CNNSum: Exploring Long-Context Summarization with Large Language Models in Chinese Novels

Lingxiao Wei[†] **He Yan**[✠] **Xiangju Lu**[✠] **Junmin Zhu**[✠]
Jun Wang[†] **Wei Zhang**[†‡*]
[†]East China Normal University [✠]iQIYI Inc [‡]Shanghai Innovation Institute
`51265901053@stu.ecnu.edu.cn` `zhangwei.thu2011@gmail.com`

Abstract

Large language models (LLMs) have been well-researched in various long-context tasks. However, the scarcity of long-context summarization datasets hinders progress in this area. To address this, we introduce CNNSum, a multi-scale long-context summarization benchmark based on Chinese novels, featuring human-driven annotations across four subsets totaling 695 samples, with lengths ranging from 16k to 128k. We benchmark numerous LLMs and conduct detailed human assessments to summarize abnormal output types. Furthermore, we extensively explore how to improve long-context summarization. In our study: (1) Advanced LLMs may generate much subjective commentary, leading to vague summaries. (2) Currently, long-context summarization mainly relies on memory ability. The advantages of Large LLMs are hard to utilize, thus small LLMs are more cost-effective. (3) Different prompt types paired with various version models may cause large performance gaps. In further fine-tuning, these can be mitigated, and the Base version models perform better. (4) LLMs with RoPE-base scaled exhibit strong extrapolation potential; using short-context data can significantly improve long-context summarization performance. However, further applying other interpolation methods requires careful selection. (5) CNNSum provides more reliable evaluation results than other benchmarks. We release CNNSum to advance future research.[1]

1 Introduction

Long-context large language models are advancing rapidly (Wang et al., 2024; Pawar et al., 2024), making complex long-context tasks that were once considered difficult achieve revolutionary progress (Li et al., 2024a). However, long-context summarization tasks progress slowly, which is crucial for assessing long-context ability and forms the basis for complex contextual understanding (Pagnoni et al., 2021; Bhandari et al., 2020; Kim et al., 2024). (Pu et al., 2023) declare that traditional summarization research is dead, but creating new high-quality datasets remains essential. Unlike most tasks, summarization is more subjective, without a gold standard. And obtaining high-quality data is challenging for long texts, which requires a global context understanding and memory, and even human experts struggle to annotate directly (Qiu et al., 2024).

Recent studies on LLMs summarization (Tam et al., 2023; Song et al., 2024; Li et al., 2024b) are still based on previous short datasets (Narayan et al., 2018; Nallapati et al., 2016). (Bertsch et al., 2024) proposes unlimited-input transformers based on summarization task, using only BookSum (Kryściński et al., 2022) as sufficiently long data, which remains the latest book-length dataset. The maximum length of newer Chinese long summarization datasets (Liu et al., 2022) is only 6k. (Chang et al., 2024; Kim et al., 2024) study book-length summarization, but both are expensive in annotating their own datasets to avoid leakage risks.

Most studies focus on hallucination-related issues evaluations. However, implementing these evaluations holds limited significance if models cannot work properly. Although the 128k context length has been common, LLMs exhibit large performance degradation on various tasks as sequence length increases (Hsieh et al., 2024). In this study, we indeed find that outputs may be disorganized, meaningless, or fail to follow the instructions. The core is how LLMs handle long-context summarization tasks. Due to a lack of datasets, there is still insufficient systematic research and guidance.

Despite many long-context LLMs benchmarks emerging (An et al., 2024; Bai et al., 2024b; Kwan et al., 2024; Zhang et al., 2024; Ni et al., 2024b; Qiu et al., 2024), their summarization tasks still have considerable room for improvement in terms of construction. Regardless of language and domain,

*Corresponding author.
[1] `https://github.com/CxsGhost/CNNSum`

Findings of the Association for Computational Linguistics: ACL 2025, pages 8034–8062
July 27 - August 1, 2025

we summarize the shortcomings: (1) Building on old datasets (Fabbri et al., 2019; Huang et al., 2021; Sharma et al., 2019; Kryściński et al., 2022; Wang et al., 2021) leads to high leakage risks (Xu et al., 2024; Li et al., 2024c). (2) Data volume is small, maybe just dozens. (3) The average and maximum lengths are short, typically under 16k. (4) Lacking multi-scale length-based subsets limits evaluation in varying contexts, and our experiments indicate a wide range of mixed lengths may lead to misleading results. (5) Annotations are either collected from the web, carrying high leakage risk and low quality, or synthesized by LLMs and may have various errors (Chang et al., 2024; Song et al., 2024; Kim et al., 2024). Each benchmark mentioned above has at least two of these shortcomings.

To better explore long-context summarization with LLMs, we introduce Chinese Novel Summarization (CNNSum), a new multi-scale dataset that improves all the above shortcomings. Based on a newly collected corpus, we built four subsets: L(16k), XL(32k), 2XL(64k), and 3XL(128k), totaling 695 samples. The annotations are completed by human experts with assistance from LLMs. Since prompt formats greatly impact LLMs (Sclar et al., 2024) and beginning and ending tokens are crucial in long contexts (Han et al., 2024; Liu et al., 2024b), we define two prompt types for baselines: Instruction at the Beginning (Prompt-IB) and Instruction at the End (Prompt-IE), as shown in Appendix I.1. Although ROUGE-L (Lin, 2004) remains the primary metric in most benchmarks, it differs significantly from human preferences. Even LLMs are not yet reliable enough to replace humans (Shen et al., 2023). While referencing automatic metrics, we conduct fine-grained inspection and analysis of the outputs to categorize error types. Our observations are as follows: (1) GPT-4o (OpenAI, 2024b) prefers more subjective narration leading to vague summaries, so focusing on objective plots remains the key. (2) The two prompts can result in a significant output quality gap across various LLMs. (3) The Chat or Instruction version may harm the Base model's long-context summarization ability. (4) The large LLMs struggle to fully leverage their advantages in reasoning and comprehension, while small LLMs are more cost-effective.

The abnormal behaviors are usually related to training configurations of LLMs; we further conduct extensive fine-tuning experiments. The LLMs may have acquired longer-context abilities during pre-training (Xiong et al., 2024; Fu et al., 2024), requiring only activation. Consequently, we concatenate short summarization data into longer ones for training. Our findings indicate that: (1) The performance gaps caused by prompts can be mitigated, allowing models to fully perform. (2) The Base models are better for further fine-tuning and extrapolation than the Instruction or Chat version. (3) Fine-tuning with short-context data can significantly improve long-context summarization performance. (4) Current LLMs widely use Adjusted Base Frequency (ABF) (Xiong et al., 2024), improving the extrapolation potential and maintaining decent extrapolation performance on samples several times longer than the training length. However, when further combined with other interpolation methods, they show interesting differences. (5) CNNSum could provide more reliable evaluations, as it employs a more rigorous and reasonable length sampling strategy. The only relatively new and large-scale Chinese summarization benchmark is CLongEval-LStSum (Qiu et al., 2024), which reuses BookSum (Kryściński et al., 2022) and annotations are synthesized by GPT-4. However, we find it may produce misleading results compared to CNNSum, especially in extrapolation evaluation.

In summary, we introduce CNNSum, an entirely new long-context Chinese novel summarization benchmark, which is significantly superior in design and construction compared to existing datasets. We carry out comprehensive explorations of long-context summarization tasks for LLMs, providing guiding insights for future research in this field.

2 Related Work

2.1 Summarization Evaluation of LLMs

ROUGE (Lin, 2004) is still popular due to its simple implementation. It measures the information overlap between outputs and references. (Chang et al., 2024) proposes a protocol for coherence evaluation of book-length summaries generated by LLMs and achieves an LLM-based automatic implementation. It does not focus on faithfulness and thus does not rely on gold summaries but advanced LLMs like GPT-4 (Achiam et al., 2023). (Tam et al., 2023) constructs a benchmark for evaluating factual consistency by manually re-annotating previous news summarization datasets (Nallapati et al., 2016). (Krishna et al., 2023) proposes guidelines for human evaluation of faithfulness. (Song et al., 2024) introduces a prompt-guided fine-grained evaluator for faithfulness and completeness. It also relies on advanced LLMs for reliable evaluation. (Ravaut et al., 2024) studies how LLMs utilize entire context in summarization, still based on those previous datasets. (Kim et al., 2024) is the first to conduct a large-scale evaluation of book-length summarization faithfulness, which mainly focuses on commercial models. (Haldar and Hockenmaier, 2024) analyzes code summarization performance of LLMs based on short-context datasets.

2.2 Long-Context Extension for LLMs

RoPE (Su et al., 2024) is widely used in LLMs, but its extrapolation performance is limited. Position Interpolation (Fei et al., 2024) achieves efficient extension but still needs fine-tuning on the target length. NTK methods (u/LocalLLaMA, 2023; r/LocalLLaMA, 2024) mitigate the high-frequency information loss caused by interpolation, enabling extrapolation directly. YaRN (Peng et al., 2024) combines the above methods and attains 2x extrapolation. CLEX (Chen et al., 2024a) achieves more efficient dynamic scaling, and LongRoPE (Ding et al., 2024) leverages RoPE's non-uniformity to search better interpolations for more times of extrapolation, but the iteration process is complex. The above methods are mostly implemented on early LLMs (Touvron et al., 2023a,b) for experiments. The Adjusted Base Frequency (ABF) (Xiong et al., 2024) has been widely used (Yang et al., 2024; Qwen et al., 2024; Young et al., 2024; Cai et al., 2024), significantly improving extrapolation potential. (Liu et al., 2024c) further proposes scaling laws for RoPE base. There is still no systematic research on applying other interpolation methods to ABF-scaled LLMs.

