meaning that the boundary values were not included in the bins.

To ensure that our findings are not the result of the specific binning choice, we additionally report results using a finer-grained setting with ten bins in Table 12. The same trend persists even with this increased resolution, confirming that the *Hallucinate at the Last* phenomenon is robust to the choice of bin granularity.

## D More Results with Different Methods

We adopt a diverse set of methods to validate the robustness of our faithfulness evaluation framework. In this section, we present experimental results for faithfulness assessment models beyond MiniCheck, incorporating approaches based on NLI and ROUGE-L. Specifically, we substitute the scoring model $\mathcal{M}$ in Equation 2 of Section 3 with these alternative methods. Additionally, we evaluate using FineSurE, which assesses faithfulness at the sentence level using an LLM, rather than at the atomic fact level. Furthermore, we compare the metric scores with human judgments and find that MiniCheck exhibits a substantially higher correlation than FineSurE (Spearman's $\rho = 0.875$ vs. $0.375$), indicating that fine-grained atomic fact-level evaluation more closely aligns with human perception of factual consistency.

**Minicheck Based** We utilize MiniCheck-Flan-T5-Large Model[6] for the main evaluation method. We report the faithfulness scores and sensitivity for **standard** summaries generated by different models across output bins on the Wikipedia dataset in Table 17, and those of **long** summaries in Table 18.

**NLI Based** For the NLI model, we use the state-of-the-art hallucination evaluation model (Bao et al., 2024)[7].

We report the faithfulness scores and sensitivity assessed by NLI model for **standard** summaries

[3] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
[4] https://huggingface.co/google/gemma-3-12b-it
[5] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct
[6] https://huggingface.co/lytang/MiniCheck-Flan-T5-Large
[7] https://huggingface.co/vectara/hallucination_evaluation_model

| MODELS | GENERATED SUMMARY BINS | | | | | | | | | | SENSITIVITY |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | |
| Llama3.1-8B | 0.80 | 0.79 | **0.83** | 0.77 | 0.71 | 0.74 | 0.72 | 0.83 | 0.76 | 0.70 | 7.22 |
| Qwen2.5-7B | 0.85 | 0.85 | 0.80 | 0.82 | 0.85 | 0.82 | **0.89** | 0.81 | 0.89 | 0.86 | -1.78 |

Table 12: Faithfulness scores and sensitivity for **long** summaries generated by Llama and Qwen.

generated by different models across output bins on the Wikipedia dataset in Table 19, and those of **long** summaries in Table 20. As observed in the experimental results, consistent with the main results, the Llama and GPT models consistently exhibit high sensitivity, whereas the Qwen model occasionally demonstrates sensitivity values even below zero.

**ROUGE Based** We report the faithfulness scores and sensitivity assessed by ROUGE-L (Lin, 2004) for **standard** summaries generated by different models across output bins on the Wikipedia dataset in Table 23, and those of **long** summaries in Table 24.

**LLM Based** We present the faithfulness scores assessed by FineSurE framework for **standard** summaries generated by different models across output bins on the Wikipedia dataset in Table 21, and those of **long** summaries in Table 22. Note that unlike the other methods, which operate at the atomic fact level, FineSurE evaluates faithfulness at the sentence level using an LLM.

As observed in the experimental results, consistent with the main results, the Llama and GPT models consistently exhibit the highest error rates in the final bin. In contrast, the Qwen model does not yield the highest error rates in the final bin for both the cases of standard and long summaries.

## E More Results on Varying Datasets

**Minicheck Based** We report the faithfulness scores and sensitivity assessed by MiniCheck-Flan-T5-Large model of **standard** summaries generated by the Llama3.1-8B-Instruct model across output bins in multiple domains in Table 25, and those of **long** summaries in Table 26.

The experimental results reveal that in the case of standard summaries, sensitivities below zero are frequently observed; notably, in the Pubmed dataset, more than half of the values fall below zero. In contrast, such negative sensitivities are rarely observed in long summaries, with the Pubmed dataset in particular showing no sensitivities below zero, unlike the standard summary setting.

**NLI Based** We report the faithfulness scores and sensitivity assessed by NLI model for **standard** summaries generated by the Llama3.1-8B-Instruct model across output bins in multiple domains in Table 27, and those of **long** summaries in Table 28.

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

where $\texttt{attn}(x_i \rightarrow x_j)$ 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[8] with greedy decoding and `bfloat16` precision.

## G 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** Standard 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 13, and the prompt template employed for summary generation is shown in Table 15.

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

| CONTEXT LENGTH | WORDS RANGE |
|---|---|
| 4K | 800 to 1000 |
| 5K | 1000 to 1250 |
| 6K | 1200 to 1500 |
| 7K | 1400 to 1750 |
| 8K | 1600 to 2000 |

Table 13: Words range used in the prompt for long output generation.

| MODELS | AVG. WORDS | |
|---|---|---|
| | standard | long |
| Llama3.1-8B-Instruct | 279±24 | 834±399 |
| Qwen2.5-7B-Instruct | 153±39 | 751±204 |
| Mistral-7B-Instruct-v0.3 | 285±292 | 794±390 |
| GPT-4o mini | 172±8 | 925±132 |
| LongWriter-Llama3.1-8B | 398±171 | 3291±979 |

Table 14: The average word counts of both **standard** and **long** summaries generated by each model on the Wikipedia dataset.

**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 often produce excessively long summaries containing repeated sentences or even entire paragraphs. To ensure fair and reliable evaluation, we excluded such outputs and considered only summaries with comparable lengths in our analysis. Specifically, we compared summaries within each sentence-count range and selected those with the highest sentence counts as length-matched summaries. For example, in our setting, the length-matched summaries correspond to 28 sentences for the 4K-context case and 40 sentences for the 8K-context case.

Table 14 presents the average word counts of both standard 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.

