LongNovel: A Multi-Scale Benchmark for Hallucination Detection in Long-Context Novel Summarization

Anonymous ACL submission

Abstract

Although context windows have expanded significantly in recent years, hallucinations in long-context summarization remain a challenge. Long novels are better suited than news or papers for researching these hallucinations, due to their intrinsic information and detailed descriptions of events and dialogues. However, current research lacks a multi-scale benchmark for hallucination detection in long-context novel summarization and does not fully explore how hallucinations change as the context grows longer. In this study, we propose LongNovel, a multi-scale long-context Chinese novel benchmark for hallucination detection. This benchmark is constructed from 29 books, ranging from 2k to 100k tokens. We design 8 hallucination types and employ a combination of Multi-Model Arbitration and Entity-Referenced Hallucination Generation to ensure both data authenticity and a balanced distribution of hallucination categories. Furthermore, we manually revise the contents in the test set to guarantee data reliability. Extensive experimental results demonstrate that LongNovel is a challenging benchmark. We release LongNovel for future research.[1]

1 Introduction

While the expansion of context windows for Large Language Models (LLMs) to 100k tokens or more (Chen et al., 2024b; Peng et al., 2024; Ding et al., 2024) has enabled the processing of long-form content, this increased capacity does not inherently resolve the issue of hallucinations in long-context summarization (Kim et al., 2024; Belém et al., 2025; Pal et al., 2023). Novel summarization is well-suited for researching hallucinations in long-context summarization because it requires inferring implicit information from dialogues and events, which is more complex than processing the explicit data found in news or academic papers (Karpinska et al., 2024a; Kryscinski et al., 2022; Kim and Kim, 2025). While traditional metrics like ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2020) are limited to lexical or semantic similarity, other NLI-based approaches such as SummaC (Laban et al., 2022) and AlignScore (Zha et al., 2023) often fail to detect long-context hallucinations due to their on short input windows. Consequently, there is an urgent need for more precise evaluation models that can identify hallucinations in long-context scenarios. However, the development of robust evaluation models relies on the availability of high-quality benchmarks. Therefore, constructing a long-text, multi-scale hallucination detection dataset will facilitate the identification of more reliable evaluation models.

However, existing research faces two primary challenges. First, constructing high-quality benchmarks for hallucination detection in long-context novels is hindered by the cost of manual annotation. Traditional datasets such as NOCHA (Karpinska et al., 2024b) and StorySumm (Subbiah et al., 2024) rely on human-labeled hallucination data, which is labor-intensive and time-consuming. While synthetic datasets like LCHD (Liu et al., 2025) effectively reduce annotation costs, they fail to reach the 100k-token scale. Second, the evolution of hallucinations as context length increases remains largely unexplored. Most existing benchmarks lack a multi-scale design capable of evaluating model robustness across varying lengths. While Fables (Kim et al., 2024) covers the 100k-token scale, it is largely restricted to a single length. Although Clipper (Pham et al., 2025) introduces a multi-scale approach with book-level and chapter-level claims at the 100k scale, it is not designed for summary hallucination detection.

To address these issues, we introduce **LongNovel**, a multi-scale benchmark for hallucination

[1] https://anonymous.4open.science/r/LongNovel-60B194

Benchmark	Summ. Halluc.	100K	Auto. Label	Diff. Len.
Nocha (2024a)	✗	✓	✗	✗
StorySumm (2024)	✓	✗	✗	✗
FABLES (2024)	✓	✓	✗	✗
LCHD (2025)	✓	✗	✓	✗
CLIPPER (2025)	✗	✓	✓	✓
LongNovel (Ours)	✓	✓	✓	✓

Table 1: Comparison of our benchmark with other benchmarks in novel. 'Summ. Halluc.', '100K', 'Auto. Gen.', and 'Diff. Len.' mean whether it is a summarization hallucination detection dataset, whether it reaches up to 100K tokens, whether the hallucinated data is generated through automated methods, and whether it encompasses different levels of length, respectively. LCHD (Liu et al., 2025) refers to the long-context hallucination detection dataset.

detection in long-context novel summarization. Based on a corpus of 29 books, we construct five long context scenarios: S(2k∼4k), M(16k), L(32k), XL(64k), and XXL(100K), totaling 650 samples in the test set. Building on these scenarios, eight hallucination types have been designed. Each summary includes: whether the summary contains hallucinations, hallucination types, identified hallucinated sentences, the corrected content, and the reasons for identifying the hallucinations. Notably, we use human-written summaries as ground truth to guide the LLM generation to ensure data reliability. The framework of our benchmark is illustrated in Fig. 1.

Considering data authenticity and a balanced distribution of hallucination types, we use two complementary methods to construct the benchmark. One method is Multi-Model Arbitration, using Qwen2.5-7B-instruct (Yang et al., 2024) for summary generation and cross-model verification for data labeling, which captures more authentic hallucinations found in LLM outputs. Another is Entity-Referenced Hallucination Construction, where we perturb the human-written summary by extracting entities and use LLMs to craft hallucinations based on eight specific prompts, ensuring a balanced distribution across all hallucination types. Finally, to ensure data reliability, we conduct human revision of the test set. The advantages of ours compared with previous long-context novel/-book benchmarks are listed in Table 1. We evaluate several state-of-the-art models and conduct experiments using different methods on LongNovel. The results show that LongNovel serves as a challenging benchmark for current models. Overall, our contributions are as follows:

- We introduce LongNovel, a multi-scale Chinese benchmark for hallucination detection in long-context novel summarization across five levels of length, containing eight hallucination types.
- We implement a construction approach that combines Multi-Model Arbitration with Entity-Referenced Hallucination Construction. This methodology ensures the dataset features real-world authenticity while covering various hallucination types.
- We evaluate state-of-the-art models and methods on LongNovel, revealing the challenges of long-context hallucination detection and providing a challenging benchmark for future research.

2 RELATED WORK

2.1 Hallucination Detection Benchmarks

Current methodologies for constructing hallucination datasets can be categorized into two approaches. The first involves generation by LLM, followed by manual annotation to identify hallucinated samples (Laban et al., 2023; Karpinska et al., 2024a; Subbiah et al., 2024; Akbar et al., 2024; Tang et al., 2024b; Chen et al., 2024a; Bao et al., 2025; Abdaljalil et al., 2025). The advantage of this approach is that the resulting hallucination patterns closely align with the model's performance in the real world. However, it relies heavily on high-quality annotation, making it both time-consuming and resource-intensive (Qi et al., 2025).

The second approach is automated hallucination injection based on existing reference materials, such as books or summaries. Early methods like (Kryscinski et al., 2020) generate negative samples by employing entity substitution and negation insertion. Cao and Wang (2021) select system outputs with low likelihood scores as negative samples. Recent works perform hallucination synthesis based on the instruction-following capabilities of LLMs (Tang et al., 2024a; Liu et al., 2025; Ming et al., 2025). Pham et al. (2025) extract summaries or outlines from original book content, thereby inducing LLMs to generate true-false claim pairs and corresponding reasoning chains.

2.2 Hallucination Detection Methods

Existing hallucination detection methods can be categorized into short-text and long-text detection based on the length of the processed content. In

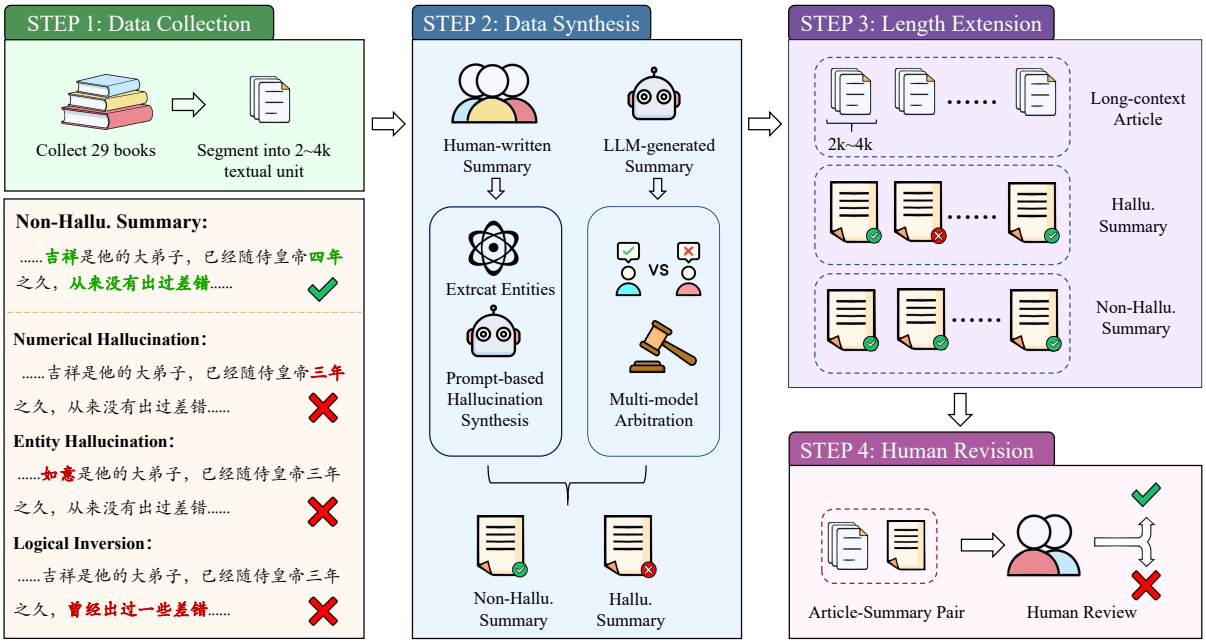

Figure 1: The framework of LongNovel.

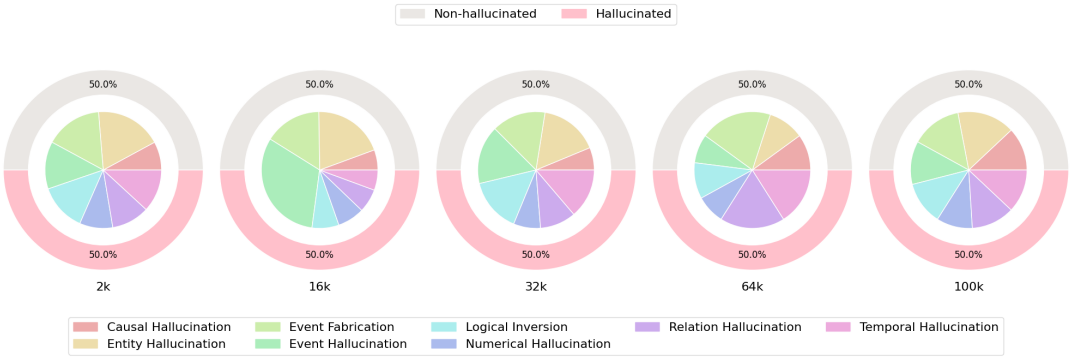

Figure 2: Distribution of hallucination types across different context lengths.

the short-text detection, Laban et al. (2022) evaluate factual consistency by decomposing documents into sentence pairs and computing NLI-based entailment scores. Zha et al. (2023) enhance cross-task generalization through large-scale alignment pre-training, and Liu et al. (2023c) introduce an evaluation framework based on Atomic Content Units. Additionally, QA-based methods (Deutsch et al., 2021; Scialom et al., 2021) verify informational faithfulness by measuring the answer consistency between source texts and generated summaries.

In long-text hallucination detection, many short-text methods are constrained by input window limitations. Approaches address this by either directly leveraging LLMs for binary classification or employing Chain-of-Thought (CoT) (Wei et al., 2022) to provide step-by-step analysis before reaching a final judgement. Liu et al. (2023b) utilize LLMs to score content based on preset indicators, demonstrating a high correlation with human judgment. Min et al. (2025) introduce a debate-based framework by assigning specific roles to LLMs, such as Advocate, Skeptic, and Adjudicator, to enhance data reliability. RAG-based systems, which are used to verify faithfulness in QA tasks (Zhang et al., 2025; Hu et al., 2025; Laban et al., 2024), can also detect hallucinations in summarization (Min et al., 2023).

3 Longnovel Construction

3.1 Data Collection

We collect 29 books from open-source data on the Chinese internet. Each book possesses a coherent plot, making it highly suitable for long-text consistency detection. Each book $B = \{u_1, u_2, \ldots, u_n\}$

is segmented into a series of textual units (chapters or paragraphs), where the length of each unit u_i ranges from 2k to 4k tokens. We employ 16 annotators. For each textual unit, one annotator drafts an initial summary, which is then revised by two other annotators to ensure accuracy and faithfulness to the source text. Ultimately, each unit u_i is paired with a corresponding human-written summary, denoted as s_i. For each unit, we also generate a summary using the Qwen 2.5-7B model (Yang et al., 2024). These model-generated summaries may contain hallucinations.

3.2 Data Synthesis

We categorize hallucinations into eight types: Entity Hallucination, Numerical Hallucination, Relation Hallucination, Logical Inversion, Event Hallucination, Temporal Hallucination, Causal Hallucination, and Event Fabrication (detailed in Appendix A). To ensure data authenticity and a balanced distribution of hallucination types, we employ two methods to generate our LongNovel benchmark.