3 CNNSum Construction

3.1 Chinese Book Corpus Collection

We collect 103 books from the Chinese Internet open-source data, each having a clear chapter structure. We remove books that are composed of multiple independent short stories or that lack main storylines, as their long-context dependencies are weak. Qwen2-72B-Instruct (Yang et al., 2024) is used to detect potentially popular books, as it has a high leakage risk (Ni et al., 2024a; Xu et al., 2024). We randomly select chapters from each book to generate summaries. If the model identifies source books and outputs extra information, we remove them as they may also be extensively leaked in other LLMs; this filters out 27 books. Our corpus has many web-serialized novels, often exhibiting unique writing habits or styles of authors. Common cases include non-standard punctuation usage and randomly interspersed irrelevant extras, as shown in Appendix H.1. These may disrupt the context coherence or lead to unexpected behavior in model outputs. We format these via regular expressions and manual checks.

3.2 Multi-Scale Sampling

We calculate the efficiency of various tokenizers in tokenizing Chinese text based on our corpus; details are in Appendix C.1. We use Yi (Young et al., 2024) tokenizer to process chapters and define four token-level target lengths T and their ranges. Each range has a specific lower boundary to avoid scoring bias from overly short data. The upper boundary is uniformly set as $T + 2k$ ($k = 1024$), able to evaluate extrapolation without causing score collapse. The specific settings are as follows:

$$
\begin{aligned}
T_L &= 16k, & Range_L &= [16k - 4k, 16k + 2k] \\
T_{XL} &= 32k, & Range_{XL} &= [32k - 6k, 32k + 2k] \\
T_{2XL} &= 64k, & Range_{2XL} &= [64k - 10k, 64k + 2k] \\
T_{3XL} &= 128k, & Range_{3XL} &= [128k - 16k, 128k + 2k]
\end{aligned}
$$

The book B is composed of multiple chapters c, represented as $B = \{c_0, c_1, \ldots, c_i\}$. For a target range R and a book B, the sampling method is as follows: (1) Initialize a variable sliding window w starting from c_0, expanding chapter by chapter. (2) When $w = [c_j, c_{j+1}, \ldots, c_{k-1}, c_k]$ $(j \le k < i)$:

i. If the length of $w < \inf R$, continue expanding w with chapter c_{k+1}.

ii. If the length of $w \in R$, gather as a data point, set $w = [c_{k+1}]$.

iii. If the length of $w > \sup R$, remove chapters sequentially from c_j onward until i or ii.

(3) Follow the procedure until B has no remaining chapters, then process the next book. This method is applied to each range R across our book corpus, yielding four candidate sets.

Despite candidate sets being large, the sample volume gaps per book vary considerably. For example, books with longer total lengths but shorter chapters have more samples. Conversely, shorter books may have no or few samples in 3XL candidate set. We prioritize data from books with fewer samples to preserve the diversity. Regarding length, we prefer data closer to the target T. Table 1 shows

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

Model Series	CNNSum-L (Count=190) Source Books=76				CNNSum-XL (Count=195) Source Books=71				CNNSum-2XL (Count=200) Source Books=60				CNNSum-3XL (Count=110) Source Books=45			
	Min	Q1	Mean	Max	Min	Q1	Mean	Max	Min	Q1	Mean	Max	Min	Q1	Mean	Max
Yi / Yi-1.5	13,447	14,841	15,721	18,323	27,586	30,932	31,844	34,703	56,323	63,683	64,303	67,403	115,351	128,673	128,751	132,688
Llama3 / Llama3.1	15,128	18,203	19,097	23,408	31,795	37,344	38,557	42,659	67,952	76,326	77,845	85,690	140,920	152,603	155,339	166,504
Qwen1.5 / Qwen2.5	12,938	14,529	15,324	18,501	27,130	30,189	31,026	34,104	54,760	61,718	62,526	66,844	110,373	123,792	124,995	130,707
GPT-4o-2024-08-06	15,564	17,974	18,814	22,792	32,428	36,676	37,984	42,177	68,641	74,996	76,598	85,231	138,417	150,226	152,875	164,939

Table 1: Token sequence length of CNNsum. Note that GPT-4o uses tiktoken[2]and Q1 denotes quarter quartile.

the final statistics of CNNSum. Each subset includes shorter data to maintain moderate length diversity, but the mean length is close to the target T, with the first quartile also kept near it. The final data volume is 695, based on two considerations: (1) The Annotation cost of long-context summaries. (2) Due to sample volume gaps across books, increasing the final volume will lead to imbalance, further introducing new biases from book content.

The key to long-context data is long dependencies within content, not just long sequence length. We assess using ProLong (Chen et al., 2024b) and the results indicate that CNNSum performs strong long dependencies compared to other summarization datasets. Details are provided in Appendix A.

3.3 Summary Annotation

Our approach is similar to **Incremental Updating** (Chang et al., 2024), but with manual efforts. For a sample, we first generate a plot synopsis for each chapter using the prompt in Appendix I.3, mainly consisting of key plots and character information. To reduce bias from a single model, we use various commercial LLMs as listed in Appendix B. Annotators then read each synopsis, select key plots by their judgment, and merge them into a final summary. If finding conflicts, annotators will ask LLMs to locate the source plots and correct them manually. We have two requirements for annotators: (1) Do not merely delete the unnecessary parts and merge the rest, but rewrite in your own words. (2) Avoid subjective commentary, focusing instead on objective plots. Although this may sacrifice some coherence, it increases the effective information density. For L and XL, we limit a maximum of 500 words, and 600 for 2XL and 3XL. Adding more words for coherence is not cost-effective, as most models do not produce overly long summaries.

Our annotation team has 23 experts. To reduce individual bias, each sample in 2~3XL is annotated by one expert and then reviewed by another for consistency. This annotation strategy leverages the strengths of LLMs in efficiently processing multiple texts concurrently, alongside the logical thinking and error-correction abilities of human experts. This combination complements the weaknesses of each other, making it one of the best annotation strategies for long-context summarization tasks.

4 Experiments

4.1 Baselines

Chinese has one of the largest corpora, but many advanced LLMs still support it poorly, as shown in Appendix C.1. Such as Claude3.5 (Anthropic, 2024) and GPT-4o cannot process 3XL. We select models as follows; details are in Appendix C.2.

Commercial Models (1) OpenAI flagship, GPT-4o-2024-08-06 (OpenAI, 2024b) and the lite version GPT-4o-mini-2024-07-18 (OpenAI, 2024a). (2) Moonshot-v1-128k (MoonshotAI, 2024). (3) Qwen-plus-2024-09-19 (Cloud, 2024). (4) Doubao-pro-128k (Doubao, 2024). All the above have a 128k context length. (5) Gemini-1.5-pro (Team et al., 2024a), with 2 million context length.

Open-Source Models (1) Yi (Young et al., 2024) and Yi-1.5 series. (2) Qwen1.5, Qwen2 (Yang et al., 2024), and Qwen2.5 (Qwen et al., 2024) series. (3) ChatGLM3-6B-128K (Du et al., 2022) and special version GLM4-9B-Chat-1M (GLM et al., 2024). (4) InternLM2.5-20B (Cai et al., 2024) series and special version InternLM2.5-7B-Chat-1M. (5) Llama3.1 series (Dubey et al., 2024). (6) Ministral-8B-Instruct-2410 (AI, 2023). (7) LWM-Text-1M (Liu et al., 2025).

4.2 Experimental Setup

We set generation tokens to 400 for L and XL, and 500 for 2~3XL. We use jieba[3] for Chinese segmentation and then report ROUGE-L (Lin, 2004). For the BERTScore (Zhang et al., 2020), we use the Chinese XLNet-Large (Cui et al., 2020). This architecture can handle long texts, and the computation is based on the hidden states of the last layer.

Baseline Evaluation We use vLLM (Kwon et al., 2023) for open-source models and apply greedy

[3]https://github.com/fxsjy/jieba

Model	Ctx. Len.	CNNSum L (16k)	XL (32k)	2XL (64k)	3XL (128k)	MSE (*P-IE vs P-IB*)
Commercial Models						
GPT-4o-2024-08-06	128k	15.5	14.2	12.5	-	0.0
GPT-4o-mini-2024-07-18	128k	15.7	13.7	12.8	-	0.3
Gemini-1.5-pro	2m	19.3	18.1	16.8	14.6	0.1
Qwen-plus-2024-09-19	128k	20.5	18.5	16.4	14.8	0.1
Doubao-pro-128k	128k	19.6	17.8	15.5	13.2	0.6
Moonshot-v1-128k	128k	**22.4**	**20.3**	**18.0**	**15.2**	0.9
Open-Source Models						
Yi-6B	4k	8.8	4.4	2.7	2.3	1.1
Yi-6B-Chat	4k	13.5	6.9	2.1	2.1	0.4
Yi-6B-200K	200k	9.9	9.4	8.8	4.0	4.5
Yi-34B-200K	200k	12.1	11.3	10.8	10.0	0.1
Yi-1.5-34B-32K	32k	11.6	10.5	9.6	0.1	14.5
Yi-1.5-34B-Chat-16K	16k	13.8	12.3	10.9	0	7.8
InternLM2.5-7B-Chat-1M	1m	18.0	17.1	14.7	13.0	0.9
InternLM2.5-20B	256k	18.2	16.3	14.4	12.7	0.1
InternLM2.5-20B-Chat	32k	18.9	17.3	16.2	14.2	0.1
ChatGLM3-6B-128K	128k	17.2	16.1	14.8	13.7	0.1
GLM4-9B-Chat-1M	1m	19.0	**17.9**	**16.5**	**15.4**	0.2
Llama3.1-8B	128k	7.8	8.0	8.4	3.1	7.2
Llama3.1-8B-Instruct	128k	15.6	14.3	12.8	9.9	1.4
Llama3.1-70B	128k	12.2	8.9	6.9	7.5	5.0
Llama3.1-70B-Instruct	128k	17.9	15.7	13.8	10.6	0.5
LWM-Text-1M	1m	3.3	3.0	2.5	1.1	0.2
Ministral-8B-Instruct-2410	128k	16.2	14.0	11.3	3.5	0.1
Qwen1.5-7B	32k	12.1	10.9	5.7	2.6	34.4
Qwen1.5-7B-Chat	32k	15.1	13.6	10.3	2.7	0.3
Qwen1.5-32B-Chat	32k	15.7	14.2	5.0	4.2	0.3
Qwen2-7B	128k	9.8	8.4	8.0	6.9	1.5
Qwen2-7B-Instruct	32k	15.2	13.3	12.3	11.3	1.5
Qwen2-72B	128k	14.8	13.2	11.7	10.4	16.3
Qwen2-72B-Instruct	32k	15.5	13.2	11.5	9.2	19.0
Qwen2.5-7B-Instruct	32k	15.7	14.0	12.7	9.6	3.4
Qwen2.5-72B-Instruct	32k	**19.6**	17.6	13.6	13.4	1.3