## H Use of Large Language Model

This paper primarily utilized the AI assistant (ChatGPT-5.1) for grammar correction and refinement of clarity. It was not employed for content generation or research ideation.

| **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).<br>Use an unbiased and journalistic tone.<br>Text: {{Text}} |

Table 15: Prompt template used to generate summaries from the original document in Section 2.

| **Input Prompt** |
| --- |
| Decompose the text into atomic facts.<br>{{Text}} |

Table 16: Prompt template used to generate atomic facts from the summary in Section 2.

| MODELS | CONTEXT LENGTH | GENERATED SUMMARY BINS 1 | 2 | 3 | 4 | 5 | SENSITIVITY |
|---|---|---|---|---|---|---|---|
| Llama3.1-8B-Instruct | 1K | 0.81 | 0.79 | **0.85** | 0.82 | 0.79 | 2.75 |
| | 2K | **0.84** | 0.79 | 0.81 | 0.81 | 0.77 | 4.25 |
| | 3K | **0.87** | 0.85 | 0.84 | 0.83 | 0.79 | 5.75 |
| | 4K | **0.91** | 0.90 | 0.88 | 0.86 | 0.88 | 0.75 |
| | 5K | 0.85 | 0.85 | **0.86** | 0.83 | 0.83 | 1.75 |
| | 6K | **0.89** | 0.85 | 0.85 | 0.83 | 0.81 | 4.5 |
| | 7K | **0.91** | 0.81 | 0.85 | 0.87 | 0.81 | 5.0 |
| | 8K | **0.87** | 0.84 | 0.82 | 0.82 | 0.79 | 4.75 |
| Llama3.1-70B-Instruct | 1K | **0.87** | 0.83 | 0.86 | 0.80 | 0.77 | 7.0 |
| | 2K | **0.92** | 0.81 | 0.83 | 0.83 | 0.72 | 12.75 |
| | 3K | **0.90** | 0.82 | 0.86 | 0.77 | 0.83 | 0.75 |
| | 4K | **0.94** | 0.85 | 0.85 | 0.86 | 0.77 | 10.5 |
| | 5K | **0.91** | 0.86 | 0.85 | 0.83 | 0.89 | -2.75 |
| | 6K | **0.89** | 0.85 | 0.84 | 0.87 | 0.83 | 3.25 |
| | 7K | 0.87 | 0.88 | **0.90** | 0.83 | 0.84 | 3.0 |
| | 8K | **0.92** | 0.91 | 0.84 | 0.84 | 0.85 | 2.75 |
| Qwen2.5-7B-Instruct | 1K | **0.89** | 0.78 | 0.71 | 0.77 | 0.85 | -6.25 |
| | 2K | **0.90** | 0.89 | 0.83 | 0.78 | 0.84 | 1.0 |
| | 3K | **0.87** | 0.84 | 0.79 | 0.83 | 0.82 | 1.25 |
| | 4K | **0.91** | 0.88 | 0.90 | 0.79 | 0.89 | -2.0 |
| | 5K | 0.86 | **0.89** | 0.85 | 0.76 | **0.89** | -5.0 |
| | 6K | **0.93** | 0.91 | 0.92 | 0.80 | 0.90 | -1.0 |
| | 7K | **0.90** | 0.89 | 0.89 | 0.84 | 0.82 | 6.0 |
| | 8K | 0.89 | **0.93** | 0.88 | 0.89 | 0.92 | -2.25 |
| GPT-4o mini | 1K | **0.85** | 0.78 | 0.81 | 0.80 | 0.77 | 4.0 |
| | 2K | **0.87** | 0.86 | **0.87** | 0.82 | 0.80 | 5.5 |
| | 3K | **0.89** | 0.88 | 0.86 | 0.85 | 0.80 | 7.0 |
| | 4K | 0.79 | 0.80 | **0.86** | 0.83 | 0.82 | 0.0 |
| | 5K | **0.91** | 0.86 | 0.88 | 0.87 | 0.82 | 6.0 |
| | 6K | 0.85 | 0.89 | 0.85 | **0.90** | 0.85 | 2.25 |
| | 7K | **0.92** | 0.87 | 0.84 | 0.90 | 0.84 | 4.25 |
| | 8K | **0.90** | **0.90** | 0.84 | 0.85 | 0.82 | 5.25 |
| Gemma3-4B-Instruct | 1K | 0.88 | **0.89** | 0.85 | 0.71 | 0.77 | 6.25 |
| | 2K | 0.81 | **0.91** | 0.89 | 0.84 | 0.83 | 3.25 |
| | 3K | 0.80 | 0.88 | 0.87 | **0.92** | 0.77 | 9.75 |
| | 4K | 0.85 | 0.87 | **0.92** | 0.88 | 0.77 | 11.0 |
| | 5K | 0.90 | 0.83 | 0.83 | **0.92** | 0.82 | 5.0 |
| | 6K | **0.93** | 0.92 | 0.89 | 0.93 | 0.80 | 11.75 |
| | 7K | **0.91** | **0.91** | 0.90 | 0.83 | 0.81 | 7.75 |
| | 8K | 0.89 | 0.88 | 0.88 | **0.91** | 0.84 | 5.0 |
| Gemma3-12B-Instruct | 1K | 0.83 | **0.84** | 0.81 | 0.77 | 0.78 | 3.25 |
| | 2K | 0.83 | 0.74 | 0.79 | **0.85** | 0.72 | 8.25 |
| | 3K | **0.85** | 0.82 | 0.77 | **0.85** | 0.76 | 6.25 |
| | 4K | 0.91 | 0.87 | **0.93** | 0.85 | 0.84 | 5.0 |
| | 5K | 0.85 | 0.87 | **0.93** | 0.83 | 0.81 | 6.0 |
| | 6K | **0.94** | 0.89 | 0.93 | 0.90 | 0.87 | 4.5 |
| | 7K | **0.91** | 0.85 | 0.86 | 0.82 | 0.77 | 9.0 |
| | 8K | **0.90** | 0.85 | 0.80 | 0.85 | 0.78 | 7.0 |