Multi-Model Arbitration. To obtain more realistic hallucination data, inspired by MSum-Bench (Min et al., 2025), we implement a Multi-Model Arbitration method to label summaries generated by Qwen2.5-7B-instruct (Yang et al., 2024). In this process, GPT-4o (OpenAI, 2024) and Claude 3 Opus (Anthropic, 2024) first conduct independent evaluations. Specifically, they determine whether the output is a hallucinated summary h_i, which contains inconsistencies, or a non-hallucinated summary g_i, which remains faithful to the original source. If their outputs match, the result is accepted. Otherwise, Gemini-2.5-Pro (Comanici et al., 2025) serves as the arbitrator to resolve the inconsistency.

Entity-Referenced Hallucination Generation. We first extract entities such as names, organizations, and numbers from the human-written summary s_i. Then, with the extracted entities as references, GPT-4o is employed to rewrite s_i into a corresponding hallucinated summary h_i based on eight prompts, each corresponding to a specific hallucination type, as shown in Appendix I.1. Simultaneously, GPT-4o generates a non-hallucinated summary g_i to enhance data diversity.

In the construction of our dataset, summaries without hallucinations are labeled as non-hallucinated summaries. For hallucinated samples, we provide the hallucination type, the hallucinated sentence, the corrected version, and the reason for the judgment.

3.3 Length Extension

To evaluate the model's performance across various lengths, we use Qwen3 (Yang et al.) tokenizer to process textual units and categorize the dataset into five ranges: 2k, 16k, 32k, 64k, and 100k.

Assuming the target length requires k consecutive textual units, let the input context be $X = [u_t, u_{t+1}, \ldots, u_{t+k-1}]$. Each unit u_i corresponds to three summary versions: a human-written summary s_i, a non-hallucinated version g_i generated by GPT-4o, and a hallucinated version h_i. We construct $Y_{pos}^{(j)}$ and $Y_{neg}^{(j)}$ as the hallucination-free and hallucinated summary sequences for X respectively by replacing the j-th textual unit's summary:

$$Y_{pos}^{(j)} = [s_t, \ldots, s_{j-1}, h_j, s_{j+1}, \ldots, s_{t+k-1}]$$

$$Y_{neg}^{(j)} = [s_t, \ldots, s_{j-1}, g_j, s_{j+1}, \ldots, s_{t+k-1}]$$

The hallucination types and the annotations of the hallucinated sentences in the negative samples are identical to those in h_j.

3.4 Human Revision

To guarantee the reliability of the benchmark, an annotator reviews the entire test set with assistance from Gemini-2.5-Pro to identify potential hallucinations and corrects them if necessary. To ensure the quality, two annotators validate 186 summaries from the test set. The inter-annotator agreement reaches 0.946 for the binary classification of whether a summary contains hallucinations, demonstrating the high quality of the final benchmark. More details w.r.t. human revision are shown in Appendix C.

3.5 Dataset Statistics

The benchmark consists of 8,104 samples, split into a training set of 7,454 samples and a test set of 650 samples. Table 2 shows the statistics of LongNovel. The token sequence length of the test set is detailed in Table 3.

To evaluate model performance across different error types, we analyze the distribution of hallucination categories within the test set. As illustrated in Fig. 2, the distribution of hallucination types remains relatively balanced across various

[2] https://github.com/openai/tiktoken

Context	Train			Test		
Length	Hallu.	Non-H.	Total	Hallu.	Non-H.	Total
S	1,511	1,107	2,618	75	75	150
M	1,147	1,010	2,157	75	75	150
L	1,379	1,300	2,679	75	75	150
XL	–	–	–	50	50	100
XXL	–	–	–	50	50	100
Total		**7,454**			**650**	

Table 2: Statistics of LongNovel. Hallu. and Non-H. represent the counts of hallucinated and non-hallucinated samples. Ratio indicates the ratio of the source article length to the summary length.

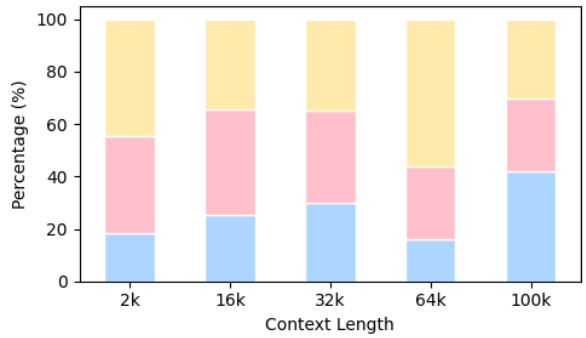

Figure 3: Proportional distribution of hallucination positions (beginning, middle, and end) within the input context. Blue, red, and yellow represent the beginning, middle, and end, respectively.

text lengths. Furthermore, we examine the relative position of hallucinations within the input context. As shown in Fig. 3, hallucinations in the test set are uniformly distributed across the beginning, middle, and end segments of the text. This uniform distribution is critical for long-context evaluation as it compels the model to process the entire context comprehensively rather than relying on positional shortcuts, such as the lost-in-the-middle phenomenon (Liu et al., 2024), where models tend to focus only on the beginning and end of a document.

4 Experiments

4.1 Baselines

We evaluate several state-of-the-art models with strong Chinese language capabilities on the benchmark. The Open-Source Models include the InternLM series (Cai et al.) (InternLM2.5-20B-chat, InternLM3-8B-instruct), GLM-4-9B (GLM et al., 2024), Llama3.1-8B (Grattafiori et al., 2024), and the Qwen3 series (Yang et al.) (8B, 14B, and 32B). For Commercial Models, we include the GPT series (GPT-4.1 and GPT-5), Claude-4-sonnet, the DeepSeek series (DeepSeek-v3 (DeepSeek-AI et al.), DeepSeek-r1 (DeepSeek-AI et al., 2025)), and Gemini-2.5-Flash (Comanici et al., 2025).

4.2 Experimental Setup

We employ vLLM (Kwon et al., 2023) for the inference of open-source models across all benchmarks. For the 32k, 64k, and 100k tests, we implement YaRN (Peng et al., 2024) scaling, expanding the context length from 32k to 128k. For commercial APIs, the temperature is set to 0 to ensure consistent outputs. To guarantee the reliability of our results, each experiment for the open-source models is repeated at least three times. Details are in Appendix E.1.

4.3 Evaluation Metrics

We evaluate performance using binary accuracy for both hallucinated and non-hallucinated classes. Specifically, we compute the Average Accuracy (Acc.), defined as the arithmetic mean of the accuracies achieved on hallucinated and non-hallucinated instances, to provide a balanced measure of classification performance. In addition to Average Accuracy, we employ the Matthews Correlation Coefficient (MCC.) to further assess the model's performance in hallucination detection. It balances the trade-off between false positives and false negatives by considering all categories of the confusion matrix.

4.4 Compared Methods

We conduct experiments using different methods on LongNovel.

Zero-shot Prompting. We provide both the source article and the summary to the LLMs, along with a prompt defining the criteria that guide the model in distinguishing between hallucinated and non-hallucinated summaries. We define two prompt types: target summary at the Beginning (Prompt-B) and target summary at the End (Prompt-E), as shown in Appendix I.2.

Chain-of-Thought (CoT). Based on a zero-shot setting, we implement a CoT (Wei et al., 2022) prompting strategy to elicit the model's reasoning capabilities. The model is required to generate a step-by-step analysis of the factual consistency between the source text and the summary before providing the final hallucination detection result. The prompt is shown in Appendix I.2.

Model	S (n=150)			M (n=150)			L (n=150)			XL (n=100)			XXL (n=100)		
	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max
InternLM-3	2.21	3.48	4.76	13.15	17.41	20.42	33.48	35.99	40.12	66.81	70.24	71.41	97.25	107.66	111.33
GLM-4	1.99	3.06	4.26	12.13	15.28	17.53	29.09	31.44	34.52	60.11	62.28	64.12	89.48	96.34	98.78
Qwen-3	2.01	3.15	4.34	12.76	15.67	17.99	29.63	32.38	35.49	62.99	64.85	66.21	94.31	100.44	102.68
GPT-4o	3.52	5.35	7.16	20.99	26.22	30.02	53.75	58.01	63.92	105.29	116.53	124.05	154.99	170.02	175.64

Table 3: Token sequence length of LongNovel test set (values are in thousands, i.e., k). GPT-4o uses tiktoken[2].

Supervised Fine-Tuning (SFT). To enable models to adapt to more positions and activate extrapolation ability, we fine-tune the models with the LongNovel training set by gradually increasing the context length follow findings from (Wei et al., 2025). More details are in Appendix E.2.

Retrieval-Augmented Generation (RAG). In RAG-based hallucination detection, a common approach is to retrieve the most similar chunks from the source article to serve as evidence supporting each sentence in the summary. However, a significant limitation of this method is that the retrieved segments may lack temporal coherence, failing to preserve the corresponding chronological changes of events. For instance, if a summary sentence describes an early-stage event but is matched with text chunks from a later stage, models may fail to identify whether there is a hallucination due to the misalignment.

Accordingly, we design a RAG framework using a sliding-window mechanism guided by semantic similarity, which dynamically anchors summary segments to their most relevant article context. With this framework, we can obtain the summary chunks and their corresponding article chunks. This process allows us to extract paired summary segments and article chunks, where a summary is classified as non-hallucinated if and only if every individual segment is verified as consistent with its respective document window. We set the block size to 5,500 characters for the source article and 150 characters for the summary. More details are shown in Appendix D and Appendix G.

5 Experimental Results and Analysis

5.1 Main Results

We calculate Accuracy and MCC. across various context lengths for both open-source and commercial models, and the detailed performance statistics are presented in Table 4. Full results are shown in Appendix F. Negative MCC. values such as Llama-3.1-8B-instruct result from formatting failures, primarily repetitive outputs, that prevent valid JSON generation. We treat such instances as incorrect predictions. From the results, we can draw the following conclusions:

(1) Performance exhibits a downward trend as context length increases. Almost all models show a significantly higher average accuracy at 2k context compared to 100k. For instance, InternLM2.5-20B drops from 68.67% at 2k to 49.33% at 32k. At the 64k and 100k stages, some negative Matthews Correlation Coefficient scores shows the decline in instruction-following capability. Models produce repetitive outputs or generate novel summaries instead of answering the questions in JSON format, as detailed in Appendix H.4 and Appendix H.5.

(2) Chain-of-Thought prompts enhance hallucination detection accuracy in short-context scenarios. However, this performance gain diminishes in 64k and 100k context and even leads to a reduction in accuracy. For example, DeepSeek-v3 raises from 58.00% to 74.67% at 2k but drops from 51.00% to 44.00% at 100k. This indicates that reasoning steps help resolve hallucinations in short-context scenarios.

(3) In the 2k to 100k range, the SFT models achieve average accuracy significantly higher than both the base and CoT versions, reaching 60.00% at 100k compared to only 53.00% for the base version. This demonstrates that specialized fine-tuning of open-source models for length extrapolation is an effective strategy for hallucinations detection in long-context tasks.

(4) RAG strategy shows limitation in 64k context and although its performance scales slightly better at 100k, where GPT-4.1 with RAG achieves a MCC. of 0.2521, outperforming its base and CoT counterparts, it still fails to match the overall performance of SFT at both scales. This may be because dividing summaries into smaller chunks increases the number of decisions the model must gives, leading to the accumulation of errors from individual judgments.