Table 2: Average ROUGE-L score of CNNSum. **Bold** denotes the best score, underline for the second-best, and wavy underline for the third-best. Red highlights the large value (≥ 5.0) in **MSE (*P-IE vs P-IB*)**.

sampling. We set $temperature = 0$ for commercial APIs. The details can be seen in Appendix D.1. The prompts we used are shown in Appendix I.1.

Fine-tuning Data Our dataset contains about 9,000 novel summarization samples, independent of the CNNSum corpus. The length ranges from 2k~4k with Yi tokenizer. The prompts are shown in Appendix I.2. We randomly concatenated the samples, similar to Section 3.2, setting ranges to 14k~18k and 30k~34k, with average lengths of about 16k and 32k. The main motivation is to enable models to adapt to more positions and activate extrapolation ability. (Tian et al., 2024) contemporaneously proposes a similar data augmentation strategy, proving it enhances long-context ability.

Fine-tuning Experiments We follow findings from (Chen et al., 2024c), use LoRA (rank=8) (Hu et al., 2021), and unfreeze embedding and normalization layers. Each experiment is repeated at least three times. More details are in Appendix D.2.

4.3 Main Results on CNNSum

We calculate mean ROUGE and use Mean Squared Error (MSE) to measure performance gaps caused by prompts, as shown in Table 2 (Full results are shown in Appendix F). Overall, commercial models outperform most open-source models. The low MSE demonstrates a more stable long-context summarization ability. Moonshot-v1 performs the best, but it suffers a decrease of 32.1% at 3XL. Gemini-pro, Qwen-plus, and Doubao-pro have similar performance. Doubao-pro shows the largest decrease of 32.7% from L to 3XL, while Gemini-pro shows the smallest decrease of 24.4%. Unexpectedly, GPT-4o performs the worst, lagging behind other

Model	Ctx. Len.	L ($16k$)			XL ($32k$)			2XL ($64k$)			3XL ($128k$)		
		CNNSum											
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
GPT-4o-2024-08-06	$128k$	61.2	59.4	60.3	60.9	59.3	60.1	60.1	56.8	58.0	-	-	-
GPT-4o-mini-2024-07-18	$128k$	61.8	61.9	61.8	60.8	60.0	60.4	60.6	59.3	59.9	-	-	-
Gemini-1.5-pro	$2m$	66.6	63.9	65.1	66.6	64.4	65.4	65.7	65.2	65.4	65.0	65.3	**65.1**
Qwen-plus-2024-09-19	$128k$	68.8	63.1	65.7	67.9	61.6	64.5	66.6	60.8	63.5	65.3	60.3	62.6
Doubao-pro-128k	$128k$	67.3	63.1	65.0	66.1	60.6	63.1	64.7	58.0	61.1	61.5	55.6	58.3
Moonshot-v1-128k	$128k$	67.4	67.8	**67.6**	67.0	67.4	**67.1**	65.8	67.3	**66.5**	64.1	64.8	64.4

Table 3: Average BERTScore of commercial models. **Bold** denotes the best score, underline for the second-best, and wavy underline for the third-best. **P** stands for precision, and **R** stands for recall.

Model	Ctx. Len.	L	XL	2XL	3XL
		CNNSum			
Yi-1.5-34B-32K	$32k$	**14.4**	**12.7**	**10.8**	0.1
		8.8	8.2	8.3	0.1
Yi-1.5-34B-Chat-16K	$16k$	**15.4**	**14.2**	**12.1**	0.0
		12.1	10.4	9.7	0.0
Llama3.1-8B	$128k$	6.1	6.6	7.0	2.2
		9.4	**9.4**	**9.7**	**3.9**
Llama3.1-70B	$128k$	11.8	7.5	5.2	7.0
		12.5	**10.2**	**8.6**	**7.9**
Qwen1.5-7B	$32k$	**16.5**	**14.5**	**7.0**	2.5
		7.7	7.2	4.4	**2.6**
Qwen2-72B	$128k$	12.1	10.7	10.3	9.5
		17.5	**15.7**	**13.1**	**11.3**
Qwen2-72B-Instruct	$32k$	12.4	10.5	10.6	**10.6**
		18.5	**15.8**	**12.3**	7.8

Table 4: Gray for Prompt-IE and white for Prompt-IB. **Bold** denotes the best score of each model.

models by 20%~30% on average from L to 2XL. Given that the output quality of commercial APIs is better, and their ROUGE scores are close, we specifically compute the BERTScore for them, as shown in Table 3. However, they almost have the same trend, providing limited additional meaningful insights. This indicates the limitations of simple automatic metrics. Therefore, in the subsequent parts, we only use ROUGE as the basic guiding metric and conduct detailed human inspection.

GLM4-9B-Chat-1M performs best in the open-source models, with only an 18.9% decrease from L to 3XL. InternLM2.5-20B-Chat shows the most robust extrapolation ability to 3XL, with a 24.8% decrease. In the Qwen series, Qwen2.5-72B-Instruct performs the best but decreases by 31.6% when extrapolated to 3XL. Others perform similarly on L and XL, with the Qwen2 series decreasing 20%~40% to 3XL. Due to lower Chinese encoding efficiency, Llama3.1 and Ministral require extrapolation on 3XL, with Llama3.1 series showing an average decrease exceeding 40% and Ministral reaching as high as 78.3%. Yi-34B-200K generally outputs just half or one sentence or meaningless repetitions. And Yi-6B-200K performs even worse.

4.4 Analysis and Case Study

Why Do GPT-4o Fail? As shown in Table 1 and Table 3, GPT-4o actually needs to process longer sequences. Despite this, its performance on L still trails other commercial models on XL on average by 16%. GPT-4o and Qwen-plus generate more concise summaries, while Moonshot-v1 and Gemini-pro generate longer ones. Moonshot-v1 outputs significantly longer than others, with many samples reaching the length limit without repetition. Longer outputs contain richer information, are more likely to overlap with references, leading to higher scores. For GPT-4o, its already shorter outputs often include subjective commentary. These contents are extensions of the plot rather than itself, usually not overlapping with references, as shown in Appendix H.2. On the other hand, such highly generalized content may indicate that the model has limited memory of key plots. Other commercial models rarely exhibit this and focus on objective plots, regardless of the output length.

Which Prompt? Which Model Version? Results of models with large MSE (in red) are detailed in Table 4. Models with shorter context lengths perform better with Prompt-IE, namely Qwen1.5 and Yi-1.5. We inspect their outputs and find that the instruction may be forgotten after reading long contexts in Prompt-IB. Many outputs copy from novels or are meaningless repetitions. Conversely, using Prompt-IE can summarize normally even in extrapolation, examples are in Appendix H.3. The Base model is more stable than its Chat version with Prompt-IE, which is not directly reflected by ROUGE. Yi-1.5-34B generates more concise and clean summaries, while the Chat model outputs more content, but it is more prone to disorganization and repetition. This observation also occurs in models with lower MSE, such as InternLM2.5-20B series. Examples are shown in Appendix H.4.

Models with longer context lengths seem to recognize the instruction at the beginning easily. However, it still may forget the instruction when extrapolating on a very long context with Prompt-IB, such as Qwen2-72B-Instruction. Examples are in Appendix H.5. Llama3.1 is more sensitive to prompt types. They sometimes even fail to produce any meaningful content with Prompt-IE, see Appendix H.6; this is more frequent in longer samples. But when using Prompt-IB, it does not occur at all.

The types of abnormal outputs are summarized below: (1) Initially outputs normally, but at random moments it starts repeating a sentence or paragraph. (2) After generating a summary, it fails to stop and starts repeating extra prompt instructions. (3) Copy from novels randomly until reaching the length limit. (4) Fail to follow the instructions and output randomly. This may depend on the prompt template. (5) Crash directly and repeat meaninglessly. (6) A combination of the above situations. The repetition issues partly arise from a greedy sampling strategy (Meister et al., 2023; Xu et al., 2022), but they actually reflect the ability deficiencies in handling long text. On the one hand, these stem from the training phase configuration, including data volume of summarization tasks, prompt diversity, and so on. On the other hand, the primary reason is the challenge of length extrapolation. When position embeddings exceed the maximum training sequence length, it may lead to a breakdown in the attention mechanism (Han et al., 2024). Although most baseline LLMs have been extended for extrapolation during pretraining, their stability in summarization tasks remains inadequate.