Table 17: Comparison of MiniCheck-Flan-T5-Large based faithfulness scores and Sensitivity for **standard** 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.

| MODELS | CONTEXT LENGTH | GENERATED SUMMARY BINS 1 | 2 | 3 | 4 | 5 | SENSITIVITY |
|---|---|---|---|---|---|---|---|
| Llama3.1-8B-Instruct | 1K | 0.78 | **0.82** | 0.75 | 0.75 | 0.75 | 2.5 |
| | 2K | **0.82** | 0.79 | 0.75 | 0.76 | 0.73 | 5.0 |
| | 3K | 0.77 | **0.81** | 0.80 | **0.81** | 0.71 | 8.75 |
| | 4K | 0.79 | **0.80** | 0.73 | 0.77 | 0.74 | 3.25 |
| | 5K | 0.80 | 0.83 | 0.83 | **0.84** | 0.79 | 3.5 |
| | 6K | 0.80 | 0.82 | 0.77 | **0.83** | 0.72 | 8.5 |
| | 7K | 0.84 | **0.87** | 0.84 | 0.79 | 0.77 | 6.5 |
| | 8K | 0.84 | **0.85** | 0.82 | 0.82 | 0.77 | 6.25 |
| Llama3.1-70B-Instruct | 1K | 0.84 | 0.81 | **0.86** | 0.77 | 0.75 | 7.0 |
| | 2K | 0.79 | 0.83 | **0.89** | 0.84 | 0.74 | 9.75 |
| | 3K | 0.80 | **0.85** | 0.83 | 0.83 | 0.79 | 3.75 |
| | 4K | 0.83 | 0.81 | **0.85** | 0.82 | 0.74 | 8.75 |
| | 5K | **0.84** | 0.83 | 0.80 | 0.77 | 0.80 | 1.0 |
| | 6K | **0.83** | 0.81 | 0.80 | 0.79 | 0.79 | 1.75 |
| | 7K | 0.81 | 0.75 | 0.76 | **0.84** | 0.77 | 2.0 |
| | 8K | **0.87** | 0.86 | 0.84 | 0.85 | 0.79 | 6.5 |
| Qwen2.5-7B-Instruct | 1K | **0.88** | 0.76 | 0.81 | 0.83 | 0.83 | -1.0 |
| | 2K | 0.83 | **0.89** | 0.84 | 0.83 | 0.82 | 2.75 |
| | 3K | 0.83 | 0.84 | **0.86** | 0.79 | 0.84 | -1.0 |
| | 4K | **0.87** | 0.83 | 0.79 | 0.83 | 0.84 | -1.0 |
| | 5K | 0.87 | **0.89** | **0.89** | 0.80 | 0.85 | 1.25 |
| | 6K | **0.89** | 0.81 | 0.78 | 0.78 | 0.82 | -0.5 |
| | 7K | **0.86** | 0.78 | 0.79 | 0.83 | 0.84 | -2.5 |
| | 8K | **0.86** | 0.85 | 0.85 | 0.85 | 0.85 | 0.25 |
| GPT-4o mini | 1K | **0.85** | 0.81 | 0.81 | 0.74 | 0.80 | 0.25 |
| | 2K | **0.81** | 0.80 | 0.80 | 0.77 | 0.75 | 4.5 |
| | 3K | **0.83** | 0.81 | 0.82 | 0.74 | 0.73 | 7.0 |
| | 4K | **0.83** | 0.82 | 0.79 | 0.78 | 0.78 | 2.5 |
| | 5K | **0.84** | 0.82 | **0.84** | 0.81 | 0.75 | 7.75 |
| | 6K | **0.85** | 0.82 | 0.81 | 0.79 | 0.77 | 4.75 |
| | 7K | **0.82** | 0.79 | 0.81 | 0.79 | 0.76 | 4.25 |
| | 8K | **0.83** | **0.83** | 0.82 | **0.83** | 0.78 | 4.75 |
| Gemma3-4B-Instruct | 1K | 0.77 | 0.82 | 0.81 | **0.84** | 0.73 | 8.0 |
| | 2K | **0.86** | 0.72 | 0.80 | 0.85 | 0.76 | 4.75 |
| | 3K | **0.87** | 0.82 | 0.84 | 0.84 | 0.79 | 5.25 |
| | 4K | **0.86** | 0.85 | 0.78 | **0.86** | 0.80 | 3.75 |
| | 5K | **0.88** | 0.83 | 0.83 | 0.84 | 0.78 | 6.5 |
| | 6K | 0.83 | 0.85 | **0.86** | 0.85 | 0.81 | 3.75 |
| | 7K | 0.82 | 0.82 | **0.88** | 0.86 | 0.82 | 2.5 |
| | 8K | 0.87 | **0.89** | 0.84 | 0.83 | 0.84 | 1.75 |
| Gemma3-12B-Instruct | 1K | **0.90** | 0.83 | 0.80 | 0.82 | 0.81 | 2.75 |
| | 2K | 0.82 | 0.83 | 0.87 | **0.88** | 0.80 | 5.0 |
| | 3K | **0.87** | 0.84 | 0.84 | 0.84 | 0.76 | 8.75 |
| | 4K | 0.80 | 0.76 | **0.83** | 0.79 | 0.76 | 3.5 |
| | 5K | **0.86** | 0.79 | **0.86** | 0.79 | 0.80 | 2.5 |
| | 6K | 0.85 | **0.87** | 0.81 | 0.81 | 0.81 | 2.5 |
| | 7K | 0.88 | 0.88 | **0.89** | 0.86 | 0.82 | 5.75 |
| | 8K | **0.85** | 0.84 | **0.85** | 0.84 | 0.77 | 7.5 |

Table 18: Comparison of MiniCheck-Flan-T5-Large based 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.