Model	S (2k ∼ 4k)		M (16k)		L (32k)		XL (64k)		XXL (100k)	
	Acc.	MCC.	Acc.	MCC.	Acc.	MCC.	Acc.	MCC.	Acc.	MCC.
Open-Source Models										
InternLM2.5-20B	68.67	0.3734	60.67	0.2287	49.33	-0.0142	49.00	-0.1005	44.00	-0.1267
InternLM3-8B-instruct	60.67	0.3282	50.00	0.0000	49.33	-0.0819	51.00	0.1005	51.00	0.1005
GLM-4-9B	55.33	0.1833	50.89	0.0934	49.56	-0.0546	50.00	0.0000	50.00	0.0000
Llama3.1-8B-instruct	45.56	-0.1041	40.89	-0.1831	3.33	-0.9341	0.00	-1.0000	0.00	-1.0000
——SFT	73.56	0.4965	59.11	0.1824	50.00	0.0000	12.00	-0.7829	0.00	-1.0000
——RAG	–	–	–	–	–	–	50.00	0.0000	50.00	0.0000
Qwen3-8B	65.78	0.3164	61.56	0.2907	48.89	-0.0228	53.00	0.0673	53.00	0.0693
——CoT	68.44	0.3698	60.89	0.2597	57.11	0.1456	54.33	0.0971	52.67	0.0540
——SFT	77.33	0.5516	65.56	0.3137	**66.00**	**0.3333**	**61.00**	**0.2368**	**60.00**	**0.3077**
——RAG	–	–	–	–	–	–	50.00	0.0000	50.00	0.0000
Qwen3-14B	64.22	0.3216	55.33	0.1183	54.22	0.1189	45.67	-0.1696	52.67	0.0456
——CoT	66.00	0.3563	62.22	0.2564	56.00	0.1497	53.00	0.0638	46.67	-0.0737
Qwen3-32B	69.11	0.4217	67.11	0.3423	49.11	-0.0220	53.00	0.0645	54.00	0.1153
——CoT	74.89	0.5245	64.44	0.2891	51.78	0.0420	43.00	-0.1430	54.67	0.0971
——RAG	–	–	–	–	–	–	50.00	0.0000	50.00	0.0000
Commercial Models										
DeepSeek-v3	58.00	0.2949	50.00	0.0000	48.00	-0.0426	50.00	0.0000	51.00	0.0459
——CoT	74.67	0.4998	60.67	0.2343	57.33	0.1909	57.00	0.1960	44.00	-0.1286
DeepSeek-r1	63.33	0.3922	52.00	0.0711	48.67	-0.0278	52.00	0.1429	51.00	0.0349
——CoT	74.00	0.4874	64.00	0.2905	53.33	0.0923	48.00	-0.0737	45.00	-0.1400
Gemini-2.5-Flash	70.67	0.4455	50.67	0.0371	50.67	0.0819	47.00	-0.1176	49.00	-0.021
——CoT	68.67	0.4002	50.67	0.0476	50.67	0.0819	49.00	-0.1005	49.00	-0.1005
Claude-4-sonnet	59.33	0.3024	61.33	0.2584	54.00	0.0849	54.00	0.0937	50.00	0.0000
——CoT	80.67	0.6160	64.67	0.3580	49.33	-0.0281	52.00	0.0436	51.00	0.0266
——RAG	–	–	–	–	–	–	50.00	0.0000	56.00	0.1562
GPT-4.1	79.33	0.5867	62.67	0.2541	52.67	0.0533	53.00	0.0765	51.00	0.1005
——CoT	**81.33**	**0.6614**	64.67	0.3764	56.00	0.2133	51.00	0.1005	51.00	0.1005
——RAG	–	–	–	–	–	–	51.00	0.1005	59.00	0.2521
GPT-5	67.33	0.3860	**68.00**	**0.3766**	61.33	0.2287	59.00	0.1845	51.00	0.0220
——CoT	69.33	0.4219	64.00	0.2985	54.67	0.0936	54.00	0.0873	44.00	-0.1209
——RAG	–	–	–	–	–	–	50.00	0.0000	57.00	0.2081

Table 4: Accuracy and Matthews Correlation Coefficient of LongNovel. **Acc.** denotes the binary accuracy calculated as the mean of accuracy for hallucinated and non-hallucinated instances and is reported as a percentage (%). **MCC.** represents Matthews Correlation Coefficient . The best performing score is highlighted in **bold** and second-best is underline.

5.2 Analysis and Case Study

Which Hallucination Types are Easier to Detect? As shown in Figure 4, Event Fabrication stands out as the most detectable hallucination type for many models. For instance, Qwen3-8B-SFT achieves its highest recall of 0.74 in this category. Temporal Hallucinations are tend to be the most difficult to detect. Many models, such as InternLM3-8B-instruct and GLM-4-9B-chat, show recall rates near 0.00 for these types. This indicates that content completely absent from the source text is relatively easy to detect. However, when the content exist in the source but the chronological order of events is altered, it becomes significantly more challenging to identify.

Which Hallucination Positions are Easier to Detect? To mitigate the potential influence of hallucination types on position-based performance, we selected four hallucination types (Numerical Hallucination, Logical inversion, Event Hallucination and Entity Hallucination) with median Recall scores based on the results in Fig. 6. We then calculate the hallucination detection recall at different positions (Beginning, Middle, and End) in various context lengths (2k, 16k, 32k, 64k, and 100k), as illustrated in Fig. 5.

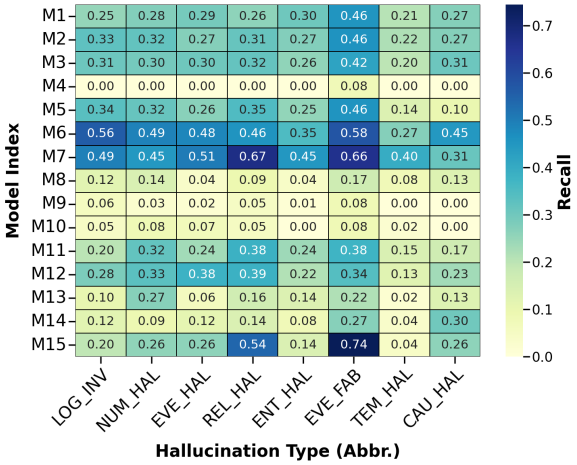

Figure 4: Recall performance of various LLMs across different hallucination types. Note that models are abbreviated as follows: M1: Claude-4-sonnet, M2: DeepSeek-r1, M3: DeepSeek-v3, M4: GLM-4-9B-chat, M5: GPT-4.1, M6: GPT-5, M7: Gemini-2.5-Flash, M8: InternLM2.5-20B-chat, M9: InternLM3-8B-instruct, M10: Llama3.1-8B-instruct, M11: Qwen3-14B, M12: Qwen3-32B, M13: Qwen3-8B, M14: Llama3.1-8B-instruct-SFT, M15: Qwen3-8B-SFT. The hallucination types are abbreviated as follows: EVE_HAL: Event Hallucination, REL_HAL: Relation Hallucination, LOG_INV: Logical Inversion, CAU_HAL: Causal Hallucination, ENT_HAL: Entity Hallucination, NUM_HAL: Numerical Hallucination, TEM_HAL: Temporal Hallucination, and EVE_FAB: Event Fabrication.

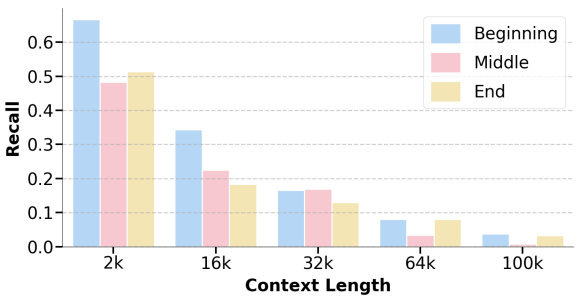

Figure 5: Recall score of hallucination detection across different positions, averaged over four hallucination types.

Hallucinations in the middle position yield the lowest recall at 2k, 64k, and 100k lengths. At the 16k scale, the recall for the middle also remains significantly lower than that of the beginning. While this pattern exhibits some fluctuations across different scales, for instance, at 32k, the middle recall (0.165) is slightly higher than that of the end (0.142), the overall results confirm that the hallucinations at the beginning of the context are the most detectable.

Why Hallucination Detection Fails in Long-context Scale? By analyzing outputs of models at the long-context scale, which are shown in Appendix H, we find several error patterns.

First, models may fail to comprehensively process or may misread both the source text and its corresponding summary, directly resulting in wrong detection outcomes. Second, in some cases, the model identifies a hallucination yet produces a correction that is identical to the original sentence. However, most of these cases are faithful and do not need corrections. Third, a deficiency in reasoning capabilities prevents models from successfully mapping a series of specific actions or dialogues to a correct abstract generalization. Furthermore, repetitive output patterns sometimes cause the model response to exceed length constraints, leading to truncated and incomplete answers. Additionally, some cases fail to complete the hallucination detection, indicating a drop in instruction-following performance. We also observe internal logical contradictions where a model initially identifies a hallucination but ultimately concludes that no such hallucination exists when giving the reason for the judgment. Finally, models often fail to recognize semantic equivalence, misidentify a summary as a hallucination due to lexical changes or omitted peripheral information, despite the core summary remaining factually accurate.

6 Conclusion

We introduce LongNovel, a Chinese long-context dataset for hallucination detection in novels, based on human-annotated summaries. It comprises five subsets ranging from 2k to 100k tokens, totaling 650 human-revised samples. Our extensive experiments on LongNovel reveal that current large language models still lack sufficient capability in long-context hallucination detection tasks. We hope that LongNovel will provide useful insights for future research in this field.

Limitations

While the LongNovel dataset ensures plot coherence through concatenated sequences, the construction of negative samples for hallucination detection is primarily constrained within a $2k$ to $4k$ token range, which may impose a ceiling on the model's performance. Furthermore, our study is

limited to open-source models with up to 32B parameters. Consequently, the generalization capabilities of large-scale models, such as Qwen2.5-72B and Llama-3.1-70B, have not yet been investigated. Finally, as LongNovel is a Chinese-language dataset, the applicability of our findings to cross-lingual tasks and non-novel domains remains to be further validated in future research.

References

Samir Abdaljalil, Hasan Kurban, and Erchin Serpedin. 2025. Halluverse25: Fine-grained multilingual benchmark dataset for LLM hallucinations. *CoRR*, abs/2503.07833.

Shayan Ali Akbar, Md Mosharaf Hossain, Tess Wood, Si-Chi Chin, Erica Salinas, Victor Alvarez, and Erwin Cornejo. 2024. Hallumeasure: Fine-grained hallucination measurement using chain-of-thought reasoning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 15020–15037. Association for Computational Linguistics.

Anthropic. 2024. Introducing the next generation of claude. Accessed: 2024-03-04.

Forrest Sheng Bao, Miaoran Li, Renyi Qu, Ge Luo, Erana Wan, Yujia Tang, Weisi Fan, Manveer Singh Tamber, Suleman Kazi, Vivek Sourabh, Mike Qi, Ruixuan Tu, Chenyu Xu, Matthew Gonzales, Ofer Mendelevitch, and Amin Ahmad. 2025. Faithbench: A diverse hallucination benchmark for summarization by modern llms. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 2: Short Papers, Albuquerque, New Mexico, April 29 - May 4, 2025*, pages 448–461. Association for Computational Linguistics.

Catarina G. Belém, Pouya Pezeshkpour, Hayate Iso, Seiji Maekawa, Nikita Bhutani, and Estevam Hruschka. 2025. From single to multi: How llms hallucinate in multi-document summarization. In *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 5276–5309. Association for Computational Linguistics.

Zheng Cai, Maosong Cao, Haojiong Chen, and others. Internlm2 technical report.

Shuyang Cao and Lu Wang. 2021. CLIFF: contrastive learning for improving faithfulness and factuality in abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6633–6649. Association for Computational Linguistics.

Xiang Chen, Duanzheng Song, Honghao Gui, Chenxi Wang, Ningyu Zhang, Yong Jiang, Fei Huang, Chengfei Lyu, Dan Zhang, and Huajun Chen. 2024a. Factchd: Benchmarking fact-conflicting hallucination detection. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*, pages 6216–6224. ijcai.org.

Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. 2024b. Longlora: Efficient fine-tuning of long-context large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, and others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Tri Dao. 2024. Flashattention-2: Faster attention with better parallelism and work partitioning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, and others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, and others. Deepseek-v3 technical report.

Daniel Deutsch, Tania Bedrax-Weiss, and Dan Roth. 2021. Towards question-answering as an automatic metric for evaluating the content quality of a summary. *Trans. Assoc. Comput. Linguistics*, 9:774–789.

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. 2024. Longrope: Extending LLM context window beyond 2 million tokens. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Team GLM, :, Aohan Zeng, Bin Xu, Bowen Wang, and others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, and others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Wentao Hu, Wengyu Zhang, Yiyang Jiang, Chen Jason Zhang, Xiaoyong Wei, and Qing Li. 2025. Removal of hallucination on hallucination: Debate-augmented RAG. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages

15839–15853. Association for Computational Linguistics.

Marzena Karpinska, Katherine Thai, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024a. One thousand and one pairs: A "novel" challenge for long-context language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17048–17085, Miami, Florida, USA. Association for Computational Linguistics.

Marzena Karpinska, Katherine Thai, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024b. One thousand and one pairs: A "novel" challenge for long-context language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 17048–17085. Association for Computational Linguistics.

Hyuntak Kim and Byung-Hak Kim. 2025. Nexussum: Hierarchical LLM agents for long-form narrative summarization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 10120–10157. 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. *Preprint*, arXiv:2404.01261.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 9332–9346. Association for Computational Linguistics.

Wojciech Kryscinski, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. 2022. BOOKSUM: A collection of datasets for long-form narrative summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6536–6558, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles, SOSP 2023, Koblenz, Germany, October 23-26, 2023*, pages 611–626. ACM.

Philippe Laban, Alexander R. Fabbri, Caiming Xiong, and Chien-Sheng Wu. 2024. Summary of a haystack: A challenge to long-context llms and RAG systems. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 9885–9903. Association for Computational Linguistics.

Philippe Laban, Wojciech Kryscinski, Divyansh Agarwal, Alexander R. Fabbri, Caiming Xiong, Shafiq Joty, and Chien-Sheng Wu. 2023. Summedits: Measuring LLM ability at factual reasoning through the lens of summarization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 9662–9676. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

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.

Hao Liu, Matei Zaharia, and Pieter Abbeel. 2023a. Ring attention with blockwise transformers for near-infinite context. *CoRR*, abs/2310.01889.