Rise of Small Long-Context LLMs In the InternLM2.5 series, 7B-1M slightly outperforms 20B and achieves 94% of 20B-Chat average performance. Similar observations are in Qwen, GPT-4o, and Llama3.1. Despite larger LLMs having better reasoning ability, they hardly apply to long-context novel plots. In Table 3, Gemini-pro, which supports longer context, maintains a high recall across all subsets, indicating its stronger ability to memorize the key plots. In contrast, GPT-4o shows a significant decline. We summarize the primary challenges: (1) A longer context length means stable memory ability for key plots. (2) Models with supervised fine-tuning better recognize and follow the instructions after reading long novel texts. For many samples, we observe that GPT-4o summarizes with identical structures and plots to GPT-4o mini, differing only in phrasing, as shown in Appendix H.7. Training small long-context models is significantly less challenging and cheaper, and thus they are cost-effective for novel summarization.

Base Model	CLongEval-LStSum			CNNSum			
	Small	Medium	Large	L	XL	2XL	3XL
Qwen1.5-7B	17.2	**15.2**	10.0	**18.4**	**16.1**	11.4	**3.1**
	17.4	14.8	**11.1**	18.1	15.7	**12.1**	2.9
Qwen1.5-7B-Chat	16.7	14.7	7.4	17.7	15.6	8.5	2.6
	16.8	14.2	8.8	17.2	15.0	9.2	2.6
Yi-6B	**17.2**	14.8	11.9	16.8	**15.6**	12.4	**9.0**
	17.0	**15.0**	12.0	**17.1**	15.5	**12.5**	8.3
Yi-6B-Chat	15.5	14.6	**12.9**	16.6	15.5	**12.5**	6.0
	15.6	14.5	12.4	16.1	15.5	12.3	6.0

Table 5: Gray for Prompt-IE, white for Prompt-IB. **Bold** denotes the best score within each model series.

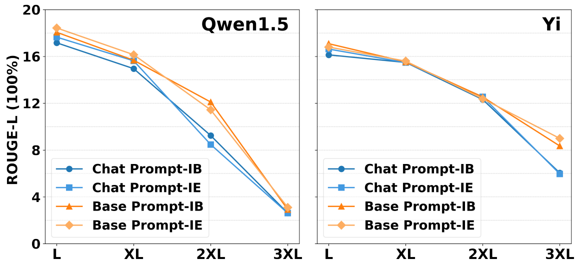


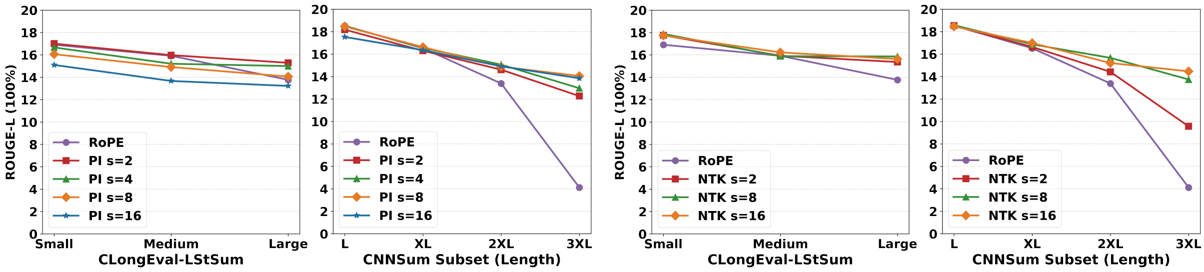
Figure 1: Results of fine-tuned models on CNNSum.

4.5 Ablation Study for Fine-Tuning

We have performed analyses based on various general LLMs. However, fine-tuning powerful models for summarization tasks needs further exploration. We fine-tune Yi-6B series and Qwen1.5-7B series on our training sets with an average length of 16k and evaluate on CNNSum and CLongEval-LStSum (Qiu et al., 2024). CLongEval has three subsets: Small (1k~16k), Medium (16k~50k), and Large (50k~100k). Unlike CNNSum, its sample length distribution is uniform. More statistics are in Appendix E. Results are shown in Table 5.

Extrapolation Potential Compared to Table 2, the performance of Yi-6B on L and XL significantly improves by 2~3 times, while Qwen1.5 also improves by 50%. Despite Yi-6B having the shortest default context length (4k), its RoPE base is scaled up via ABF (Xiong et al., 2024), enabling greater potential for long context. In contrast, the improvement in Qwen1.5 mainly comes from the enhanced ability to follow instructions for summarization tasks, reducing the meaningless outputs. With the largest RoPE base in our baselines (details in Appendix C.2), Yi-6B shows better extrapolation potential on 2~3XL.

Model Version Within each model series, the best scores for most subsets are achieved by the Base version. Fig. 1 indicates that the Base models



(a) Train Average Len. $32k$ with PI
(b) Train Average Len. $32k$ with NTK

Figure 2: Results comparison of Fine-tuned Qwen1.5-7B on CNNSum and CLongEval-LStSum.

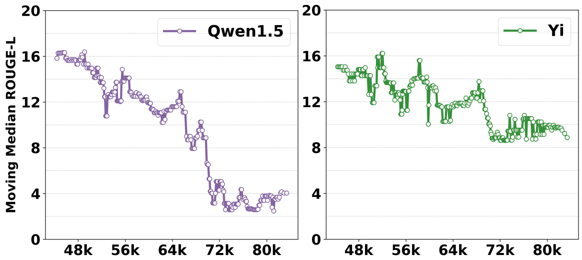


Figure 3: Moving Median ROUGE-L (window size=15) of fine-tuned models on CLongEval-LStSum Large

Train Len. (Ave.)	YaRN($s=16$) $\alpha(Slow)$	$\beta(Fast)$	CNNSum L	XL	2XL	3XL
$14k \sim 18k$ ($16k$)	1	32	17.5	**16.0**	**14.1**	5.8
	1	4	**17.7**	**16.0**	13.6	5.7
	0.1	32	17.0	14.3	11.5	5.5
	RoPE		16.8	15.6	12.4	**9.0**
	PI ($s=16$)		16.1	13.9	12.0	4.1
	NTK ($s=16$)		16.2	14.7	11.3	5.1

Table 6: Gray denotes YaRN default settings. Low β means more no-interpolation dimensions, low α means less "low-frequency" dimensions using PI.

demonstrate better extrapolation potential and are more suitable for efficient fine-tuning. For example, we observe that Qwen1.5-7B-Chat exhibits worse stability when extrapolating on 2XL, having more completely meaningless repetition.

Prompt-IB vs Prompt-IE In Table 4, Qwen1.5-7B shows the largest performance gaps. However, these gaps nearly disappear after fine-tuning. This is confirmed by Fig. 1, which intuitively shows that a fine-tuned model has consistent performance, regardless of the prompt templates. We check that the outputs have no significant difference in quality.

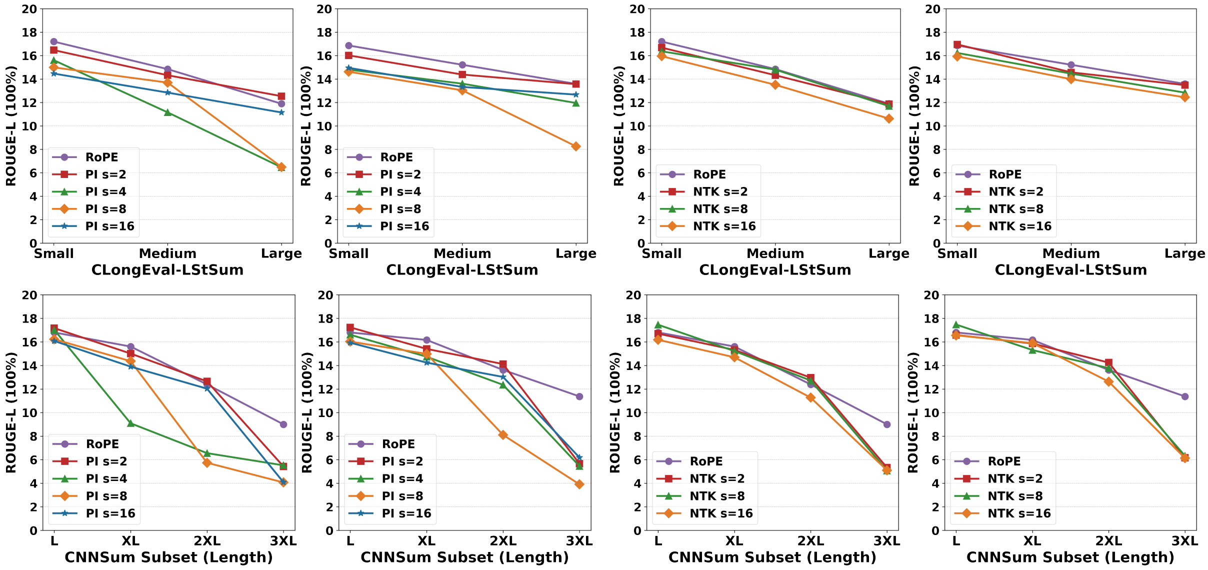
4.6 Reliable Evaluation of Extrapolation

In Table 5, the scores of CLongEval-LStSum Large (50k~100k) and 2XL (64k) are higher than 3XL (128k). We find the output quality has decreased significantly on 2XL, with the type (1) summarized in Section 4.4 occurring frequently. The shorter samples have much better output, leading to misleading final scores, as shown in Fig. 3. To further evaluate long-context summarization extrapolation reliably, we fine-tune models using RoPE scaling methods. Full results can be seen in Appendix G.

The misleading nature of CLongEval further increases. In Fig. 2 (a), CLongEval shows that original RoPE and PI (Fei et al., 2024) perform comparably, obtaining high scores. PI demonstrates a slight decrease in the overall score as s increases, which is a known limitation. Conversely, CNNSum demonstrates that the extrapolation ability of the original RoPE is weaker, leading to a 66% decrease on 3XL. PI with larger s performs poorer on L and XL, but as context increases, smaller s shows a faster decrease due to insufficient interpolation. In Fig. 2 (b), CNNSum also clearly shows more significant differences. Although PI theoretically cannot extrapolate on length without training due to distance resolution variations, models with ABF seem to adapt. NTK (u/LocalLLaMA, 2023) interpolates more moderately than PI, leading to a faster decrease with the same s on 3XL (e.g., $s = 2$).