| 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 | 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 |
| Qwen2.5-7B-Instruct | 1K | **0.84** | 0.75 | 0.77 | 0.76 | 0.71 | 7.0 |
| | 2K | 0.81 | **0.82** | 0.78 | 0.69 | 0.71 | 6.5 |
| | 3K | **0.78** | 0.76 | 0.72 | 0.73 | 0.68 | 6.75 |
| | 4K | **0.90** | 0.83 | 0.76 | 0.80 | 0.79 | 3.25 |
| | 5K | 0.83 | **0.86** | 0.74 | 0.76 | 0.79 | 0.75 |
| | 6K | **0.87** | 0.82 | 0.77 | 0.77 | 0.83 | -2.25 |
| | 7K | **0.86** | 0.82 | 0.79 | 0.79 | 0.75 | 6.5 |
| | 8K | 0.85 | **0.88** | 0.84 | 0.80 | 0.76 | 8.25 |
| GPT-4o mini | 1K | **0.79** | 0.69 | 0.76 | 0.73 | 0.65 | 9.25 |
| | 2K | **0.79** | 0.75 | 0.73 | 0.66 | 0.66 | 7.25 |
| | 3K | **0.84** | 0.81 | 0.80 | 0.73 | 0.71 | 8.5 |
| | 4K | 0.74 | **0.78** | 0.75 | 0.72 | 0.68 | 6.75 |
| | 5K | **0.86** | 0.81 | 0.79 | 0.75 | 0.68 | 12.25 |
| | 6K | 0.77 | **0.80** | 0.76 | 0.77 | 0.67 | 10.5 |
| | 7K | **0.87** | 0.85 | 0.75 | 0.77 | 0.68 | 13.0 |
| | 8K | **0.80** | **0.80** | 0.76 | 0.77 | 0.68 | 10.25 |
| Gemma3-12B-Instruct | 1K | 0.75 | 0.73 | **0.76** | 0.75 | 0.71 | 3.75 |
| | 2K | **0.80** | 0.65 | 0.70 | 0.68 | 0.65 | 5.75 |
| | 3K | **0.78** | 0.76 | 0.70 | 0.75 | 0.66 | 8.75 |
| | 4K | **0.82** | 0.70 | 0.76 | 0.76 | 0.69 | 7.0 |
| | 5K | 0.76 | 0.80 | **0.86** | 0.60 | 0.72 | 3.5 |
| | 6K | 0.81 | 0.79 | **0.84** | 0.78 | 0.69 | 11.5 |
| | 7K | **0.83** | 0.81 | 0.78 | 0.79 | 0.74 | 6.25 |
| | 8K | **0.83** | 0.81 | 0.78 | 0.79 | 0.74 | 6.25 |

Table 19: Comparison of NLI model based faithfulness scores and Sensitivity for **standard** 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.

| 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** | 0.74 | 0.76 | 0.75 | 0.68 | 9.0 |
| | 2K | 0.76 | **0.80** | 0.79 | 0.76 | 0.71 | 6.75 |
| | 3K | 0.79 | 0.79 | **0.81** | 0.71 | 0.75 | 2.5 |
| | 4K | 0.77 | 0.75 | 0.79 | **0.80** | 0.73 | 4.75 |
| | 5K | 0.83 | **0.84** | 0.79 | 0.76 | 0.76 | 4.5 |
| | 6K | **0.82** | 0.74 | 0.76 | 0.73 | 0.70 | 6.25 |
| | 7K | **0.80** | 0.74 | 0.72 | 0.78 | 0.78 | -2.0 |
| | 8K | 0.79 | 0.79 | **0.80** | 0.79 | 0.75 | 4.25 |
| 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 |
| Gemma3-12B-Instruct | 1K | **0.81** | 0.79 | 0.72 | 0.72 | 0.73 | 3.0 |
| | 2K | 0.77 | 0.75 | **0.79** | 0.78 | 0.74 | 3.25 |
| | 3K | 0.75 | **0.77** | 0.71 | 0.73 | 0.68 | 6.0 |
| | 4K | 0.72 | 0.67 | 0.70 | **0.74** | 0.65 | 5.75 |
| | 5K | 0.77 | 0.78 | **0.80** | 0.75 | 0.71 | 6.5 |
| | 6K | 0.77 | **0.78** | 0.77 | 0.74 | 0.70 | 6.5 |
| | 7K | **0.82** | 0.81 | **0.82** | 0.79 | 0.71 | 10.0 |
| | 8K | 0.74 | 0.74 | **0.77** | 0.75 | 0.63 | 12.0 |

Table 20: Comparison of NLI model based 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.