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

Siyi Liu, Kishaloy Halder, Zheng Qi, Wei Xiao, Nikolaos Pappas, Phu Mon Htut, Neha Anna John, Yassine Benajiba, and Dan Roth. 2025. Towards long context hallucination detection. In *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 7827–7835. Association for Computational Linguistics.

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

Yixin Liu, Alexander R. Fabbri, Yilun Zhao, Pengfei Liu, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023c. Towards interpretable and efficient automatic reference-based summarization evaluation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 16360–16368. Association for Computational Linguistics.

Hyangsuk Min, Yuho Lee, Minjeong Ban, Jiaqi Deng, Nicole Hee-Yeon Kim, Taewon Yun, Hang Su, Jason Cai, and Hwanjun Song. 2025. Towards

multi-dimensional evaluation of LLM summarization across domains and languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14417–14450, Vienna, Austria. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei 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, EMNLP 2023, Singapore, December 6-10, 2023*, pages 12076–12100. Association for Computational Linguistics.

Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. 2025. Faitheval: Can your language model stay faithful to context, even if "the moon is made of marshmallows". In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024. Fine-grained hallucination detection and editing for language models. *Preprint*, arXiv:2401.06855.

OpenAI. 2024. Hello gpt-4o. Accessed: 2024.

Yury Orlovskiy, Camille Thibault, Anne Imouza, Jean-François Godbout, Reihaneh Rabbany, and Kellin Pelrine. 2024. Uncertainty resolution in misinformation detection. *Preprint*, arXiv:2401.01197.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.

Arka Pal, Deep Karkhanis, Manley Roberts, Samuel Dooley, Arvind Sundararajan, and Siddartha Naidu. 2023. Giraffe: Adventures in expanding context lengths in llms. *CoRR*, abs/2308.10882.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2024. Yarn: Efficient context window extension of large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Chau Minh Pham, Yapei Chang, and Mohit Iyyer. 2025. CLIPPER: compression enables long-context synthetic data generation. *CoRR*, abs/2502.14854.

Siya Qi, Rui Cao, Yulan He, and Zheng Yuan. 2025. Evaluating llms' assessment of mixed-context hallucination through the lens of summarization. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 16480–16503. Association for Computational Linguistics.

Jie Ren, Samyam Rajbhandari, Reza Yazdani Aminabadi, Olatunji Ruwase, Shuangyan Yang, Minjia Zhang, Dong Li, and Yuxiong He. 2021. Zero-offload: Democratizing billion-scale model training. In *Proceedings of the 2021 USENIX Annual Technical Conference, USENIX ATC 2021, July 14-16, 2021*, pages 551–564. USENIX Association.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. Questeval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6594–6604. Association for Computational Linguistics.

Melanie Subbiah, Faisal Ladhak, Akankshya Mishra, Griffin Adams, Lydia B. Chilton, and Kathleen R. McKeown. 2024. STORYSUMM: evaluating faithfulness in story summarization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 9988–10005. Association for Computational Linguistics.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024a. Minicheck: Efficient fact-checking of llms on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 8818–8847. Association for Computational Linguistics.

Liyan Tang, Igor Shalyminov, Amy Wing-mei Wong, Jon Burnsky, Jake W. Vincent, Yuan Yang, Siffi Singh, Song Feng, Hwanjun Song, Hang Su, Lijia Sun, Yi Zhang, Saab Mansour, and Kathleen McKeown. 2024b. Tofueval: Evaluating hallucinations of llms on topic-focused dialogue summarization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 4455–4480. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Lingxiao Wei, He Yan, Xiangju Lu, Junmin Zhu, Jun Wang, and Wei Zhang. 2025. Cnnsum: Exploring long-context summarization with large language models in chinese novels. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 8034–8062. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, and others. Qwen3 technical report.

An Yang, Baosong Yang, Beichen Zhang, and others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. Alignscore: Evaluating factual consistency with A unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 11328–11348. Association for Computational Linguistics.

Jiajie Zhang, Yushi Bai, Xin Lv, Wanjun Gu, Danqing Liu, Minhao Zou, Shulin Cao, Lei Hou, Yuxiao Dong, Ling Feng, and Juanzi Li. 2025. LongCite: Enabling LLMs to generate fine-grained citations in long-context QA. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5098–5122, Vienna, Austria. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Haosheng Zou, Xiaowei Lv, Shousheng Jia, Lin Li, Xiaochun Gong, and Xiangzheng Zhang. 2025. 360-llama-factory: Plug & play sequence parallelism for long post-training.

A Hallucination Type

To construct a comprehensive hallucination detection dataset, we categorize hallucinations observed in novel summarization into eight distinct types by referring to related research (Kryscinski et al., 2020; Pagnoni et al., 2021; Orlovskiy et al., 2024; Mishra et al., 2024).

Entity Hallucination Entity Hallucination occurs when a summary contains factual errors regarding the entities mentioned in the source text. This encompasses pronominal errors, subject-object role reversals, and erroneous descriptions of entities such as characters, organizations, or locations.

Numerical Hallucination Numerical Hallucination occurs when a summary introduces numerical data that is inconsistent with the source text. This encompasses difference in quantities, ages, or monetary values.

Relation Hallucination Relation Hallucination occurs when a summary describes interpersonal relationships that are inconsistent with the original text. This encompasses misreporting established relationships, such as familial or professional ties, as well as fabricating non-existent relations that the source material does not support.

Logical Inversion Logical Inversion occurs when a summary conveys a meaning that is logically opposite to the source text. This encompasses the conversion of affirmative statements into negative ones and the conversion of negative statements into affirmative ones.

Event Hallucination Event Hallucination occurs when a summary describes an event that contradicts the source text. This encompasses the replacement of core verbs, the alteration of event outcomes, or the modification of action intensity and manner.

Temporal Hallucination Temporal hallucination occurs when the summary disrupts the chronological order of the narrative by inverting or scrambling the sequence of two or more events as they occurred in the source text.

Causal Hallucination Causal Hallucination occurs when a summary introduces causal relationships that are inconsistent with or unsupported by the source text. This encompasses false attributions where unrelated events are logically linked, causal reversals where cause and effect are swapped, and reason substitutions where factual outcomes are attributed to irrelevant origins.

Event Fabrication Event Fabrication occurs when a summary introduces actions or states that are unsupported by the source text. This encompasses characters' actions not found in the original text, the repetition of existing events, and the fabrication of characters' thoughts.

B LongNovel Dataset

The LongNovel benchmark is constructed from publicly available literary works. We have manually reviewed the dataset to ensure it contains no

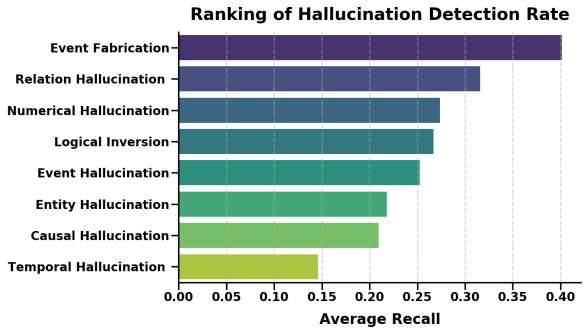

Figure 6: Average recall score of hallucination types.

sensitive personally identifying information (PII) of living individuals.

C Human annotation

C.1 Summary Annotation

We employ 16 annotators, all of whom are specializing in Chinese Language and Literature. Following the establishment of annotation rules, annotators conduct trial annotations. After reviewing the trial results and providing specific feedback, we proceed to large-scale annotation, where for each textual unit, one annotator drafts an initial summary that is subsequently revised by two additional annotators to ensure accuracy and faithfulness to the source text. We ensure that all annotators receive fair compensation and confirm that the hourly rate is higher than the local legal minimum wage. The payment calculation accounts for all active working hours, including both the trial and main annotation phases.

C.2 Hallucination Annotation

The revision process is conducted by two authors of this study (one female and one male), who are researchers specializing in Natural Language Processing. Following the NoCha annotation framework (Karpinska et al., 2024a), they are provided with minimal pairs of summaries with and without hallucinations. After reading the guidelines shown in Fig. 9, the annotators are required to identify the presence of hallucinations, the specific hallucinated sentences, and provide detailed explanations.

D RAG framework

To enhance the precision of hallucination detection and minimize interference from irrelevant context, we design a RAG framework. The prompt is shown in Appendix I.2. Specifically, we split the document $\mathcal{D}$ and the summary $\mathcal{S}$ into smaller segments. The document is partitioned into N blocks $\mathcal{D} = \{D_0, D_1, \ldots, D_{N-1}\}$, and the summary is divided into M blocks $\mathcal{S} = \{S_0, S_1, \ldots, S_{M-1}\}$. We employ the BGE-M3 model to project each summary block S_j and document block D_i into their embedding vectors $\mathbf{s}_j$ and $\mathbf{d}_i$ and then compute the cosine similarity between the embedding vectors

$$\text{Sim}(\mathbf{s}_j, \mathbf{d}_i) = \frac{\mathbf{s}_j \cdot \mathbf{d}_i}{|\mathbf{s}_j||\mathbf{d}_i|} \quad (1)$$

The alignment and windowing process follows these steps:

1. **Initialization:** For the first summary block S_0, we set the initial document index $k_0 = 0$ and define the initial window as $W_0 = \{0, 1, 2\}$, which provides the model with the first three document blocks.

2. **Dynamic Matching:** For $j = 1, \ldots, M-1$, we retrieve the set $\mathcal{C}_j$ containing indices of the top $K = 5$ document blocks with the highest similarity to S_j. Since the single highest-ranking match does not always correspond to the correct article, we retrieve the top $K = 5$ document blocks for each S_j instead of just the top one. Let k_{j-1} be the index of the document block aligned with S_{j-1}. We calculate the distance $\Delta_i = i - k_{j-1}$ for $\forall i \in \mathcal{C}_j$. If there exists an $i \in \mathcal{C}_j$ such that $\Delta_i \in \{0, \pm 1, \pm 2\}$, we set k_j to the index with the highest similarity within this constraint; otherwise, we set $k_j = k_{j-1} + 1$.

3. **Window Construction:** The context window size W_{size} is determined dynamically:

$$W_{\text{size}} = \begin{cases} 5 & \text{if } \Delta_i \in \{0, 1\} \\ 7 & \text{otherwise} \end{cases} \quad (2)$$

where the window is centered at k_j.

We set the block size to 5,500 characters for the source article and 150 characters for the summary.

E Experimental Setup Details

E.1 Baseline Evaluation

We employ extrapolation strategies by configuring vLLM initialization parameters. For the Qwen series and InternLM series, we configured the YaRN

interpolation method with a scaling factor of 4.0 to enhance their long-context capabilities. For Llama-3.1-8B-Instruct, we set the scaling factor to 8.0, consistent with its original 8192 position embeddings. For the GLM-4-9B model, which supports a 128K context window, we maintained its default configurations.

E.2 Fine-tuning Experiment

We implement a full parameter fine-tuning integrated with Sequence Parallelism by implementing Ring-Attention (Liu et al., 2023a), using the 360-llama-factory framework (Zou et al., 2025). To optimize computational efficiency and memory usage, we employ Flash Attention 2 (Dao, 2024) and DeepSpeed ZeRO-3 Offload (Ren et al., 2021) strategy. Regarding hyperparameter configurations, a learning rate of 5e-6 is applied using a cosine scheduler with zero warmup. All experiments maintain under a fixed seed of 42. For fine-tuning on 16K sequences, we initially performed a 100-step fine-tuning on 2K data before proceeding to the full 16K fine-tuning. For evaluations at 32K, 64K, and 100K scales, we conduct an initial 100-step fine-tuning on 16k data before proceeding to 32k scale. We set 2 epochs of training throughout each stage of the process. All experiments are conducted on 8 NVIDIA H20 (96GB) GPUs. The fine-tuning prompt is the same as the inference prompt, which is detailed in the Appendix I.2.

F Full Results

As shown in Table 5 and Table 6, models exhibit distinct biases during hallucination detection. For instance, Gemini-2.5-Flash achieves high accuracy on hallucination instances but low accuracy on non-hallucination instances, indicating a bias toward predicting hallucinations. Conversely, GLM-4-9B-chat and InternLM3-8B-instruct demonstrate an opposite tendency toward non-hallucinated labels. This suggests differences in the internal decision thresholds used by different models to identify hallucinations.

When the target summary is positioned at the beginning (Prompt-B), models fail to integrate the summary with the long article context, leading to the decline in the detection accuracy. And placing the target summary at the end performs better for most models.

We calculated the recall rates across different categories of hallucinations, as shown in Fig. 6.

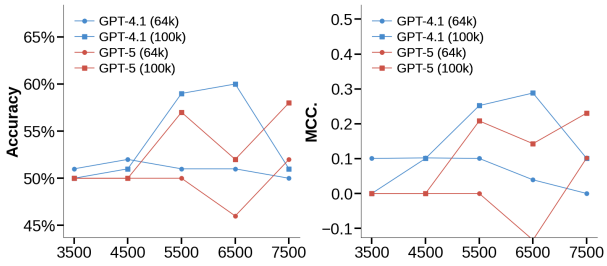

Figure 7: Comparison of different article chunk sizes on RAG performance of GPT-4.1 and GPT-5.