Yi-6B shows some interesting differences. The original RoPE with a larger base performs best on 3XL. In Fig. 4 (a) and (b), CLongEval shows that PI exhibits an abnormal decrease. CNNSum provides a clear explanation. In Fig. 4 (a), PI with $s = 2$ extrapolates to 2XL (64k). But for $s = 4, 8, 16$, extrapolate respectively to 16k, 32k, and 64k, exactly matching interpolation based on its default 4k context length. Fig. 4 (c) and (d) indicate it is more sensitive to high-frequency information of RoPE in extrapolation. The larger base (50,000,000) significantly diminishes the rotation speed of RoPE. With the default settings of YaRN (Peng et al., 2024), 58% "low-frequency" dimensions (37/64) use PI, and only 20% "high-frequency" dimensions (13/64) are without interpolation. Reducing interpolation leads to slight improvements, as shown in Table 6.

These observations suggest that interpolation methods should be used cautiously for models with RoPE-base scaled. Fine-tuning without interpola-



(a) Train Ave. Len. $16k$ (b) Train Ave. Len. $32k$ (c) Train Ave. Len. $16k$ (d) Train Ave. Len. $32k$

Figure 4: Results comparison of Fine-tuned Yi-6B on CNNSum and CLongEval-LStSum.

tion appears to be stable and straightforward for models with a large RoPE. We also have training observations that differ from (Peng et al., 2024). With the same scale s, PI led to a higher initial loss and slower convergence than NTK and original RoPE. This grows more pronounced as s increases.

5 Conclusion

We present CNNSum, a new long-context Chinese novel summarization benchmark dataset. It has four subsets ranging from L (16k) to 3XL (128k), comprising 695 samples with human annotations. We conduct extensive experiments on CNNSum to investigate the long-context summarization. We conduct a fine-grained human assessment and analysis of the model outputs, revealing the deficiencies and issues of current LLMs in handling long-context summarization. We further explore how to improve the performance of LLMs through fine-tuning with concatenated short-context data and confirm that CNNSum can offer more reliable evaluation results. We hope CNNSum advances future research and provides valuable insights.

6 Limitations

Most advanced evaluation methods require powerful LLMs support, so we do not employ them. The more complex and diverse prompt templates like Appendix I.3 may still introduce subtle variations, which warrant further exploration. This requires more computational resources and detailed human inspections for meaningful conclusions.

Although fine-tuning with concatenated data generally achieves favorable results, its potential drawbacks, such as interference between these short datasets, have not been investigated, which may affect the convergence upper bound. Our fine-tuning experiments are based on parameter-efficient methods. Full fine-tuning may lead to different observations. CNNSum is a Chinese dataset, and we plan to extend future summarization research to English corpora and other domains. The long-context extension techniques are developing rapidly. Our exploration of scaling position embedding and extrapolation provides fundamental guidance, but numerous tasks remain unexplored. We leave these investigations for future work.

7 Acknowledgment

We would like to thank the annotation experts from iQIYI for providing high-quality human annotations and corrections. This work was supported in part by the Key Laboratory of Advanced Theory and Application in Statistics and Data Science, Ministry of Education.

References

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

Mistral AI. 2023. Mistralai announces ministrax. Accessed: 2024-11-10.

Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. 2024. L-eval: Instituting standardized evaluation for long context language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 14388–14411.

Anthropic. 2024. Claude 3.5: Sonnet. Accessed: 2024-6-21.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Yushi Bai, Xin Lv, Jiajie Zhang, Yuze He, Ji Qi, Lei Hou, Jie Tang, Yuxiao Dong, and Juanzi Li. 2024a. Longalign: A recipe for long context alignment of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1376–1395.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024b. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Amanda Bertsch, Uri Alon, Graham Neubig, and Matthew Gormley. 2024. Unlimiformer: Long-range transformers with unlimited length input. *Advances in Neural Information Processing Systems*, 36.

Manik Bhandari, Pranav Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. 2024. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*.

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

Guanzheng Chen, Xin Li, Zaiqiao Meng, Shangsong Liang, and Lidong Bing. 2024a. Clex: Continuous length extrapolation for large language models. In *The Twelfth International Conference on Learning Representations*.

Longze Chen, Ziqiang Liu, Wanwei He, Yinhe Zheng, Hao Sun, Yunshui Li, Run Luo, and Min Yang. 2024b. Long context is not long at all: A prospector of long-dependency data for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. 2024c. Longlora: Efficient fine-tuning of long-context large language models. In *The International Conference on Learning Representations*.

Alibaba Cloud. 2024. Tongyi: Ai large model platform. Accessed: 2024-12-09.

Cohere For AI. 2024. c4ai-command-r-08-2024.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Tri Dao. 2024. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations*.

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*.

Doubao. 2024. Doubao: Ai large model platform. Accessed: 2024-12-09.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

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

Alexander Richard Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084.

Weizhi Fei, Xueyan Niu, Pingyi Zhou, Lu Hou, Bo Bai, Lei Deng, and Wei Han. 2024. Extending context window of large language models via semantic compression. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Yao Fu, Rameswar Panda, Xinyao Niu, Xiang Yue, Hannaneh Hajishirzi, Yoon Kim, and Hao Peng. 2024. Data engineering for scaling language models to 128k context. In *Forty-first International Conference on Machine Learning*.

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

Rajarshi Haldar and Julia Hockenmaier. 2024. Analyzing the performance of large language models on code summarization. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 995–1008.

Chi Han, Qifan Wang, Hao Peng, Wenhan Xiong, Yu Chen, Heng Ji, and Sinong Wang. 2024. Lm-infinite: Zero-shot extreme length generalization for large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 3991–4008.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. 2024. RULER: What's the real context size of your long-context language models? In *First Conference on Language Modeling*.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

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.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

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*.

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

Wojciech Kryściński, 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.

Wai-Chung Kwan, Xingshan Zeng, Yufei Wang, Yusen Sun, Liangyou Li, Lifeng Shang, Qun Liu, and Kam-Fai Wong. 2024. M4le: A multi-ability multi-range multi-task multi-domain long-context evaluation benchmark for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

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*.

Mo Li, Songyang Zhang, Yunxin Liu, and Kai Chen. 2024a. Needlebench: Can llms do retrieval and reasoning in 1 million context window? *arXiv preprint arXiv:2407.11963*.

Taiji Li, Zhi Li, and Yin Zhang. 2024b. Improving faithfulness of large language models in summarization via sliding generation and self-consistency. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 8804–8817.

Yucheng Li, Yunhao Guo, Frank Guerin, and Chenghua Lin. 2024c. An open-source data contamination report for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 528–541.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. 2025. World model on million-length video and language with ringattention. In *International Conference on Learning Representations*.

Hao Liu, Matei Zaharia, and Pieter Abbeel. 2024a. Ring attention with blockwise transformers for near-infinite context. *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.

Xiaojun Liu, Shunan Zang, Chuang Zhang, Xiaojun Chen, and Yangyang Ding. 2022. Clts+: a new chinese long text summarization dataset with abstractive summaries. In *International Conference on Artificial Neural Networks*, pages 73–84. Springer.

Xiaoran Liu, Hang Yan, Chenxin An, Xipeng Qiu, and Dahua Lin. 2024c. Scaling laws of rope-based extrapolation. In *The Twelfth International Conference on Learning Representations*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. 2023. Locally typical sampling. *Transactions of the Association for Computational Linguistics*, 11:102–121.

MoonshotAI. 2024. Moonshot ai platform. Accessed: 2024.

Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*.

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. *arXiv preprint arXiv:1808.08745*.

Shiwen Ni, Xiangtao Kong, Chengming Li, Xiping Hu, Ruifeng Xu, Jia Zhu, and Min Yang. 2024a. Training on the benchmark is not all you need. *arXiv preprint arXiv:2409.01790*.

Xuanfan Ni, Hengyi Cai, Xiaochi Wei, Shuaiqiang Wang, Dawei Yin, and Piji Li. 2024b. Xl $\hat{2}$ˆ 2 bench: A benchmark for extremely long context understanding with long-range dependencies. *arXiv preprint arXiv:2404.05446*.

OpenAI. 2024a. Gpt-4o mini: Advancing cost-efficient intelligence. Accessed: 2024.

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

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.

Saurav Pawar, SM Tonmoy, SM Zaman, Vinija Jain, Aman Chadha, and Amitava Das. 2024. The what, why, and how of context length extension techniques in large language models–a detailed survey. *arXiv preprint arXiv:2401.07872*.

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*.

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead. *arXiv preprint arXiv:2309.09558*.

Zexuan Qiu, Jingjing Li, Shijue Huang, Xiaoqi Jiao, Wanjun Zhong, and Irwin King. 2024. CLongEval: A Chinese benchmark for evaluating long-context large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

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*, pages 2764–2781.

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 *2021 USENIX Annual Technical Conference*, pages 551–564.

r/LocalLLaMA. 2024. Dynamically scaled rope further increases context length. Accessed: 2024-11-08.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*.

Eva Sharma, Chen Li, and Lu Wang. 2019. Bigpatent: A large-scale dataset for abstractive and coherent summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2204–2213.

Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. 2023. Large language models are not yet human-level evaluators for abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

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*, pages 906–922.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Derek Tam, Anisha Mascarenhas, Shiyue Zhang, Sarah Kwan, Mohit Bansal, and Colin Raffel. 2023. Evaluating the factual consistency of large language models through news summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5220–5255.