| MODELS | CONTEXT LENGTH | GENERATED SUMMARY BINS 1 | 2 | 3 | 4 | 5 | SENSITIVITY |
|---|---|---|---|---|---|---|---|
| Llama3.1-8B-Instruct | 1K | 0.76 | 0.71 | **0.86** | 0.76 | 0.67 | 10.25 |
| | 2K | 0.79 | 0.79 | **0.84** | 0.76 | 0.63 | 16.5 |
| | 3K | 0.80 | 0.70 | 0.80 | **0.90** | 0.43 | 38.5 |
| | 4K | **0.94** | 0.83 | 0.77 | 0.83 | 0.74 | 10.5 |
| | 5K | 0.69 | **0.80** | **0.80** | 0.64 | 0.67 | 11.25 |
| | 6K | 0.79 | 0.75 | 0.71 | **0.86** | 0.50 | 22.75 |
| | 7K | 0.76 | 0.83 | **0.86** | 0.69 | 0.34 | 41.5 |
| | 8K | 0.72 | **0.76** | 0.69 | **0.76** | 0.31 | 25.0 |
| Qwen2.5-7B-Instruct | 1K | 0.93 | **1.00** | 0.87 | **1.00** | 0.67 | 19.75 |
| | 2K | 0.87 | 0.80 | 0.87 | **1.00** | 0.87 | 3.25 |
| | 3K | 0.73 | 0.80 | 0.73 | 0.80 | **1.00** | -23.25 |
| | 4K | 0.73 | **1.00** | 0.80 | 0.87 | 0.80 | 5.0 |
| | 5K | 0.87 | **0.93** | 0.67 | 0.80 | 0.73 | 17.0 |
| | 6K | **0.93** | **0.93** | 0.73 | 0.80 | 0.87 | 4.5 |
| | 7K | 0.69 | **1.00** | 0.92 | **1.00** | **1.00** | -7.25 |
| | 8K | 0.69 | **0.77** | **0.77** | **0.77** | **0.77** | -8.0 |
| GPT-4o mini | 1K | 0.95 | **1.00** | **1.00** | **1.00** | 0.95 | 2.5 |
| | 2K | **1.00** | **1.00** | 0.90 | **1.00** | 0.90 | 7.5 |
| | 3K | 0.95 | **1.00** | **1.00** | **1.00** | 0.90 | 8.75 |
| | 4K | **1.00** | **1.00** | 0.95 | 0.95 | 0.85 | 8.75 |
| | 5K | 0.85 | **0.95** | **0.95** | **0.95** | 0.90 | 1.25 |
| | 6K | 0.90 | 0.90 | **1.00** | 0.95 | 0.85 | 10.0 |
| | 7K | 0.90 | **1.00** | **1.00** | 0.95 | 0.90 | 6.25 |
| | 8K | 0.85 | 0.85 | **0.90** | 0.80 | 0.80 | 6.25 |
| Gemma3-12B-Instruct | 1K | 0.88 | **1.00** | 0.88 | **1.00** | 0.88 | 6.0 |
| | 2K | 0.75 | 0.75 | **1.00** | 0.88 | 0.88 | -9.25 |
| | 3K | **1.00** | **1.00** | **1.00** | 0.88 | **1.00** | 6.0 |
| | 4K | **1.00** | **1.00** | **1.00** | 0.89 | **1.00** | 5.5 |
| | 5K | 0.88 | **1.00** | **1.00** | **1.00** | 0.88 | 9.25 |
| | 6K | **1.00** | **1.00** | **1.00** | **1.00** | 0.86 | 14.0 |
| | 7K | **1.00** | **1.00** | **1.00** | **1.00** | 0.67 | 33.25 |
| | 8K | **1.00** | **1.00** | **1.00** | **1.00** | 0.83 | 17.25 |

Table 21: Comparison of FineSurE faithfulness scores for **standard** 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.

| MODELS | CONTEXT LENGTH | GENERATED SUMMARY BINS 1 | 2 | 3 | 4 | 5 | SENSITIVITY |
|---|---|---|---|---|---|---|---|
| Llama3.1-8B-Instruct | 1K | 0.47 | 0.73 | 0.40 | **0.80** | 0.20 | 40.0 |
| | 2K | 0.20 | **0.47** | 0.33 | 0.20 | 0.00 | 30.0 |
| | 3K | 0.14 | **0.44** | **0.44** | 0.00 | 0.00 | 25.5 |
| | 4K | 0.11 | 0.34 | 0.23 | **0.56** | 0.00 | 31.0 |
| | 5K | 0.40 | **0.47** | 0.44 | 0.13 | 0.00 | 36.0 |
| | 6K | 0.17 | **0.33** | **0.33** | **0.33** | 0.00 | 29.0 |
| | 7K | 0.10 | **0.40** | 0.00 | 0.10 | 0.00 | 15.0 |
| | 8K | **0.33** | 0.08 | 0.25 | 0.17 | 0.00 | 20.75 |
| Qwen2.5-7B-Instruct | 1K | 0.77 | 0.69 | 0.77 | **0.85** | **0.85** | -5.75 |
| | 2K | 0.69 | 0.62 | 0.46 | 0.62 | **0.77** | -17.0 |
| | 3K | 0.75 | 0.75 | 0.58 | 0.67 | **0.83** | -14.0 |
| | 4K | 0.50 | **0.80** | **0.80** | 0.50 | 0.30 | 35.0 |
| | 5K | 0.80 | 0.70 | 0.70 | 0.60 | **1.00** | -30.0 |
| | 6K | **0.75** | 0.08 | 0.33 | 0.33 | 0.67 | -38.5 |
| | 7K | 0.36 | 0.18 | 0.18 | 0.36 | **0.73** | -28.5 |
| | 8K | **0.90** | **0.90** | 0.70 | 0.70 | **1.00** | -22.5 |
| GPT-4o mini | 1K | 0.80 | 0.75 | **0.90** | 0.80 | 0.85 | -3.75 |
| | 2K | **0.90** | 0.80 | 0.70 | 0.85 | 0.60 | 21.25 |
| | 3K | 0.57 | 0.64 | 0.43 | **0.79** | 0.36 | 33.75 |
| | 4K | 0.33 | 0.72 | 0.61 | **0.83** | 0.28 | 41.25 |
| | 5K | 0.69 | 0.50 | 0.75 | **0.88** | 0.69 | 6.25 |
| | 6K | **0.61** | 0.39 | 0.44 | 0.50 | 0.28 | 27.5 |
| | 7K | 0.50 | 0.29 | **0.64** | **0.64** | 0.50 | 14.25 |
| | 8K | **0.33** | 0.25 | 0.08 | 0.08 | 0.25 | -12.75 |
| Gemma3-12B-Instruct | 1K | **1.00** | **1.00** | **1.00** | **1.00** | 0.80 | 20.0 |
| | 2K | **1.00** | **1.00** | **1.00** | **1.00** | 0.88 | 12.0 |
| | 3K | 0.70 | 0.70 | **1.00** | 0.70 | 0.70 | 7.5 |
| | 4K | **0.60** | 0.40 | 0.20 | **0.60** | 0.00 | 45.0 |
| | 5K | 0.75 | **1.00** | 0.75 | 0.75 | 0.25 | 56.25 |
| | 6K | **0.88** | **0.88** | 0.75 | 0.62 | **0.88** | -6.25 |
| | 7K | **1.00** | **1.00** | **1.00** | **1.00** | 0.86 | 14.0 |
| | 8K | **1.00** | 0.83 | **1.00** | **1.00** | 0.67 | 45.75 |