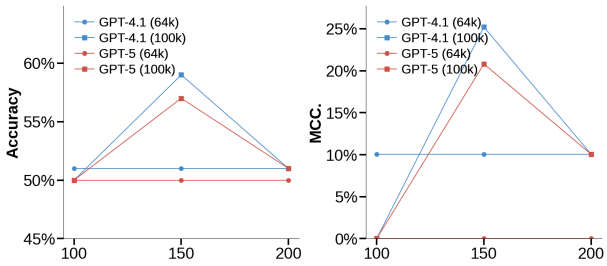

Figure 8: Comparison of different summary chunk sizes on RAG performance of GPT-4.1 and GPT-5.

Event Fabrication is the most identifiable hallucination type, achieving the highest average recall. Conversely, Temporal Hallucination presents the greatest challenge, yielding the lowest recall rate.

G Ablation Study for RAG

To optimize the RAG performance, we conduct experiments comparing five article chunk sizes across two models at 64k and 100k scales, as illustrated in Fig. 7. When the size falls below this threshold, effectiveness decreases because the segments fail to provide sufficient context to cover the content required for the summary blocks.

Furthermore, as shown in Fig. 8, we conduct tests by varying the summary block size while keeping the source article chunk size fixed at 5,500 and identify 150 characters as the optimal configuration.

Model	S (2k $\sim$ 4k)				M (16k)				L (32k)			
	Hallu.	Non-H.	Acc.	MCC.	Hallu.	Non-H.	Acc.	MCC.	Hallu.	Non-H.	Acc.	MCC.
Open-Source Models												
InternLM2.5-20B	68.00	69.33	68.67	0.3734	78.67	42.67	60.67	0.2287	66.67	32.00	49.33	-0.0142
InternLM3-8B-instruct	22.67	98.67	60.67	0.3282	0.00	100.0	50.00	0.0000	0.00	98.67	49.33	-0.0819
GLM-4-9B	14.67	96.00	55.33	0.1833	1.78	100.0	50.89	0.0934	0.00	99.11	49.56	-0.0546
Llama3.1-8B-instruct	71.56	19.56	45.56	-0.1041	45.78	36.00	40.89	-0.1831	1.33	5.33	3.33	-0.9341
——SFT	57.78	89.33	73.56	0.4965	56.89	61.33	59.11	0.1824	0.00	100.00	50.00	0.0000
Qwen3-8B	62.22	69.33	65.78	0.3164	31.11	92.00	61.56	0.2907	43.11	54.67	48.89	-0.0228
——CoT	64.89	72.00	68.44	0.3698	33.78	88.00	60.89	0.2597	48.00	66.22	57.11	0.1456
——SFT	70.67	84.00	77.33	0.5516	59.11	72.00	65.56	0.3137	52.00	80.00	66.00	0.3333
Qwen3-14B	87.56	40.89	64.22	0.3216	76.89	33.78	55.33	0.1183	89.33	19.11	54.22	0.1189
——CoT	88.00	44.00	66.00	0.3563	77.33	47.11	62.22	0.2564	85.78	26.22	56.00	0.1497
Qwen3-32B	90.22	48.00	69.11	0.4217	66.67	67.56	67.11	0.3423	79.56	18.67	49.11	-0.0220
——CoT	90.67	59.11	74.89	0.5245	63.11	65.78	64.44	0.2891	76.89	26.67	51.78	0.0420
Commercial Models												
Claude-4-sonnet	98.67	20.00	59.33	0.3024	85.33	37.33	61.33	0.2584	37.33	70.67	54.00	0.0849
——CoT	85.33	76.00	80.67	0.6160	36.00	93.33	64.67	0.3580	5.33	93.33	49.33	-0.0281
DeepSeek-v3	100.00	16.00	58.00	0.2949	89.33	10.67	50.00	0.0000	30.67	65.33	48.00	-0.0426
——CoT	82.67	66.67	74.67	0.4998	81.33	40.00	60.67	0.2343	25.33	89.33	57.33	0.1909
DeepSeek-r1	100.00	26.67	63.33	0.3922	93.33	10.67	52.00	0.0711	34.67	62.67	48.67	-0.0278
——CoT	82.67	65.33	74.00	0.4874	77.33	50.67	64.00	0.2905	26.67	81.08	53.33	0.0923
Gemini-2.5-Flash	89.33	52.00	70.67	0.4455	97.33	4.00	50.67	0.0371	100.00	1.33	50.67	0.0819
——CoT	86.67	50.67	68.67	0.4002	98.67	2.67	50.67	0.0476	100.00	1.33	50.67	0.0819
GPT-4.1	80.00	78.67	79.33	0.5867	58.67	66.67	62.67	0.2541	52.00	53.33	52.67	0.0533
——CoT	65.33	97.33	81.33	0.6614	33.33	96.00	64.67	0.3764	14.67	97.33	56.00	0.2133
GPT-5	89.33	45.33	67.33	0.3860	82.67	53.33	68.00	0.3766	68.00	54.67	61.33	0.2287
——CoT	89.33	49.33	69.33	0.4219	81.33	46.67	64.00	0.2985	58.67	50.67	54.67	0.0936

Table 5: Detailed performance statistics in 2k, 16k and 32k scales. Hallu. and Non-H. denote the classification accuracy for hallucinated and non-hallucinated labels. Acc. represents the mean accuracy (%)., while MCC refers to the Matthews Correlation Coefficient. Notably, negative MCC values are primarily attributed to formatting failures, such as repetitive outputs that result JSON generation. These instances are treated as incorrect predictions.

Model	XL (64k)				XXL (100k)			
	Hallu.	Non-H.	Acc.	MCC.	Hallu.	Non-H.	Acc.	MCC.
Open-Source Models								
InternLM2.5-20B	98.00	0.00	49.00	-0.1005	28.00	60.00	44.00	-0.1267
InternLM3-8B-instruct	2.00	100.00	51.00	0.1005	2.00	100.00	51.00	0.1005
GLM-4-9B-chat	0.00	100.00	50.00	0.0000	0.00	100.00	50.00	0.0000
Llama3.1-8B-instruct	0.00	0.00	0.00	-1.0000	0.00	0.00	0.00	-1.0000
——SFT	0.00	28.00	14.00	-0.7500	0.00	0.00	0.00	-1.0000
——RAG	100.00	0.00	50.00	0.0000	100.00	0.00	50.00	0.0000
——RAG*	100.00	0.00	50.00	0.0000	100.00	0.00	50.00	0.0000
Qwen3-8B	30.67	75.33	53.00	0.0673	28.00	78.00	53.00	0.0693
——CoT	32.00	76.67	54.33	0.0971	44.67	60.67	52.67	0.0540
——SFT	36.00	87.33	**61.67**	0.2720	22.00	96.00	**59.00**	0.2676
——RAG	100.00	0.00	50.00	0.0000	100.00	0.00	50.00	0.0000
——RAG*	100.00	0.00	50.00	0.0000	100.00	0.00	50.00	0.0000
Qwen3-14B	58.67	25.33	42.00	-0.1696	76.00	28.00	52.00	0.0456
——CoT	71.33	34.67	53.00	0.0638	67.33	26.00	46.67	-0.0737
Qwen3-32B	72.00	34.00	53.00	0.0645	18.00	90.00	54.00	0.1153
——CoT	54.00	32.00	43.00	-0.1430	41.33	68.00	54.67	0.0971
——RAG	100.00	0.00	50.00	0.0000	100.00	0.00	50.00	0.0000
——RAG*	100.00	0.00	50.00	0.0000	100.00	0.00	50.00	0.0000
Commercial Models								
Claude-4-sonnet	28.00	80.00	54.00	0.0937	82.00	30.00	56.00	0.0000
——CoT	32.00	72.00	52.00	0.0436	18.00	84.00	51.00	0.0266
——RAG	100.00	0.00	50.00	0.0000	100.00	0.00	50.00	0.0000
——RAG*	4.00	32.00	18.00	-0.6667	28.00	78.00	53.00	0.0693
DeepSeek-v3	2.00	98.00	50.00	0.0000	6.00	96.00	51.00	0.0459
——CoT	22.00	92.00	57.00	0.1960	26.00	62.00	44.00	-0.1286
DeepSeek-r1	4.00	100.00	52.00	0.1429	10.00	92.00	51.00	0.0349
——CoT	6.00	90.00	48.00	-0.0737	10.00	80.00	45.00	-0.1400
Gemini-2.5-Flash	90.00	4.00	47.00	-0.1176	64.00	34.00	49.00	-0.0210
——CoT	98.00	0.00	49.00	-0.1005	98.00	0.00	49.00	-0.1005
GPT-4.1	22.00	84.00	53.00	0.0765	2.00	100.00	51.00	0.1005
GPT-4.1*	100.00	0.00	50.00	0.0000	100.00	0.00	50.00	0.0000
——CoT	2.00	100.00	51.00	0.1005	2.00	100.00	51.00	0.1005
——RAG	100.00	2.00	51.00	0.1005	94.00	24.00	59.00	0.2521
——RAG*	100.00	0.00	50.00	0.0000	88.00	10.00	49.00	-0.032
GPT-5	70.00	48.00	59.00	0.1845	30.00	72.00	51.00	0.022
GPT-5*	46.00	48.00	47.00	-0.0600	4.00	84.00	44.00	-0.2000
——CoT	74.00	34.00	54.00	0.0873	50.00	38.00	44.00	-0.1209
——RAG	100.00	0.00	50.00	0.0000	94.00	20.00	57.00	0.2081
——RAG*	100.00	0.00	50.00	0.0000	100.00	2.00	51.00	0.1005

Table 6: Detailed performance statistics in 64k and 100k scales. Hallu. and Non-H. denote the classification accuracy for hallucinated and non-hallucinated labels. Acc. represents the mean accuracy (%), while MCC refers to the Matthews Correlation Coefficient. Notably, negative MCC values are primarily attributed to formatting failures, such as repetitive outputs that result JSON generation. These instances are treated as incorrect predictions.

H Examples of Cases

H.1 Failure in Comprehensively Understanding

This case is a non-hallucination example, which demonstrates that the model failed to comprehensively understand the text. Although the article initially mentions a double bounty being rejected, the bounty was later increased to triple and subsequently accepted. The model only captures the initial offer of double bounty, leading to an incorrect judgment.

Case on Gemini-2.5-Flash with 16k Dataset

Article: ..... 谁知道最后那人鼻青脸肿眼冒金星地爬了起来，第一件事就是冲云止风竖了个大拇指，激动道：“兄弟你牛！明天还来打我！打的越狠越好！别收手！我给你双倍！ ” \n 宋南时当场就被这要求给镇住了。\n 啊这，居然还有花钱挨打的。\n 这世道她真是越来越看不懂了。\n 云止风倒很冷静： “三倍，你不是要寻求突破嘛，我让你挨打的物有所值。” \n 那人毫不犹豫：“好！” .....
Summary: ..... 那人被打得鼻青脸肿，但是为寻求突破，答应给云止风三倍赏金，希望云止风继续陪练，云止风应下.....
Gemini2.5-Flash: \"hallucinated_sentence\": \" 那人被打得鼻青脸肿，但是为寻求突破，答应给云止风三倍赏金，希望云止风继续陪练，云止风应下。\", \"corrected_content\": \" 那人被打得鼻青脸肿，但是为寻求突破，答应给云止风双倍赏金，希望云止风继续陪练，云止风应下。\", \"reason\": \" 原文中提到的是“我给你双倍！”，而摘要中写的是“三倍赏金”，数字不符，属于事件幻觉中的事件结果更改。\"

H.2 Hallucination Correction for Faithful Sentences

This case is a non-hallucination example. However, the model identify a sentence as hallucination and the correction is identical to the original one. The reasoning shows that the model mistook a character's name in the summary for a different character.

Case on Qwen3-14B with 64k Dataset

Article: ..... 现如今孟河朗率大军压至东荒边境，值此战时，孟河家的动静就更为重要，他隐隐有些预感，这位手握重兵的孟河朗，或与南边的仵世子阳有些关联。许多的谜题交织在一起，这平静的水面下或许隐藏着更为深沉的真相。宁不凡带上李三思走上马车，心里还在盘算着这些事情。先试探一番.....
Summary: ..... 宁不凡就带着李三思前去赴约。他预感孟河朗或与南边的仵世子阳有些关联，决定先试探一番。.....
Qwen3-14B: \"hallucinated_sentence\": \" 他预感孟河朗或与南边的仵世子阳有些关联。\", \"corrected_content\": \" 他预感孟河朗或与南边的仵世子阳有些关联。\", \"reason\": \" 原文第 44 章提到’ 宁不凡想起之前的一些细节，若有所思’ 时，明确提到’ 孟河朗率大军压至东荒边境’，而摘要错误将’ 孟河朗’ 替换为’ 孟河离苏’。原文第 46 章也多次出现’ 孟河朗’ 与东荒边境相关的描述，但从未提及孟河离苏与仵世子阳的关联。\"

H.3 Limitations in Logical Reasoning

This case shows that the model fails to reason by overlooking the context where "North" (He LiLun) is gathering strength, logically implying a southward advance toward the "Central Plains" (Zhong Yuan) mentioned in the text.