Chaofan Tao, Qian Liu, Longxu Dou, Niklas Muennighoff, Zhongwei Wan, Ping Luo, Min Lin, and Ngai Wong. 2024. Scaling laws with vocabulary: Larger models deserve larger vocabularies. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

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

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024b. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Junfeng Tian, Da Zheng, Yang Cheng, Rui Wang, Colin Zhang, and Debing Zhang. 2024. Untie the knots: An efficient data augmentation strategy for long-context pre-training in language models. *arXiv preprint arXiv:2409.04774*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: open and efficient foundation language models. arxiv. *arXiv preprint arXiv:2302.13971*.

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

u/LocalLLaMA. 2023. Ntk-aware scaled rope allows llama models to have context lengths exceeding 4k tokens. https://www.reddit.com/r/LocalLLaMA/comments/14lz7j5/ntkaware_scaled_rope_allows_llama_models_to_have/. Accessed: 2024-11-23.

Danqing Wang, Jiaze Chen, Xianze Wu, Hao Zhou, and Lei Li. 2021. Cnewsum: a large-scale chinese news summarization dataset with human-annotated adequacy and deducibility level. *arXiv preprint arXiv:2110.10874*.

Xindi Wang, Mahsa Salmani, Parsa Omidi, Xiangyu Ren, Mehdi Rezagholizadeh, and Armaghan Eshaghi. 2024. Beyond the limits: a survey of techniques to extend the context length in large language models. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 8299–8307.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, et al. 2024. Effective long-context scaling of foundation models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4643–4663.

Jin Xu, Xiaojiang Liu, Jianhao Yan, Deng Cai, Huayang Li, and Jian Li. 2022. Learning to break the loop: Analyzing and mitigating repetitions for neural text generation. *Advances in Neural Information Processing Systems*, 35:3082–3095.

Ruijie Xu, Zengzhi Wang, Run-Ze Fan, and Pengfei Liu. 2024. Benchmarking benchmark leakage in large language models. *arXiv preprint arXiv:2404.18824*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Hao, Xu Han, Zhen Thai, Shuo Wang, Zhiyuan Liu, and Maosong Sun. 2024. ∞Bench: Extending long context evaluation beyond 100K tokens. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

A Long-Context Data Quality Assessment

The key to long-context data lies in its long-distance dependencies, not merely in having a long sequence length. We employ the ProLong (Chen et al., 2024b) framework for assessment, which can assign a long dependency score (LDS) to

Dataset	Ave. Len.	Win. Size	Ave. LDS
CNNSum			
- *L*	$16k$	$16k$	950.9
- *XL*	$32k$	$32k$	774.3
CLongEval-LStSum			
- *Medium*	$28k$	$32k$	763.6
- *Large*	$64k$	$32k$	753.0
LongBench-VCSUM	$9k$	$9k$	613.2
Reference from (Chen et al., 2024b)			
Book3	$165k$	$32k$	385.4
BooksCorpus2	$103k$	$32k$	327.2
PG19	$137k$	$32k$	469.9

Table 7: **Ave. Len.** denotes the average length of the dataset. **Win. Size** denotes the context window. **Ave. LDS** is the average long dependency score.

Model Series	Vocab Size	Chars / Token
Llama / Llama2 / LWM	32,000	0.65
Mistral	32,768	0.82
Llama3 / Llama3.1	128,256	1.12
Ministral-8B-Instruct-2410*	131,072	1.01
Yi / Yi-1.5	64,000	1.34
ChatGLM2 / ChatGLM3	65,024	1.42
InternLM2 / InternLM2.5	92,544	1.40
Baichuan2	125,696	**1.48**
Qwen1.5 / Qwen2 / Qwen2.5	152,064	1.38
GLM4	151,552	1.43
Command R	**256,000**	1.23
Gemma 2	**256,000**	1.36
GPT-4o-2024-08-06*	?	1.15

Table 8: We calculate the number of characters per token represented. * denotes the specific model.

each sample in a dataset. Most settings adhere to the suggested defaults, but we replace the default OPT-350M (Zhang et al., 2022) with Qwen2.5-7B (Qwen et al., 2024), which supports Chinese better and has larger parameters. The ablation study of ProLong shows that while small LLMs offer better cost-effectiveness considering computational costs, larger 7B LLMs provide superior performance. Additionally, ProLong uses a default context window of 32k for computing scores. For datasets with sufficient sample lengths, we keep this unchanged. For those with shorter lengths, we set the context window to their average length. We assess Chinese summarization benchmarks: CNNSum, CLongEval-LStSum (Qiu et al., 2024), and LongBench-VCSUM (Bai et al., 2024b). We also refer to the results of ProLong (Chen et al., 2024b) on large-scale English book datasets for comparison. The results shown in Table 7 indicate that CNNSum has high long-distance dependency, making it suitable for Chinese long-context summarization research. Consequently, the findings and insights derived from it should be fairly reliable.

B Commercial Models for Annotation

Annotation team uses various commercial APIs, including Qwen (Cloud, 2024), Doubao (Doubao, 2024), Kimi (MoonshotAI, 2024), and Gemini (Team et al., 2024a). Due to the short length of chapters and cost considerations, we opt for their short-context versions instead of the long-context versions that are evaluated in Table 2. Specifically, these include: (1) Qwen-turbo-2024-06-24, with a context length of 8k. (2) Doubao-pro-32k. (3) Moonshot-v1-8k. (4) Gemini-1.5-flash, with a context length of 100k. Notably, we do not employ any open-source LLMs for annotation.

C Long-Context LLMs for Chinese

C.1 Tokenizer Efficiency

The Chinese encoding efficiency is shown in Table 8. Mistral (Jiang et al., 2023) and LWM (Liu et al., 2025), derived from Llama (Touvron et al., 2023a,b), support context lengths of up to 128k and 1M, respectively, but still exhibit low encoding efficiency. The actual text they can process is limited compared to the context length defined by tokens. Recently released Llama3.1 (Dubey et al., 2024) and Ministral (AI, 2023) benefit from a substantial increase in vocabulary size, resulting in significantly improved encoding efficiency, though still not ideal, GTP-4o (OpenAI, 2024b,a) as is the case. Bilingual LLMs Yi (Young et al., 2024), ChatGLM3 (Du et al., 2022; GLM et al., 2024), and InternLM (Cai et al., 2024), all developed by Chinese teams, achieve high encoding efficiency with a vocabulary size of less than 10k. Qwen (Bai et al., 2023; Yang et al., 2024; Qwen et al., 2024) and GLM4 (GLM et al., 2024) support better for multiple languages with 15k vocabulary size. The vocabulary size of Command R (Cohere For AI, 2024) and Gemma 2 (Team et al., 2024b) exceeds 25k, leading to large embedding layer parameters. This may significantly increase costs for methods requiring fine-tuning the embedding layer (Tao et al., 2024), such as LongLoRA (Chen et al., 2024c).

C.2 Open-Source Models Details

(1) Yi extends context length via ABF (Xiong et al., 2024). Yi-6B (base=5,000,000) and Yi-6B-Chat default to 4k context, consistent with the training phase. Yi-6B-200K and Yi-34B-200K use a larger RoPE Base (10,000,000) to support

Model Series	Small (Count=300)					Medium (Count=400)					Large (Count=300)				
	Min	Q1	Mean	Q3	Max	Min	Q1	Mean	Q3	Max	Min	Q1	Mean	Q3	Max
Yi / Yi-1.5	1,078	3,865	6,979	9,880	14,995	13,239	20,053	28,352	37,049	53,082	42,725	53,531	63,493	73,350	87,735
Qwen1.5 / Qwen2.5	1,095	3,863	6,978	9,846	14,700	13,552	20,011	28,293	36,928	51,984	42,705	53,279	63,288	73,061	87,195

Table 9: Token sequence length of CLongEval-LStSum. Q1 and Q3 denote the first and third quartiles, respectively.

200k. Yi-1.5-34B-32K and Yi-1.5-34B-Chat-16K are the largest long-context models in the Yi-1.5 series. (2) InternLM2.5 also extends by scaling the RoPE Base (Liu et al., 2024c) and defaults to using DynamicNTK (r/LocalLLaMA, 2024). For context length, InternLM2.5-20B defaults to 256k, InternLM2.5-20B-Chat is 32k, and the special version InternLM2.5-7B-Chat-1M supports 1 million. (3) ChatGLM3-6B-128K is 128k version of ChatGLM3-6B. GLM4-9B-Chat-1M is a special version of GLM4 using LongAlign (Bai et al., 2024a) and ABF. (4) Llama3.1 is 128k version of Llama3 modified RoPE using a YaRN-like (Peng et al., 2024) method. (5) LWM-Text-1M is the language-only version of Large World Model (Liu et al., 2025), support 1 million context, training with RingAttention (Liu et al., 2024a) based on Llama2 (Touvron et al., 2023b). (6) Ministral-8B-Instruct-2410 is an edge model with 128k context from Ministral series. (7) Qwen series also uses ABF. Qwen1.5 default context lengths are all 32k. For Qwen2 (Yang et al., 2024) and Qwen2.5 (Qwen et al., 2024), the Base version defaults to 128k, while the Instruction version defaults to 32k.

D Experimental Setup Details

D.1 Baseline Evaluation

Open-Source Models We modify the vLLM source code related to positional embedding for correct extrapolation. Specifically, HuggingFace Transformers (Wolf et al., 2020) dynamically allocate more positions based on the sequence length, whereas vLLM uses a static pre-allocation of positions according to the default context length of models. Our modification simply increases the pre-allocated positions without any complex changes. The modified files will be released along with the dataset. For models that default to using RoPE scaling methods, we keep their original settings. Qwen series Instruct models can optionally enable YaRN for long context. For consistency and extrapolation evaluation, we keep the default RoPE settings.