Table 22: Comparison of FineSurE faithfulness scores 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.

| MODELS | CONTEXT LENGTH | GENERATED SUMMARY BINS 1 | 2 | 3 | 4 | 5 | SENSITIVITY |
|---|---|---|---|---|---|---|---|
| Llama3.1-8B-Instruct | 1K | **0.66** | 0.63 | 0.63 | 0.62 | 0.57 | 6.5 |
| | 2K | **0.70** | 0.62 | 0.64 | 0.60 | 0.57 | 7.0 |
| | 3K | **0.70** | 0.64 | 0.63 | 0.63 | 0.61 | 4.0 |
| | 4K | 0.60 | **0.66** | 0.63 | 0.61 | 0.58 | 4.5 |
| | 5K | **0.70** | 0.65 | 0.66 | 0.64 | 0.61 | 5.25 |
| | 6K | **0.70** | 0.67 | 0.69 | 0.66 | 0.62 | 6.0 |
| | 7K | **0.74** | 0.67 | 0.64 | 0.65 | 0.60 | 7.5 |
| | 8K | **0.68** | 0.65 | 0.60 | 0.64 | 0.55 | 9.25 |
| Qwen2.5-7B-Instruct | 1K | **0.70** | 0.58 | 0.65 | 0.61 | 0.58 | 5.5 |
| | 2K | **0.81** | 0.61 | 0.64 | 0.57 | 0.52 | 13.75 |
| | 3K | **0.70** | 0.62 | 0.57 | 0.60 | 0.54 | 8.25 |
| | 4K | **0.73** | 0.68 | 0.63 | 0.59 | 0.57 | 8.75 |
| | 5K | **0.71** | 0.65 | 0.66 | 0.57 | 0.57 | 7.75 |
| | 6K | **0.70** | 0.60 | 0.62 | 0.62 | 0.58 | 5.5 |
| | 7K | **0.72** | 0.57 | 0.61 | 0.61 | 0.54 | 8.75 |
| | 8K | **0.66** | 0.60 | 0.64 | 0.61 | 0.59 | 3.75 |
| GPT-4o mini | 1K | **0.68** | 0.60 | 0.55 | 0.51 | 0.51 | 7.5 |
| | 2K | **0.73** | 0.58 | 0.58 | 0.57 | 0.50 | 11.5 |
| | 3K | **0.72** | 0.59 | 0.58 | 0.57 | 0.53 | 8.5 |
| | 4K | **0.68** | 0.60 | 0.55 | 0.56 | 0.52 | 7.75 |
| | 5K | **0.68** | 0.61 | 0.60 | 0.56 | 0.52 | 9.25 |
| | 6K | **0.67** | 0.61 | 0.60 | 0.57 | 0.52 | 9.25 |
| | 7K | **0.68** | 0.62 | 0.58 | 0.57 | 0.53 | 8.25 |
| | 8K | **0.69** | 0.62 | 0.56 | 0.56 | 0.49 | 11.75 |
| Gemma3-12B-Instruct | 1K | **0.64** | 0.55 | 0.56 | 0.51 | 0.62 | -5.5 |
| | 2K | **0.65** | 0.56 | 0.52 | 0.53 | 0.54 | 2.5 |
| | 3K | **0.65** | 0.63 | 0.55 | 0.55 | 0.53 | 6.5 |
| | 4K | **0.63** | 0.59 | 0.57 | 0.56 | 0.52 | 6.75 |
| | 5K | 0.61 | 0.57 | **0.62** | 0.57 | 0.53 | 6.25 |
| | 6K | **0.69** | 0.63 | 0.60 | 0.56 | 0.54 | 8.0 |
| | 7K | **0.68** | 0.61 | 0.63 | 0.67 | 0.56 | 8.75 |
| | 8K | **0.68** | 0.61 | 0.63 | 0.67 | 0.56 | 8.75 |

Table 23: Comparison of ROUGE-L based faithfulness scores and Sensitivity for **standard** 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.