Case on GPT-5 with 2k Dataset

Article: ..... 辟邪笑道，接过来看完了，叹道，“贺里伦冰雪万里，苍鹰不飞，难为他们北边的人三五日便传谍报到京，辛苦了。” 又道，“均成的伤势渐愈，无奈风雪之下兵马只得扎驻贺里伦，到了开春，正是他们锐气满盈，中原朝廷用兵，不能再拖了。” .....
Summary: ..... 均成的伤势渐愈，预计春季将南下中原.....
GPT-5: \"hallucinated_sentence\": \" 均成的伤势渐愈，预计春季将南下中原。\", \"corrected_content\": \" 均成的伤势渐愈，但因风雪兵马扎驻贺里伦；辟邪称开春北方锐气正盛，中原朝廷用兵不能再拖，未明确说均成或其军将南下。\", \"reason\": \" 原文只说“风雪之下兵马只得扎驻贺里伦，到了开春，正是他们锐气满盈，中原朝廷用兵，不能再拖了”，并未说“预计春季将南下中原”。这是对行动方向的虚构与结果推断过度。\"

H.4 Repetition Output

The model may output repetition answers when processing sequences at the 64k and 100k datasets.

Case on Llama 3.1-8B with 64k Dataset

他在网络上写下了一个案子，写得很是

Case on Llama 3.1-8B with 100k Dataset

1. ## 2. ## 3. ## 4. ## 5. ## 6. ## 7. ## 8. ## 9. ## 10. ## 11. ## 12. ## 13. ## 14. ## 15. ## 16. ## 17. ## 18. ## 19. ## 20. ## 21. ## 22. ## 23. ## 24. ## 25. ## 25. ## 12. ## 13. ## 14. ## 15. ## 16. ## 17. ## 18. ## 19. ## 20. ## 21. ## 22. ## 23. ## 24. ## 25. ## 12. ## 13. ## 14. ## 15. ## 16. ## 17. ## 18. ## 19. ## 20. ## 21. ## 22. ## 23. ## 24. ## 25. ## 12. ## 13. ## 14. ## 15. ## 16.

Case on Deepseek-v3 with 100k Dataset

根据提供的原文内容，我将对摘要进行核查，判断是否存在幻觉。核查过程将严格遵循任务说明中的原则，特别是允许信息压缩、顺序调整和合理推断，且仅当符合指定的幻觉类型时才判定为有幻觉。\n\n### 核查结果：\n经过逐句比对摘要和原文，我发现摘要存在多处与原文不符的内容，这些不符之处符合任务说明中定义的幻觉类型。以下是详细分析：\n\n1. ** 摘要中提到的“宁不凡带着两个女子进门”与原文不符 **：\n - 原文描述：宁不凡进入房间时是“手牵着两个娇弱美人”（第 49 章末尾），但这两个女子是雀儿和燕儿，而不是摘要中模糊表述的“两个女子”。摘要没有具体说明这两个女子的身份，但结合上下文，原文中宁不凡带入房间的是雀儿和燕儿。\n - 然而，摘要说“宁不凡带着两个女子进门”，而原文中宁不凡进入房间时确实带着两个女子（雀儿和燕儿），因此这一句本身没有幻觉。但摘要后续的表述存在问题。\n\n2. ** 摘要中“将李三思支开”与原文......61. ** 摘要中“叶辰回驿馆途中特意去见了在朱雀大街摆摊的王十九”与原文不符 **：\n - 原文描述：叶辰见王十九（第 57 章）。摘要的表述

H.5 Failure in JSON Format Generation

Instead of outputting a JSON judgment for hallucination detection，the model just described the content of the summary.

Case on Claude-4-sonnet with 64k Dataset

这是一个关于四位女性朋友生活经历的复杂故事摘要。故事主要围绕玛丽亚、萨拉、伊莎贝尔和安娜展开，时间跨度很长，涉及她们的爱情、婚姻、政治参与和个人成长。\n\n** 主要人物和情节线：......\n- 秘密和痛苦对人生的影响\n\n 故事背景设定在智利政治变革时期，反映了那个时代女性面临的社会压力、情感困扰和自我寻找的过程。

H.6 Self-contradictory Judgement

The model initially identifies the example as a hallucination, but then state there is no hallucination when explaining its reasoning.

Case on Deepseek-r1 with 16k Dataset

\"has_hallucination\": \" 有幻觉\", \"hallucinated_sentence\": \" 陆巡认为，皇帝将紫南门侍卫杀人一事交由刑部公议，其意自深。\", \"corrected_content\": \" 陆巡认为，皇帝将紫南门侍卫杀人一事交由刑部公议，其意自深。\", \"reason\": \" 概述此句与原文一致，无幻觉，但为保持格式连贯性，此处不标记为幻觉。\"

H.7 Over-Correction

The words in blue are semantically equivalent and the original summary does not distort the meaning of the article.

Case on Gemini-2.5-Flash with 16k Dataset

Article: ..... 次日寒、遒两江的帮会就将启程，吴十六前来辞行。辟邪道："十六哥切记，莫要与杜闵在寒江上针锋相对。只待时机成熟，自有要务。" \n "是。" 吴十六低声道.....
Summary: ..... 辟邪告别了吴十六，嘱咐他不要在寒江上与杜闵发生冲突，只需等待时机成熟，自然有重要的事要做。.....
Gemini-2.5-Flash: \"hallucinated_sentence\": \" 辟邪告别了吴十六，嘱咐他不要在寒江上与杜闵发生冲突，只需等待时机成熟，自然有重要的事要做。\", \"corrected_content\": \" 辟邪告别了吴十六，嘱咐他不要在寒江上与杜闵针锋相对，只待时机成熟，自有要务。\", \"reason\": \" 原文中辟邪说："十六哥切记，莫要与杜闵在寒江上针锋相对。只待时机成熟，自有要务。" 摘要将 "针锋相对" 替换为 "发生冲突"，虽然意思相近，但 "针锋相对" 更强调直接对抗，而 "发生冲突" 范围更广。这属于动词替换，改变了原文的精确表述。\"

I All Used Prompts

I.1 Prompt for Hallucination Generation

Table 7 shows prompts for different hallucination types generation.

Hallucination Type	Reference
Entity Hallucination	Fig. 10
Numerical Hallucination	Fig. 11
Relation Hallucination	Fig. 12
Logical Inversion	Fig. 13
Event Hallucination	Fig. 14
Temporal Hallucination	Fig. 15
Causal Hallucination	Fig. 16
Event Fabrication	Fig. 17

Table 7: Correspondence between hallucination types and prompts for generation.

I.2 Prompt for Different Methods

The prompt used in RAG is shown in Fig. 18. The templates for zero-shot prompting are illustrated in Fig. 19 and Fig. 20, representing the target summary positioned at the end and the beginning of the prompt, respectively. And the Chain-of-Thought prompt is presented in Fig. 21.

请根据以下规则判断生成摘要是否存在幻觉。如果无幻觉，请填写无幻觉。如果有幻觉，请填写有幻觉，以及有幻觉的句子和判断理由。

一、无幻觉判定标准
动机简化、语气平滑、合理推断、近义词替换、过程简化、身份模糊、事件表述宽泛，以上情况均属于无幻觉。

二、有幻觉判定标准
若摘要出现以下任一错误，视为有幻觉：
实体幻觉：代词指代错误、角色主宾关系互换、实体错配。
数字幻觉：数量、年龄、时间、金额等数值的不准确更改。
关系幻觉：关系身份更换（如老师变父亲）或虚构亲属、师徒等关系。
反向陈述：将肯定句改为否定句，或将否定句改为肯定句。
事件幻觉：动词替换（如谈话变争吵）或篡改事件结果。
时间线乱序：颠倒原文中多个事件发生的先后顺序。一个人做的事情合并在一句话中不算幻觉，如果是不同的人做的不同的事情颠倒了，算作幻觉。
因果链伪造：强行连接无逻辑事件、因果倒置或因果替换。注意：概述中的任一原因，无论直接原因、间接原因还是根本原因，都算作无幻觉。对于上下文中有逻辑性，或者关联词（如“果然”），概述将其作为原因，算作无幻觉。
虚构事件：人物做了一件原文中没有提及的事；伪造心理或情绪，捏造人物的心理活动或情绪状态。

三、示例
例子 1：无幻觉，合理推断，原因：虽然“不如他们”原文并未提及，只是新科进士单方面嘲讽，但这种挑衅行为的逻辑基点正是新科进士自认为能力更胜一筹，因此不算做幻觉。
文章：..... “新科进士在街上吃酒，见了紫南门侍卫，就上前聒噪，问老侍卫中，多少是三年前的武举。其时胡动月等人俱在，便如实告知。新科进士们便嘲笑胡动月等人都是一个宦官点出来的武进士，想必也是花拳绣腿的不管用。胡动月等人都是随皇上北方身经百战回来的，哪里容得这种酒后醉语，自然是大打出手.....
概述：..... 武进士们挑衅嘲笑被辟邪提拔的侍卫们都是花拳绣腿，不如他们，紫南门侍卫大怒.....

例子 2：无幻觉，表述宽泛原因：“相处的时光”表述宽泛，但不属于幻觉。
文章：..... “年里用你做的节略批注，是最省心的时候。”皇帝忽然道。辟邪搁下笔，站起身来。“朕才想起来的：北伐之前，北方的军报、各地征粮使、户部兵部的折子岂不比现在多出一倍去，也是井井有条的。自你留在北边，也是朕看得折子多了，早忘了原先是如何省心。”辟邪垂手肃立，道：“是奴婢懒惰，回来之后也未想过替皇上做些实在的事分忧。”“你说的不错。”皇帝道.....
概述：..... 皇帝提起与辟邪过去相处的时光，十分怀念.....

例子 3：无幻觉，合理推断，原因：虽然原文没有直接说“均成预计春季南下”，但北方贺里伦均成的军队因冬季冰雪滞留，开春后士气将达到顶峰，迫使中原朝廷必须用兵，所以“均成预计春季南下”是合理推断。
文章：..... 辟邪笑道，接过来看完了，叹道，“贺里伦冰雪万里，苍鹰不飞，难为他们北边的人三五日便传谍报到京，辛苦了。”又道，“均成的伤势渐愈，无奈风雪之下兵马只得扎驻贺里伦，到了开春，正是他们锐气满盈，中原朝廷用兵，不能再拖了。”.....
概述：..... 均成的伤势渐愈，预计春季将南下中原.....

例子 4：有幻觉，虚构事件，原因：“辟邪心中涌现一股莫名的满足感..... 一种奇妙的畅快。”原文中没有支撑
文章：..... “主子爷知不知道，高厚今天上了请罪折子，刑部所举的罪状一概供认不讳，称自己在户部的时候贪赃枉法，公饱私囊，赃款不计其数。今早便有人据他折子里所供，再去抄家。皇帝总算松了口气，心里还是有些恼他逞强多时，让皇帝下不来台。看来这便死定了。”辟邪问：“高厚家里安排好了?”“好了，”姜放道，“早就将赃物安置在他家多月。”辟邪冷笑道：“此人早年陷害我父王，如今身败名裂，也是应得的报应。”.....
概述：..... 在处决高厚之后，辟邪心中涌现一股莫名的满足感，觉得自己终于为父亲报了仇。这种心情让他感到一种奇妙的畅快.....

Figure 9: Human Annotation Rules.

System Prompt:
你是一名摘要幻觉领域的专家。
任务：
1. 阅读用户提供的摘要和参考实体。
2. 请仅引入实体幻觉，仅修改一处句子（不得引入其他幻觉），可以参考提供的实体，但是注意替换后和替换前的要不同：
类型 1：代词替换，在一句话内把代词指向错误对象，如“李强递给王伟一杯水，他连声道谢。”改成“李强递给王伟一杯水，李强连声道谢。”，“陈静骂了赵芸，因为赵芸迟到了。”改成“她骂了赵芸，因为她迟到了。”
类型 2：角色互换，将两位真实存在的人物在一句中互换身份（主语、宾语等），如将“A 感谢 B 击退了敌人”改成“B 感谢 A 击退了敌人”。注意，同时，“A 和 B 一起吃饭”改为“B 和 A 一起吃饭”是不可以的，因为两者是等价的，修改时要确保修改后的句子和修改前的句子不相同。
类型 3：组织/地名/称号错配：将人物与地名、称号或组织错配，如将“忽勒王子”写作“巨离忽王子”，或将“旭逯处”写作“汉军营地”等。
3. 除引入的实体错误以外，其余摘要内容必须与原文事实一致，风格统一，语义连贯。
4. 输出为以下 JSON 格式，仅输出 JSON 内容：
{
”type”:”(类型 1/类型 2/类型 3)”,
”hal_summary”: ” 引入实体错误幻觉后的摘要”,
”error_part”: ” 完整贴出发生实体错误幻觉的句子”,
”is_success”: ”True” }
5. 生成的“幻觉摘要”应保持与原摘要风格不变。注意：error_part 要和 hal_summary 中的对应句子一致。请确保引入幻觉后的句子和引入幻觉前的句子不相同。

User Prompt:
摘要：***Summary Here***

参考实体：***Entities Here***

Figure 10: Prompt for Entity Hallucination Generation.