Commercial Models Gemini and Qwen have strict content safety checks, which block 29% and 2% of samples, respectively. Doubao excluded 8%

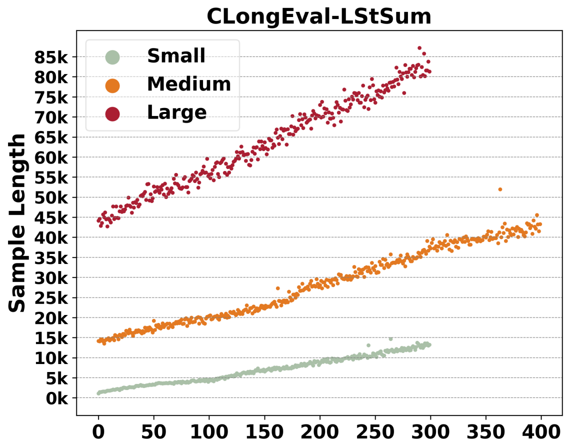


Figure 5: The sample length distribution of each subset, based on Qwen1.5 tokenizer.

of samples on 3XL due to excessive length. These cases introduce score bias.

D.2 Fine-tuning Experiments

We used Flash Attention 2 (Dao, 2024), DeepSpeed ZeRO2 / ZeRO3 with CPU offloading (Ren et al., 2021), and Adam (Loshchilov and Hutter, 2019) with a learning rate 2e-5, including a linear warmup of 30 steps. Inference also used vLLM. For fine-tuning concatenated data with an average length of 16k, we set the global batch size to 16. Due to the varying convergence rates of RoPE-based scaling methods with different scales s (Peng et al., 2024), we evaluated multiple checkpoints between 400 and 500 steps and selected the best result. For fine-tuning on concatenated data with an average length of 32k, we set the global batch size to 8 and started from a checkpoint that had been fine-tuned for 300 steps on data with an average length of 16k. We continued fine-tuning for another 200 to 300 steps and selected the best result. All experiments were conducted on NVIDIA A100 (80GB) GPUs.

E Statistics of CLongEval-LStSum

Detailed statistics based on Yi and Qwen tokenizers are shown in Table 9. Compared to our CNNSum, their average length is insufficient, and the distances between the first and third quartiles and the mean are farther. Figure 3 clearly shows the characteristic of sample length uniform distribution.

F Full Baseline Results of the Two Prompts

Model	Ctx. Len.	CNNSum			
		L ($16k$)	**XL** ($32k$)	**2XL** ($64k$)	**3XL** ($128k$)
GPT-4o-2024-08-06	$128k$	15.3	14.2	12.4	-
		15.7	14.2	12.5	-
GPT-4o-mini-2024-07-18	$128k$	16.0	13.8	13.1	-
		15.3	13.7	12.5	-
Gemini-1.5-pro	$2m$	19.1	17.9	16.7	14.6
		19.5	18.4	16.9	14.6
Doubao-pro-128k	$128k$	20.2	17.5	15.2	12.8
		19.1	18.1	15.7	13.6
Qwen-plus-2024-09-19	$128k$	20.3	18.5	16.2	14.5
		20.6	18.6	16.7	15.0
Moonshot-v1-128k	$128k$	21.8	19.7	17.6	14.7
		22.9	20.8	18.4	15.6
Yi-6B	$4k$	9.3	3.5	2.4	2.2
		8.3	5.3	2.9	2.4
Yi-6B-Chat	$4k$	13.5	7.5	2.0	2.0
		13.4	6.2	2.2	2.2
Yi-6B-200K	$200k$	11.0	10.8	9.6	4.6
		8.7	7.9	8.0	3.3
Yi-34B-200K	$200k$	12.0	11.4	10.8	10.3
		12.1	11.2	10.8	9.6
Yi-1.5-34B-32K	$32k$	14.4	12.7	10.8	0.1
		8.8	8.2	8.3	0.1
Yi-1.5-34B-Chat-16K	$16k$	15.4	14.2	12.1	0.0
		12.1	10.4	9.7	0.0
InternLM2.5-7B-Chat-1M	$1m$	17.8	17.3	15.0	13.8
		18.2	16.9	14.3	12.1
InternLM2.5-20B	$256k$	18.0	16.3	14.6	12.9
		18.3	16.3	14.1	12.5
InternLM2.5-20B-Chat	$32k$	18.8	17.5	16.3	14.2
		19.0	17.1	16.1	14.2
ChatGLM3-6B-128k	$128k$	17.3	16.2	14.9	13.9
		17.1	16.0	14.6	13.5
GLM4-9B-Chat-1M	$1m$	19.2	18.2	16.6	15.6
		18.8	17.6	16.4	15.2
Llama3.1-8B	$128k$	6.1	6.6	7.0	2.2
		9.4	9.4	9.7	3.9
Llama3.1-8B-Instruct	$128k$	16.1	15.0	12.8	10.6
		15.0	13.5	12.7	9.2
Llama3.1-70B	$128k$	11.8	7.5	5.2	7.0
		12.5	10.2	8.6	7.9
Llama3.1-70B-Instruct	$128k$	18.6	15.8	13.9	10.5
		17.3	15.6	13.7	10.7
LWM-Text-1M	$1m$	3.1	2.8	2.2	1.1
		3.5	3.2	2.8	1.0
Ministral-8B-Instruct-2410	$128k$	16.1	14.1	11.6	3.5
		16.3	14.0	11.0	3.5
Qwen1.5-7B	$32k$	16.5	14.5	7.0	2.5
		7.7	7.2	4.4	2.6
Qwen1.5-7B-Chat	$32k$	14.8	14.1	10.3	2.7
		15.4	13.1	10.3	2.7
Qwen1.5-32B-Chat	$32k$	15.8	14.2	4.5	4.2
		15.6	14.1	5.5	4.2
Qwen2-7B	$128k$	8.8	8.2	8.6	7.1
		10.8	8.5	7.3	6.6
Qwen2-7B-Instruct	$32k$	15.1	13.7	13.3	11.7
		15.3	12.9	11.2	10.8
Qwen2-72B	$128k$	12.1	10.7	10.3	9.5
		17.5	15.7	13.1	11.3
Qwen2-72B-Instruct	$32k$	12.4	10.5	10.6	10.6
		18.5	15.8	12.3	7.8
Qwen2.5-7B-Instruct	$32k$	17.0	15.2	12.7	10.1
		14.4	12.8	12.7	9.1
Qwen2.5-72B-Instruct	$32k$	19.7	18.3	13.8	12.5
		19.5	16.9	13.3	14.2

Table 10: Gray denotes using Prompt-IE. White denotes using Prompt-IB.

G Detailed Results of Extrapolation

Most of the results have been visualized and analyzed in the experimental Section 4.6.

Base Model	Training Len. (Ave.)	Pos. Emb.	CLongEval-LStSum			CNNSum			
			Small	**Medium**	**Large**	**L** ($16k$)	**XL** ($32k$)	**2XL** ($64k$)	**3XL** ($128k$)
Qwen1.5-7B	$14k \sim 18k$ $(16k)$	RoPE	17.2	15.2	10.0	18.4	16.1	11.4	3.1
		PI ($s = 2$)	16.9	14.9	15.1	18.3	16.6	14.6	7.4
		PI ($s = 4$)	15.7	14.5	14.1	17.3	15.8	14.3	12.7
		PI ($s = 8$)	16.1	14.1	13.6	17.4	15.6	14.1	12.8
		PI ($s = 16$)	15.2	13.1	12.3	16.7	14.9	13.3	12.1
		NTK ($s = 2$)	17.1	14.5	14.8	18.1	16.4	14.2	2.9
		NTK ($s = 8$)	16.9	14.3	13.5	18.4	16.2	14.5	13.3
		NTK ($s = 16$)	16.9	14.0	13.1	17.8	15.8	14.5	12.6
Qwen1.5-7B	$30k \sim 34k$ $(32k)$	RoPE	16.9	15.9	13.8	18.5	16.5	13.4	4.1
		PI ($s = 2$)	17.0	16.0	15.3	18.2	16.3	14.6	12.3
		PI ($s = 4$)	16.7	15.2	15.0	18.5	16.5	15.1	13.0
		PI ($s = 8$)	16.1	14.9	14.1	18.5	16.6	14.9	14.0
		PI ($s = 16$)	15.1	13.7	13.2	17.5	16.3	14.9	13.8
		NTK ($s = 2$)	17.7	15.9	15.4	18.5	16.6	14.4	9.6
		NTK ($s = 8$)	17.9	15.9	15.9	18.6	16.9	15.7	13.7
		NTK ($s = 16$)	17.7	16.2	15.6	18.5	17.0	15.2	14.5
Yi-6B	$14k \sim 18k$ $(16k)$	RoPE	17.2	14.8	11.9	16.8	15.6	12.4	9.0
		PI ($s = 2$)	16.5	14.3	12.5	17.2	15.0	12.6	5.4
		PI ($s = 4$)	15.6	11.2	6.5	17.0	9.1	6.6	5.5
		PI ($s = 8$)	15.0	13.7	6.5	16.2	14.4	5.7	4.1
		PI ($s = 16$)	14.5	12.8	11.1	16.1	13.9	12.0	4.1
		NTK ($s = 2$)	16.7	14.3	11.9	16.7	15.3	13.0	5.3
		NTK ($s = 8$)	16.4	14.8	11.7	17.5	15.2	12.7	5.1
		NTK ($s = 16$)	15.9	13.5	10.6	16.2	14.7	11.3	5.1
Yi-6B	$30k \sim 34k$ $(32k)$	RoPE	16.9	15.2	13.6	16.8	16.2	13.6	11.4
		PI ($s = 2$)	16.0	14.4	13.6	17.2	15.4	14.1	5.7
		PI ($s = 4$)	14.8	13.6	12.0	16.6	14.7	12.4	5.5
		PI ($s = 8$)	14.6	13.0	8.3	16.0	15.0	8.1	3.9
		PI ($s = 16$)	15.0	13.3	12.7	15.9	14.2	13.0	6.2
		NTK ($s = 2$)	17.0	14.6	13.5	16.6	15.9	14.2	6.2
		NTK ($s = 8$)	16.2	14.5	12.8	17.5	15.3	13.8	6.3
		NTK ($s = 16$)	15.9	14.0	12.4	16.6	15.9	12.6	6.1

Table 11: Results of fine-tuned models on CLongEval-LStSum and CNNSum. Training with concatenated data of average lengths $16k$ and $32k$, and position embeddings scaled by factors s.