| MODELS | CONTEXT LENGTH | GENERATED SUMMARY BINS 1 | 2 | 3 | 4 | 5 | SENSITIVITY |
|---|---|---|---|---|---|---|---|
| Llama3.1-8B-Instruct | 1K | **0.65** | 0.62 | 0.60 | 0.61 | 0.58 | 4.0 |
| | 2K | 0.65 | **0.66** | 0.64 | 0.62 | 0.57 | 7.25 |
| | 3K | 0.65 | **0.66** | 0.65 | 0.64 | 0.56 | 9.0 |
| | 4K | 0.62 | 0.61 | 0.60 | **0.64** | 0.60 | 1.75 |
| | 5K | 0.65 | **0.66** | 0.65 | 0.63 | 0.58 | 6.75 |
| | 6K | 0.65 | **0.71** | **0.71** | 0.69 | 0.61 | 8.0 |
| | 7K | **0.68** | **0.68** | 0.64 | 0.64 | 0.60 | 6.0 |
| | 8K | 0.62 | **0.67** | **0.67** | **0.67** | 0.62 | 3.75 |
| Qwen2.5-7B-Instruct | 1K | **0.69** | 0.57 | 0.58 | 0.62 | 0.52 | 9.5 |
| | 2K | **0.62** | **0.62** | 0.60 | 0.60 | 0.54 | 7.0 |
| | 3K | **0.66** | 0.60 | 0.60 | 0.62 | 0.56 | 6.0 |
| | 4K | 0.63 | 0.61 | 0.61 | **0.66** | 0.57 | 5.75 |
| | 5K | 0.64 | **0.65** | 0.63 | 0.63 | 0.60 | 3.75 |
| | 6K | **0.65** | 0.63 | 0.64 | 0.61 | 0.57 | 6.25 |
| | 7K | 0.65 | **0.66** | 0.62 | 0.60 | 0.57 | 6.25 |
| | 8K | 0.64 | **0.66** | 0.63 | 0.61 | 0.57 | 6.5 |
| GPT-4o mini | 1K | **0.69** | 0.57 | 0.55 | 0.58 | 0.52 | 7.75 |
| | 2K | **0.61** | 0.55 | 0.54 | 0.53 | 0.50 | 5.75 |
| | 3K | **0.59** | 0.55 | 0.55 | 0.50 | 0.47 | 7.75 |
| | 4K | **0.62** | 0.58 | 0.56 | 0.55 | 0.53 | 4.75 |
| | 5K | **0.59** | 0.55 | 0.55 | 0.52 | 0.46 | 9.25 |
| | 6K | **0.59** | 0.56 | 0.55 | 0.53 | 0.49 | 6.75 |
| | 7K | **0.58** | 0.56 | 0.53 | 0.51 | 0.49 | 5.5 |
| | 8K | **0.58** | 0.54 | 0.53 | 0.50 | 0.47 | 6.75 |
| Gemma3-12B-Instruct | 1K | **0.66** | 0.57 | 0.60 | 0.61 | 0.47 | 14.0 |
| | 2K | **0.63** | 0.60 | 0.60 | 0.61 | 0.57 | 4.0 |
| | 3K | **0.59** | **0.59** | 0.54 | 0.57 | 0.56 | 1.25 |
| | 4K | **0.61** | 0.56 | 0.58 | 0.58 | 0.58 | 0.25 |
| | 5K | 0.62 | 0.60 | 0.61 | **0.63** | 0.54 | 7.5 |
| | 6K | 0.60 | **0.63** | 0.61 | 0.59 | 0.57 | 3.75 |
| | 7K | **0.65** | **0.65** | 0.60 | 0.61 | 0.57 | 5.75 |
| | 8K | 0.60 | **0.64** | 0.60 | 0.60 | 0.56 | 5.0 |

Table 24: Comparison of ROUGE-L based 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 | 0.81 | 0.79 | **0.85** | 0.82 | 0.79 | 2.75 |
| | 2K | **0.84** | 0.79 | 0.81 | 0.81 | 0.77 | 4.25 |
| | 3K | **0.87** | 0.85 | 0.84 | 0.83 | 0.79 | 5.75 |
| | 4K | **0.91** | 0.90 | 0.88 | 0.86 | 0.88 | 0.75 |
| | 5K | 0.85 | 0.85 | **0.86** | 0.83 | 0.83 | 1.75 |
| | 6K | **0.89** | 0.85 | 0.85 | 0.83 | 0.81 | 4.5 |
| | 7K | **0.91** | 0.81 | 0.85 | 0.87 | 0.81 | 5.0 |
| | 8K | **0.87** | 0.84 | 0.82 | 0.82 | 0.79 | 4.75 |
| Arxiv | 1K | 0.71 | 0.79 | 0.72 | **0.80** | 0.69 | 6.5 |
| | 2K | **0.82** | 0.71 | 0.76 | 0.69 | 0.75 | -0.5 |
| | 3K | 0.74 | **0.87** | 0.78 | 0.82 | 0.77 | 3.25 |
| | 4K | **0.87** | 0.86 | 0.84 | 0.85 | 0.79 | 6.5 |
| | 5K | 0.84 | **0.85** | 0.79 | 0.82 | 0.81 | 1.5 |
| | 6K | 0.82 | 0.83 | 0.77 | 0.85 | **0.90** | -8.25 |
| | 7K | 0.86 | 0.82 | **0.89** | 0.84 | 0.79 | 6.25 |
| | 8K | 0.80 | **0.81** | 0.75 | 0.78 | **0.81** | -2.5 |
| Pubmed | 1K | **0.88** | 0.87 | **0.88** | 0.79 | 0.86 | -0.5 |
| | 2K | **0.89** | 0.85 | 0.81 | 0.87 | 0.85 | 0.5 |
| | 3K | **0.91** | 0.89 | 0.86 | **0.91** | 0.90 | -0.75 |
| | 4K | **0.93** | 0.90 | 0.92 | 0.87 | 0.92 | -1.5 |
| | 5K | **0.91** | 0.84 | 0.89 | **0.91** | 0.85 | 3.75 |
| | 6K | 0.91 | 0.88 | **0.94** | 0.91 | 0.91 | 0.0 |
| | 7K | 0.91 | 0.89 | **0.92** | 0.91 | **0.92** | -1.25 |
| | 8K | 0.91 | **0.92** | 0.89 | 0.87 | 0.90 | -0.25 |
| Govreport | 1K | **0.87** | 0.82 | 0.86 | 0.83 | 0.81 | 3.5 |
| | 2K | **0.94** | 0.88 | 0.90 | 0.93 | 0.93 | -1.75 |
| | 3K | 0.91 | **0.92** | 0.90 | 0.91 | 0.91 | 0.0 |
| | 4K | **0.94** | 0.83 | 0.88 | 0.85 | 0.84 | 3.5 |
| | 5K | **0.94** | 0.93 | **0.94** | 0.93 | 0.93 | 0.5 |
| | 6K | 0.93 | 0.91 | 0.92 | **0.94** | 0.87 | 5.5 |
| | 7K | 0.91 | 0.89 | 0.94 | **0.97** | 0.89 | 3.75 |
| | 8K | 0.90 | **0.93** | **0.93** | 0.90 | **0.93** | -1.5 |