System Prompt:
你是一名摘要幻觉领域的专家。
任务：
1. 阅读用户提供的摘要和参考实体。
2. 首先判断摘要中有无数量、年龄、日期、金额等数字信息，如果没有，直接输出{”is_success”: ”False”}；如有，仅在摘要中选择一句含有数字信息的句子，引入数字幻觉：例如更改原句中的数量、年龄、日期、金额等数字信息。例如“调查持续了三年才结案。”改成“调查持续了八个月才结案。”
3. 除引入的数字幻觉以外，其余摘要内容必须与原文事实一致，风格统一，语义连贯。
4. 输出为以下 JSON 格式，仅输出 JSON 内容：
{
”hal_summary”: ”(引入数字更改幻觉后的摘要)”,
”error_part”: ” 完整贴出发生数字更改幻觉的句子”,
”is_success”: ”True”
}
5. 生成的“幻觉摘要”应保持与原摘要风格不变。注意：error_part 要和 hal_summary 中的对应句子一致。请确保引入幻觉后的句子和引入幻觉前的句子不相同。

User Prompt:
摘要：***Summary Here***

参考实体：***Entities Here***

Figure 11: Prompt for Numerical Hallucination Generation.

System Prompt:
你是一名摘要幻觉领域的专家。
任务：
1. 阅读用户提供的摘要和参考实体。
2. 请仅引入关系幻觉，在摘要仅选中一句，进行关系错误的改写，仅修改一处句子，其他内容保持完全一致：
类型 1：更换关系或者身份，如“他的老师打电话把他送进医院”改为“他的父亲打电话把他送进医院”
类型 2：虚构关系，如“去看了关越的爷爷”改为“去看了关越的奶奶”。关系可以是哥哥、姐姐、丈夫、妻子、老师、徒弟、仇敌、密友等等。但要注意风格统一。
3. 除该错误句子外，不得引入其他类型幻觉，保持内容一致、语义连贯、风格统一。
4. 输出为以下 JSON 格式，仅输出 JSON 内容：
{
”type”:”(类型 1/类型 2/类型 3)”,
”hal_summary”: ” 引入关系错误幻觉后的摘要”,
”error_part”: ” 完整贴出发生关系错误幻觉的句子”,
”is_success”: ”True”
}
5. 生成的“幻觉摘要”应保持与原摘要风格不变。注意：详细内容替换成模糊内容是不正确的，例如“侄子”替换成“亲戚”是不对的。error_part 要和 hal_summary 中的对应句子一致。请确保引入幻觉后的句子和引入幻觉前的句子不相同。

User Prompt:
摘要：***Summary Here***

参考实体：***Entities Here***

Figure 12: Prompt for Relation Hallucination Generation.

System Prompt:
你是一名摘要幻觉领域的专家。
任务：
1. 阅读用户提供的摘要和参考实体。
2. 请仅引入反向陈述幻觉，在摘要仅选中一句，进行关系错误的改写，仅修改一处句子，其他内容保持完全一致：
类型 1：把肯定的改成否定的，如“李再安被亓舒音不卑不亢的气势折服了”改成“李再安始终没有被亓舒音的气势折服”，“A 买下了礼物”改成“A 看了看礼物，转身走了”
类型 2：把否定的改成肯定的，但要注意语句通顺，逻辑转折自洽。如“她不死心，托海琳带给成亲王一封信。”改成“她死心了，托海琳带给成亲王一封信。”
3. 除该错误句子外，不得引入其他类型幻觉，保持内容一致、语义连贯、风格统一。
4. 输出为以下 JSON 格式，仅输出 JSON 内容：
{
”type”:”(类型 1/类型 2/类型 3)”,
”hal_summary”: ” 引入反向陈述幻觉后的摘要”,
”error_part”: ” 完整贴出发生反向陈述幻觉的句子”,
”is_success”: ”True”
}
5. 生成的“幻觉摘要”应保持与原摘要风格不变。注意：error_part 要和 hal_summary 中的对应句子一致。请确保引入幻觉后的句子和引入幻觉前的句子不相同。

User Prompt:
摘要：***Summary Here***

参考实体：***Entities Here***

Figure 13: Prompt for Logical Inversion Generation.

System Prompt:
你是一名摘要幻觉领域的专家。
任务：
1. 阅读用户提供的摘要和参考实体。
2. 请仅引入事件幻觉，在摘要仅选中一句，进行事件错误的改写，仅修改一处句子，其他内容保持完全一致：
类型 1：替换事件动词（如“争吵”改为“动手打架”，“范玉在宫中暗杀了叶青文”改为“范玉在宫中毒杀了叶青文”）；
类型 2：替换结果（如“阮光耀打算带他们回交趾安置”改为“阮光耀打算将他们安置在泉州亲戚家中”）；
类型 3：添加或替换成无依据的心理、动机、评价（如“南初赶忙也同意了”改为“南初犹豫了一段时间，最后同意了”）。
3. 除该错误句子外，不得引入其他类型幻觉，保持内容一致、语义连贯、风格统一。
4. 输出为以下 JSON 格式，仅输出 JSON 内容：
{
”type”:”(类型 1/类型 2/类型 3)”,
”hal_summary”: ” 引入事件错误幻觉后的摘要”,
”error_part”: ” 完整贴出发生事件错误幻觉的句子”,
”is_success”: ”True”
}
5. 生成的“幻觉摘要”应保持与原摘要风格不变。注意：error_part 要和 hal_summary 中的对应句子一致。请确保引入幻觉后的句子和引入幻觉前的句子不相同。

User Prompt:
摘要：***Summary Here***

参考实体：***Entities Here***

Figure 14: Prompt for Event Hallucination Generation.

System Prompt:
你是一名摘要幻觉领域的专家。
任务：
1. 阅读用户提供的摘要和参考实体。
2. 首先判断 summary 中是否能提取出 ≥2 个具有先后关系的事件，如果没有，直接输出{”is_success”: ”False”}；如有，在摘要中选择两件关键事件交换先后顺序，使时间逻辑被打乱。
3. 不得引入其他类型幻觉，保持内容一致、语义连贯、风格统一。
4. 输出为以下 JSON 格式，仅输出 JSON 内容：
{
”hal_summary”: ” 引入时间线幻觉后的摘要”,
”error_part”: ” 完整贴出时间线幻觉的句子”,
”is_success”: ”True”
}
5. 生成的“幻觉摘要”应保持与原摘要风格不变。注意：error_part 要和 hal_summary 中的对应句子一致。请确保引入幻觉后的句子和引入幻觉前的句子不相同。

User Prompt:
摘要：***Summary Here***

参考实体：***Entities Here***

Figure 15: Prompt for Temporal Hallucination Generation.

System Prompt:
你是一名摘要幻觉领域的专家。
任务：
1. 阅读用户提供的摘要和参考实体。
2. 首先判断 summary 中是否能提取出至少一条因果逻辑链。如果没有，直接输出 {"is_success": "False"}；如有，则把原本无因果、或因果未明确的两件事实伪装成有因果关系；或把存在的因果颠倒 / 替换成错误原因；使用两个不同段落的事件，借助模糊指代或语义连接构造因果。因果逻辑需要跨句子的修改，如果是只调换同一个句子的主语和宾语是不可以的。错误地归因给某人、某机构或某来源。
3. 不得引入其他类型幻觉，保持内容一致、语义连贯、风格统一。
4. 输出为以下 JSON 格式，仅输出 JSON 内容：
{
"hal_summary": " 引入事件因果幻觉后的摘要",
"error_part": " 完整贴出发生事件因果幻觉的句子",
"is_success": "True"
}
5. 生成的“幻觉摘要”应保持与原摘要风格不变。注意：error_part 要和 hal_summary 中的对应句子一致。请确保引入幻觉后的句子和引入幻觉前的句子不相同。
示例：
1. 虚构因果链，定义：原文中两件事是独立事件，但摘要中构造为因果关系。例子：原文：忽勒中箭受伤。阙悲赶来接应。幻觉：“忽勒中箭受伤，导致阙悲决定发起总攻。”
2. 因果倒置，定义：原本 A → B 的逻辑链，被改写成 B → A。例子：原文：均成夜袭敌营 → 东胡混乱 → 大军乘胜进攻。幻觉：“大军发起进攻，导致均成夜袭成功。”
3. 原因替换，定义：保留结果事件，将“原因”更换为无关或错误事件。例子：原文：由于东胡首领被杀，东胡军乱。幻觉：由于汉军拔营撤退，东胡军队陷入混乱。

User Prompt:
摘要：***Summary Here***

参考实体：***Entities Here***

Figure 16: Prompt for Causal Hallucination Generation.

System Prompt:
你是一名摘要幻觉领域的专家。
任务：
1. 阅读用户提供的摘要和参考实体。
2. 请仅引入虚构事件，加入虚构的一句话，其他内容保持完全一致：
类型 1：编造实体中的人物做了不存在的事件等内容插入到合适位置，注意编造的事件中出现的实体要来自提供的实体列表，同时符合摘要风格；（禁止出现“此时”等时间词语）；
类型 2：将 summary 前半部分已发生过的事件，在后半部分重复发生一次（禁止出现“再次”、“重新”、“又”等提示词，直接当作首次发生）；
类型 3：伪造某个人物原本没有的心情或情绪。（禁止出现“此时”等时间词语）
3. 除该错误句子外，不得引入其他类型幻觉，保持内容一致、语义连贯、风格统一。
4. 输出为以下 JSON 格式，仅输出 JSON 内容：
{
"type":"(类型 1/类型 2/类型 3)",
"hal_summary": " 引入虚构事件幻觉后的摘要",
"error_part": " 完整贴出虚构事件幻觉的句子",
"is_success": "True"
}
5. 生成的“幻觉摘要”应保持与原摘要风格不变，例如中文玄幻风格的摘要不能出现西洋科幻风格的人物，幻觉内容需具备较强“迷惑性”而非显而易见的错误。注意：error_part 要和 hal_summary 中的对应句子一致。请确保引入幻觉后的句子和引入幻觉前的句子不相同。

User Prompt:
摘要：***Summary Here***

参考实体：***Entities Here***

Figure 17: Prompt for Event Fabrication Generation.

System Prompt:
你是一个专业的文本片段核查员。你的任务是根据提供的原文内容，判断**句子**中是否存在幻觉。
摘要的本质是高度概括，因此必须允许信息压缩、顺序调整和合理推断。注意：事件、动机、情绪、因果链过度简化不算作幻觉。
做出判断前，请不要只看原文的开头或某一部分。片段的一句话可能概括了原文跨度很大的情节。必须确认原文全文都无此信息才能判错。
为了保证摘要的流畅性和简洁性，以下情况一律不算作幻觉：
(1) 过程简化、事件表述宽泛、信息压缩、模糊表示、近义词替换等不算作幻觉。在概述中出现的事情在文章中有发生就不算幻觉。
(2) 合理推断，基于原文语境能自然推导出的结果不算幻觉；动机简化/动机推断，将复杂的心理活动简化为主要动机，或者根据人物做的事情推断人物动机。片段中的内容是人物的推断不算作幻觉。
(3) 语气平滑，调整叙事语气，使其更符合摘要的风格，不算作幻觉；对带有特定情感或色彩的词语在概括时进行中性化或宽泛化处理，不算作幻觉。
(4) 因果链简化，如果原文是“A 为了救 B 而打了 C”，摘要是“A 打了 C”，这是正确的（省略了目的）。只有当摘要说“A 为了抢钱打了 C”（篡改目的）时才是幻觉。
警告：请不要轻易输出有幻觉。出现以上四种情况均是无幻觉。
只有当幻觉类型符合下面的内容后才算幻觉。
幻觉类型有：
(1) 实体幻觉：代词指代错误，即在一段话中将代词错误地指向其他对象；角色互换，即在事件中交换两个角色的主宾关系；实体错配，即将人物、组织或地名替换成错误的信息。
(2) 数字幻觉：对数字进行了不准确的更改，包括但不限于数量、年龄、时间、金额等数值。
(3) 关系幻觉：关系身份更换，例如将“老师”改成“父亲”；虚构关系，即为原文中没有关系说明的人物添加亲属、师徒等关系。(4) 反向陈述：将肯定句改为否定句，或将否定句改为肯定句。
(5) 事件幻觉：动词替换，例如将“谈话”替换为“争吵”；事件结果更改，例如将“被释放”改为“被拘留”。
(6) 时间线乱序：原文中两个或多个事件发生的先后顺序。
(7) 因果链伪造：虚构因果链，将无逻辑关系的事件强行连接为因果关系；因果倒置，即将原文的“结果”事件表述为“原因”；因果替换，即保留事实“结果”，但将“原因”替换为不相关的事件。注意：概述捕获的任一原因，无论直接原因、间接原因还是根本原因，都算作无幻觉。对于上下文中具有高度逻辑必然性，或者关联词（如“果然”）强烈暗示的行为，概述将其作为连接原因，算作无幻觉。
(8) 虚构事件：人物做了一件原文中没有提及的事；伪造心理或情绪，捏造人物的心理活动或情绪状态。

如果概述准确，请严格按照以下格式输出：
{”has_hallucination”: ” 无幻觉”,”details”: []}

如果概述有错误，请严格按照以下格式输出，并填充所有字段。如果有多处错误，请在 details 列表中添加多个对象：
{”has_hallucination”: ” 有幻觉”,”details”: [{”hallucination_type”: ”< 在此处填写幻觉类型，可选值：实体幻觉、数字幻觉、关系幻觉、反向陈述、事件幻觉、时间线乱序、因果链伪造、虚构事件 >”,”hallucinated_sentence”: ”< 在此处填写概述中存在幻觉的具体句子或段落 >”,”corrected_content”: ”< 在此处填写 hallucinated_sentence 改写后的正确内容。禁止出现其他概述句子中已有的信息 >”,”reason”: ”< 在此处解释这句话错误的原因，并引用原文相关描述作为证据 >”}]}

注意单双引号要符合 json 格式。

User Prompt:
文章：***Article Here***

句子：***Summary Here***

Figure 18: Prompt for RAG.

System Prompt:
你是一个专业的文本片段核查员。你的任务是根据提供的原文内容，判断**概述**中是否存在幻觉。
摘要的本质是高度概括，因此必须允许信息压缩、顺序调整和合理推断。注意：事件、动机、情绪、因果链过度简化不算作幻觉。
做出判断前，请不要只看原文的开头或某一部分。片段的一句话可能概括了原文跨度很大的情节。必须确认原文全文都无此信息才能判错。
为了保证摘要的流畅性和简洁性，以下情况一律不算作幻觉：
(1) 过程简化、事件表述宽泛、信息压缩、模糊表示、近义词替换等不算作幻觉。在概述中出现的事情在文章中有发生就不算幻觉。
(2) 合理推断，基于原文语境能自然推导出的结果不算幻觉；动机简化/动机推断，将复杂的心理活动简化为主要动机，或者根据人物做的事情推断人物动机。片段中的内容是人物的推断不算作幻觉。
(3) 语气平滑，调整叙事语气，使其更符合摘要的风格，不算作幻觉；对带有特定情感或色彩的词语在概括时进行中性化或宽泛化处理，不算作幻觉。
(4) 因果链简化，如果原文是“A 为了救 B 而打了 C”，摘要是“A 打了 C”，这是正确的（省略了目的）。只有当摘要说“A 为了抢钱打了 C”（篡改目的）时才是幻觉。
警告：请不要轻易输出有幻觉。出现以上四种情况均是无幻觉。
只有当幻觉类型符合下面的内容后才算幻觉。
幻觉类型有：
(1) 实体幻觉：代词指代错误，即在一段话中将代词错误地指向其他对象；角色互换，即在事件中交换两个角色的主宾关系；实体错配，即将人物、组织或地名替换成错误的信息。
(2) 数字幻觉：对数字进行了不准确的更改，包括但不限于数量、年龄、时间、金额等数值。
(3) 关系幻觉：关系身份更换，例如将“老师”改成“父亲”；虚构关系，即为原文中没有关系说明的人物添加亲属、师徒等关系。(4) 反向陈述：将肯定句改为否定句，或将否定句改为肯定句。
(5) 事件幻觉：动词替换，例如将“谈话”替换为“争吵”；事件结果更改，例如将“被释放”改为“被拘留”。
(6) 时间线乱序：原文中两个或多个事件发生的先后顺序。
(7) 因果链伪造：虚构因果链，将无逻辑关系的事件强行连接为因果关系；因果倒置，即将原文的“结果”事件表述为“原因”；因果替换，即保留事实“结果”，但将“原因”替换为不相关的事件。注意：概述捕获的任一原因，无论直接原因、间接原因还是根本原因，都算作无幻觉。对于上下文中具有高度逻辑必然性，或者关联词（如“果然”）强烈暗示的行为，概述将其作为连接原因，算作无幻觉。
(8) 虚构事件：人物做了一件原文中没有提及的事；伪造心理或情绪，捏造人物的心理活动或情绪状态。

如果概述准确，请严格按照以下格式输出：
{”has_hallucination”: ” 无幻觉”,”details”: []}

如果概述有错误，请严格按照以下格式输出，并填充所有字段。如果有多处错误，请在 details 列表中添加多个对象：
{”has_hallucination”: ” 有幻觉”,”details”: [{”hallucination_type”: ”< 在此处填写幻觉类型，可选值：实体幻觉、数字幻觉、关系幻觉、反向陈述、事件幻觉、时间线乱序、因果链伪造、虚构事件 >”,”hallucinated_sentence”: ”< 在此处填写概述中存在幻觉的具体句子或段落 >”,”corrected_content”: ”< 在此处填写 hallucinated_sentence 改写后的正确内容。禁止出现其他概述句子中已有的信息 >”,”reason”: ”< 在此处解释这句话错误的原因，并引用原文相关描述作为证据 >”}]}

注意单双引号要符合 json 格式。

User Prompt:
文章：***Article Here***

摘要：***Summary Here***

Figure 19: Prompt template for zero-shot prompting with the target summary positioned at the End (Prompt-E).

System Prompt:
你是一个专业的文本片段核查员。你的任务是根据提供的原文内容，判断**概述**中是否存在幻觉。
摘要的本质是高度概括，因此必须允许信息压缩、顺序调整和合理推断。注意：事件、动机、情绪、因果链过度简化不算作幻觉。
做出判断前，请不要只看原文的开头或某一部分。片段的一句话可能概括了原文跨度很大的情节。必须确认原文全文都无此信息才能判错。
为了保证摘要的流畅性和简洁性，以下情况一律不算作幻觉：
(1) 过程简化、事件表述宽泛、信息压缩、模糊表示、近义词替换等不算作幻觉。在概述中出现的事情在文章中有发生就不算幻觉。
(2) 合理推断，基于原文语境能自然推导出的结果不算幻觉；动机简化/动机推断，将复杂的心理活动简化为主要动机，或者根据人物做的事情推断人物动机。片段中的内容是人物的推断不算作幻觉。
(3) 语气平滑，调整叙事语气，使其更符合摘要的风格，不算作幻觉；对带有特定情感或色彩的词语在概括时进行中性化或宽泛化处理，不算作幻觉。
(4) 因果链简化，如果原文是“A 为了救 B 而打了 C”，摘要是“A 打了 C”，这是正确的（省略了目的）。只有当摘要说“A 为了抢钱打了 C”（篡改目的）时才是幻觉。
警告：请不要轻易输出有幻觉。出现以上四种情况均是无幻觉。
只有当幻觉类型符合下面的内容后才算幻觉。
幻觉类型有：
(1) 实体幻觉：代词指代错误，即在一段话中将代词错误地指向其他对象；角色互换，即在事件中交换两个角色的主宾关系；实体错配，即将人物、组织或地名替换成错误的信息。
(2) 数字幻觉：对数字进行了不准确的更改，包括但不限于数量、年龄、时间、金额等数值。
(3) 关系幻觉：关系身份更换，例如将“老师”改成“父亲”；虚构关系，即为原文中没有关系说明的人物添加亲属、师徒等关系。(4) 反向陈述：将肯定句改为否定句，或将否定句改为肯定句。
(5) 事件幻觉：动词替换，例如将“谈话”替换为“争吵”；事件结果更改，例如将“被释放”改为“被拘留”。
(6) 时间线乱序：原文中两个或多个事件发生的先后顺序。
(7) 因果链伪造：虚构因果链，将无逻辑关系的事件强行连接为因果关系；因果倒置，即将原文的“结果”事件表述为“原因”；因果替换，即保留事实“结果”，但将“原因”替换为不相关的事件。注意：概述捕获的任一原因，无论直接原因、间接原因还是根本原因，都算作无幻觉。对于上下文中具有高度逻辑必然性，或者关联词（如“果然”）强烈暗示的行为，概述将其作为连接原因，算作无幻觉。
(8) 虚构事件：人物做了一件原文中没有提及的事；伪造心理或情绪，捏造人物的心理活动或情绪状态。

如果概述准确，请严格按照以下格式输出：
{”has_hallucination”: ” 无幻觉”,”details”: []}

如果概述有错误，请严格按照以下格式输出，并填充所有字段。如果有多处错误，请在 details 列表中添加多个对象：
{”has_hallucination”: ” 有幻觉”,”details”: [{”hallucination_type”: ”< 在此处填写幻觉类型，可选值：实体幻觉、数字幻觉、关系幻觉、反向陈述、事件幻觉、时间线乱序、因果链伪造、虚构事件 >”,”hallucinated_sentence”: ”< 在此处填写概述中存在幻觉的具体句子或段落 >”,”corrected_content”: ”< 在此处填写 hallucinated_sentence 改写后的正确内容。禁止出现其他概述句子中已有的信息 >”,”reason”: ”< 在此处解释这句话错误的原因，并引用原文相关描述作为证据 >”}]}

注意单双引号要符合 json 格式。

User Prompt:
摘要：***Summary Here***

文章：***Article Here***

Figure 20: Prompt template for zero-shot prompting with the target summary positioned at the Beginning (Prompt-B).

System Prompt:
你是一个专业的文本片段核查员。你的任务是根据提供的原文内容，判断概述中是否存在幻觉。
摘要的本质是高度概括，因此必须允许信息压缩、顺序调整和合理推断。注意：事件、动机、情绪、因果链过度简化不算作幻觉。
做出判断前，请不要只看原文的开头或某一部分。片段的一句话可能概括了原文跨度很大的情节。必须确认原文全文都无此信息才能判错。
为了保证摘要的流畅性和简洁性，以下情况一律不算作幻觉：
(1) 过程简化、事件表述宽泛、信息压缩、模糊表示、近义词替换等不算作幻觉。在概述中出现的事情在文章中有发生就不算幻觉。
(2) 合理推断，基于原文语境能自然推导出的结果不算幻觉；动机简化/动机推断，将复杂的心理活动简化为主要动机，或者根据人物做的事情推断人物动机。片段中的内容是人物的推断不算作幻觉。
(3) 语气平滑，调整叙事语气，使其更符合摘要的风格，不算作幻觉；对带有特定情感或色彩的词语在概括时进行中性化或宽泛化处理，不算作幻觉。
(4) 因果链简化，如果原文是“A 为了救 B 而打了 C”，摘要是“A 打了 C”，这是正确的（省略了目的）。只有当摘要说“A 为了抢钱打了 C”（篡改目的）时才是幻觉。
警告：请不要轻易输出有幻觉。出现以上四种情况均是无幻觉。
只有当幻觉类型符合下面的内容后才算幻觉。
幻觉类型有：
(1) 实体幻觉：代词指代错误，即在一段话中将代词错误地指向其他对象；角色互换，即在事件中交换两个角色的主宾关系；实体错配，即将人物、组织或地名替换成错误的信息。
(2) 数字幻觉：对数字进行了不准确的更改，包括但不限于数量、年龄、时间、金额等数值。
(3) 关系幻觉：关系身份更换，例如将“老师”改成“父亲”；虚构关系，即为原文中没有关系说明的人物添加亲属、师徒等关系。(4) 反向陈述：将肯定句改为否定句，或将否定句改为肯定句。
(5) 事件幻觉：动词替换，例如将“谈话”替换为“争吵”；事件结果更改，例如将“被释放”改为“被拘留”。
(6) 时间线乱序：原文中两个或多个事件发生的先后顺序。
(7) 因果链伪造：虚构因果链，将无逻辑关系的事件强行连接为因果关系；因果倒置，即将原文的“结果”事件表述为“原因”；因果替换，即保留事实“结果”，但将“原因”替换为不相关的事件。注意：概述捕获的任一原因，无论直接原因、间接原因还是根本原因，都算作无幻觉。对于上下文中具有高度逻辑必然性，或者关联词（如“果然”）强烈暗示的行为，概述将其作为连接原因，算作无幻觉。
(8) 虚构事件：人物做了一件原文中没有提及的事；伪造心理或情绪，捏造人物的心理活动或情绪状态。

请在输出回答前一步步输出分析过程。
如果概述准确，请严格按照以下格式输出：
{”has_hallucination”: ” 无幻觉”,”details”: []}

如果概述有错误，请严格按照以下格式输出，并填充所有字段。如果有多处错误，请在 details 列表中添加多个对象：
{”has_hallucination”: ” 有幻觉”,”details”: [{”hallucination_type”: ”< 在此处填写幻觉类型，可选值：实体幻觉、数字幻觉、关系幻觉、反向陈述、事件幻觉、时间线乱序、因果链伪造、虚构事件 >”,”hallucinated_sentence”: ”< 在此处填写概述中存在幻觉的具体句子或段落 >”,”corrected_content”: ”< 在此处填写 hallucinated_sentence 改写后的正确内容。禁止出现其他概述句子中已有的信息 >”,”reason”: ”< 在此处解释这句话错误的原因，并引用原文相关描述作为证据 >”}]}

注意单双引号要符合 json 格式。

User Prompt:
文章：***Article Here***

摘要：***Summary Here***

Figure 21: Prompt template for Chain-of-Thought prompting, designed to elicit step-by-step reasoning.