Table 25: Comparison of MiniCheck-Flan-T5-Large based faithfulness scores and Sensitivity for **standard** 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 | 0.78 | **0.82** | 0.75 | 0.75 | 0.75 | 2.5 |
| | 2K | **0.82** | 0.79 | 0.75 | 0.76 | 0.73 | 5.0 |
| | 3K | 0.77 | **0.81** | 0.80 | **0.81** | 0.71 | 8.75 |
| | 4K | 0.79 | **0.80** | 0.73 | 0.77 | 0.74 | 3.25 |
| | 5K | 0.80 | 0.83 | 0.83 | **0.84** | 0.79 | 3.5 |
| | 6K | 0.80 | 0.82 | 0.77 | **0.83** | 0.72 | 8.5 |
| | 7K | 0.84 | **0.87** | 0.84 | 0.79 | 0.77 | 6.5 |
| | 8K | 0.84 | **0.85** | 0.82 | 0.82 | 0.77 | 6.25 |
| Arxiv | 1K | 0.73 | **0.81** | **0.81** | 0.73 | 0.72 | 5.0 |
| | 2K | **0.80** | 0.75 | 0.78 | 0.72 | 0.75 | 1.25 |
| | 3K | **0.78** | 0.77 | 0.76 | 0.75 | 0.72 | 4.5 |
| | 4K | 0.74 | **0.80** | 0.70 | 0.78 | 0.74 | 1.5 |
| | 5K | 0.81 | 0.80 | **0.84** | 0.83 | 0.77 | 5.0 |
| | 6K | 0.70 | **0.77** | 0.72 | 0.72 | 0.74 | -1.25 |
| | 7K | **0.86** | 0.81 | 0.76 | 0.74 | 0.71 | 8.25 |
| | 8K | 0.79 | 0.74 | **0.81** | 0.77 | 0.68 | 9.75 |
| Pubmed | 1K | 0.89 | 0.87 | **0.93** | 0.87 | 0.89 | 0.0 |
| | 2K | 0.84 | 0.87 | 0.81 | **0.88** | 0.81 | 4.0 |
| | 3K | 0.86 | 0.89 | **0.90** | **0.90** | 0.87 | 1.75 |
| | 4K | 0.83 | **0.89** | 0.84 | 0.86 | 0.79 | 6.5 |
| | 5K | **0.87** | **0.87** | 0.86 | 0.85 | 0.85 | 1.25 |
| | 6K | 0.87 | **0.90** | 0.86 | 0.81 | 0.80 | 6.0 |
| | 7K | **0.91** | 0.89 | 0.87 | 0.82 | 0.82 | 5.25 |
| | 8K | **0.90** | 0.82 | 0.86 | 0.81 | 0.71 | 13.75 |
| Govreport | 1K | 0.88 | 0.86 | 0.86 | **0.89** | 0.85 | 2.25 |
| | 2K | 0.89 | 0.90 | 0.88 | **0.91** | **0.91** | -1.5 |
| | 3K | 0.87 | **0.92** | 0.91 | 0.91 | 0.88 | 2.25 |
| | 4K | 0.90 | 0.89 | **0.94** | 0.90 | 0.82 | 8.75 |
| | 5K | **0.94** | 0.91 | 0.90 | 0.91 | 0.89 | 2.5 |
| | 6K | 0.91 | 0.91 | 0.90 | **0.92** | **0.92** | -1.0 |
| | 7K | 0.93 | **0.94** | 0.93 | 0.93 | 0.91 | 2.25 |
| | 8K | 0.92 | 0.93 | **0.95** | 0.92 | 0.85 | 8.0 |

Table 26: Comparison of MiniCheck-Flan-T5-Large based 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 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 27: Comparison of NLI model based faithfulness scores and Sensitivity for **standard** 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 | 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 |
| Arxiv | 1K | 0.77 | **0.82** | 0.81 | 0.80 | 0.74 | 6.0 |
| | 2K | 0.76 | 0.78 | **0.80** | 0.72 | 0.71 | 5.5 |
| | 3K | 0.75 | 0.79 | **0.81** | 0.78 | 0.68 | 10.25 |
| | 4K | 0.71 | **0.81** | 0.74 | 0.77 | 0.73 | 2.75 |
| | 5K | 0.77 | 0.83 | **0.85** | 0.81 | 0.74 | 7.5 |
| | 6K | 0.69 | **0.82** | 0.79 | 0.75 | 0.75 | 1.25 |
| | 7K | **0.86** | **0.86** | 0.80 | 0.80 | 0.77 | 6.0 |
| | 8K | 0.79 | 0.81 | **0.85** | 0.82 | 0.73 | 8.75 |
| Pubmed | 1K | 0.75 | **0.85** | 0.84 | 0.80 | 0.78 | 3.0 |
| | 2K | 0.76 | **0.83** | 0.82 | **0.83** | 0.72 | 9.0 |
| | 3K | 0.81 | 0.85 | **0.86** | 0.78 | 0.76 | 6.5 |
| | 4K | 0.77 | **0.82** | 0.79 | 0.79 | 0.68 | 11.25 |
| | 5K | 0.78 | **0.81** | **0.81** | 0.78 | 0.70 | 9.5 |
| | 6K | 0.80 | **0.87** | 0.83 | 0.77 | 0.74 | 7.75 |
| | 7K | **0.85** | **0.85** | 0.84 | 0.79 | 0.77 | 6.25 |
| | 8K | 0.80 | 0.79 | **0.81** | 0.77 | 0.67 | 12.25 |
| Govreport | 1K | 0.80 | 0.80 | 0.85 | **0.87** | 0.78 | 5.0 |
| | 2K | 0.82 | **0.85** | 0.83 | **0.85** | 0.78 | 5.75 |
| | 3K | 0.81 | 0.86 | **0.89** | 0.81 | 0.82 | 2.25 |
| | 4K | 0.76 | **0.86** | **0.86** | 0.82 | 0.77 | 5.5 |
| | 5K | 0.82 | 0.85 | **0.88** | 0.82 | 0.78 | 6.25 |
| | 6K | 0.86 | **0.88** | **0.88** | 0.87 | 0.80 | 7.25 |
| | 7K | 0.86 | **0.91** | 0.89 | 0.88 | 0.83 | 5.5 |
| | 8K | 0.84 | 0.87 | **0.89** | 0.84 | 0.80 | 6.0 |

Table 28: Comparison of NLI model based 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 underlined.