H Examples of Cases

H.1 Cases in Web-Serialized Novels Corpus

Case 1 demonstrates that the author does not use standard quotation marks to enclose character dialogues but instead opts for corner brackets. This novel contains an amount of dialogue, leading to the frequent appearance of corner brackets.

Case 2 shows that the author prefers to use line breaks, where nearly every sentence, regardless of its length, begins on a new line. This increases the proportion of unnecessary characters, reducing the effective context. Additionally, it may lead the model to generate outputs containing excessive line breaks.

Case 1: Non-standard Punctuation Usage

金四爺冷冷道：「我們還真沒有想到是你。」楚留香笑了笑，道：「我也沒想到金四爺居然還認得我。」金四爺沉著臉，道：「像你這樣的人，我只要看過一眼，就絕不會忘記。」楚留香道：「哦？」金四爺道：「你有張很特別的臉。」楚留香道：「我的臉特別？」金四爺道：「無論誰有你這麼樣的一張臉，再想規規矩矩的做人都難得很。」

Case 2: Frequent Use of Line Breaks

这一瞬，那神秘青年的身上，那种霸道凌厉的凶悍之气，似乎不应该在人类中出现，只有妖兽至尊的身上才可能存在。\n
\n
爪印之上金光爆发，犹如无数金色电蛇掠出，以一种霸道无匹的姿态，势如破竹，直接摧毁绝剑王那可怕剑芒。\n
\n
势如破竹，一切溃败开去！\n
\n
漫天的符文破碎，可怕的能量风暴劲气，呈弧形扩散水域。\n
\n
可怕能量像是要将水域掀开，露出真空，大片的礁石尽数被席卷化作齑粉。\n
\n
“嗤啦......”\n
\n
爪印穿透空气，直接抓在了绝剑王手中的符器长剑上，生生将那符器长剑掠走。\n
\n
“噗嗤......”\n
\n
手中符器被生生掠走，绝剑王嘴中一口鲜血喷出，身躯踉跄向后退去，面色霎时间惨白，双眸目光在此刻似乎是感觉到了什么，骤然间就变得震骇了起来。\n
\n

Case 3 and 4 The author breaks the fourth wall, as the red-highlighted content. This content mainly involves interactions with readers, including soliciting their support, previewing update schedules, and other casual remarks. Such content typically appears at the beginning or end of the chapters.

Case 5 The author adds additional annotations at the end of chapters to explain and provide background information on some elements mentioned. This content is usually unrelated to the novel main storyline.

Case 3: Interactive Content with the Reader at the Chapter End

...这个时候，一个白衣女子的身影，悄然在石门旁浮现。“她出现了。”无需马文提醒，梅迪尔丽的法术早已准备多时！强效缠绕术！ ......ps：作为月票榜守门员（总榜第十），椰子压力好大。因为随时有被爆菊的危险，眼看后面的那位兄台月票数狂飙就要爆上来了，椰子实在是坐立不安啊。再这么下去，估计都要变成求票小王子了。大家看着办吧。这两天看看，能不能加更或者爆发。这个月的目标就是稳住第十就好。真心希望可以顶住。

...这是它的名字。当然，它还有一个更简单粗暴的外号——屠龙枪。费南赫赫有名的屠龙三件套之一。......ps：第一更送上！今天尽力多写点，看看能不能爆发！昨天写完一个策划类累晕了，一看跌到了11名，这意味着下榜了。今天，我们要重返前十！

...只不过令人惊讶的是，城墙下，居然三三五五聚集着不少难民。他们满脸愁容，不知道发生了什么事。“怎么回事？”马文心中一动。就在这个时候，城门忽然缓缓打开了一个口子。......【ps：感谢书友亡灵邪魂打赏的1w点，在暗夜最艰难的时刻，感谢所有仍然支持的书友们。这是第二更，今天依然会爆发，至少四更，双倍月票最后一天了，虽然知道榜单什么的已经暂时无缘，但是还是能拉一些就拉一些吧。另外，务必顺手丢个推荐票。谢谢。】(未完待续。)

Case 4: Interactive Content with the Reader at the Chapter Beginning

【第三更，昨天的缺章补上了。求红票，收藏。】当初渭源之战时，从古渭到渭源，长达一百多里的粮秣军资的转运，就已经耗尽了秦州泰半民力...

【今天晚上赶不下去了，就只有一章。缺的明天补上。】新磨的铜镜光可鉴人，镜中的一张如hua俏脸有着倾城之姿，却是略显憔悴...

【真的不好意思，实在吃不住了，明天还要上班，昨天缺的一章只能有空再补。】韩冈派李xiao六去章家递名帖，约时相见的时候，章却正在王安石府中...

【没能如约把欠下的两章补上，这里要说声抱歉。不过不会忘的，肯定、一定、必定会补上。】京城的冬天寒冷干燥，一不注意...

Case 5: Additional Annotations

...卷子望向父亲，看到莫名的光辉在他眼中闪动。父亲的脸上，先前那对人生的疲惫与厌倦，已全然不见踪影，取而代之的是充满生气的神采——那是一个男人有自己深爱的人，有自己必须要保护的人时才会有的神情。注释1 一种形状粗短，但底部细长形似鼠尾的萝卜，以坂城町所产最为著名。2 即通过嗅吸胶水、汽油等物质中的挥发溶剂，以产生暂时的快感，有一定的成瘾性，并且极其危险，易致窒息、器官衰竭等。上世纪中叶曾在东南亚等地泛滥，甚至成为台湾青少年首要的滥用药物。3 即用奄美大岛的特产“大岛绸”做的和服，因为采用泥染工艺，所以又俗称“泥大岛”。

H.2 Outputs Comparison of Commercial Models

The red-highlighted content is subjective commentary generated by GPT-4o. Such as the second red-highlighted part includes a key plot that is clearly and objectively described in Gemini-pro and Moonshot-v1, with a high overlap of keywords. In contrast, GPT-4o's subjective description is overly vague and includes few keywords.

GPT-4o-2024-08-06

故事讲述了年轻小医师苏小煜在职业和生活中挑战权威、寻求学术成就的艰辛历程。苏小煜在行医四世的苏家医馆中努力追求医术精进。他得到一个神秘系统协助通过任务获得医学奖励。当一名自信的杭城贵公子云天嵩前来鄞县求医时，苏小煜凭借系统提供的药方赢得了他的信任。然而，苏小煜却因为过于依赖系统而感到不安，对系统提供的奖励又心存顾虑。围绕着如何准确炮制药物，他寻求了隐士高人"高老叫花子"的帮助，展现了他的决心和悟性。最终，在家庭困境与未来发展之间，苏小煜努力寻找平衡，也在与高人交流中收获更多医学知识，开始意识到自身挑战与成长的重要性。

Gemini-1.5-pro

苏小煜凭借从系统药方中得知的云天嵩病症，赢得了云天嵩的信任，并获得高价诊金和金字招牌的承诺。苏小煜将问诊交给李老，自己则专注于炮制特殊的鳖血柴胡。在炮制过程中，苏小煜屡次失败，最终在高老叫花子的指点下成功炮制出上品鳖血柴胡，并因此拜高老叫花子为师学习针灸之术。同时，苏小煜 接受母亲的建议，打算去云德堂当学徒，但最终被大伯安排在苏庆堂打下手。最后，苏小煜完成了系统任务，获得柴胡疏肝丸和一次抽奖机会，意外抽得《内门八段锦》，并得知修炼至小成可以以气运针，但任务系统也因此暂时关闭。

Moonshot-v1-128k

苏小煜在苏庆堂的医术受到质疑，尤其是云天嵩对他的年轻和经验表示不信任。然而，苏小煜凭借自己的医术和一张神秘的药方——柴胡解毒汤，成功诊断出云天嵩的病症，赢得了云天嵩的信任。云天嵩承诺，如果苏小煜能治好他的病，将给予重赏。苏小煜决定亲自炮制一种特殊的药材——鳖血柴胡，以治疗云天嵩的病症。在炮制过程中，他遇到了困难，鳖血柴胡的颜色和质量未能达到预期，这让他感到困惑和沮丧。高老叫花子，一个看似普通的乞丐，对苏小煜的炮制方法提出了质疑，并暗示自己知道正确的炮制方法。苏小煜在多次失败后，意识到问题的关键在于鳖血的处理和水的选择。最终，高老叫花子答应在苏小煜带来新的鳖后，教授他正确的炮制方法。苏小煜在这一过程中，不仅学习到了医术知识，还体会到了家族内部的复杂关系和权力斗争。他的母亲柳氏建议他去云德堂当学徒，以避免在苏庆堂受到打压，苏小煜 决定 接受这个建议，开始自己的医学之路。

CNNSum Annotation

苏小煜从云天嵩那里得了五两银子准备买药材，他在鄞县买了两只鳖和柴胡。在破庙里按照药方炮制鳖血柴胡，却因为火候控制不当而失败。他因此被高老叫花子嘲讽，但高老叫花子随后指出问题所在，炮制鳖血柴胡需要考虑火候、鳖血处理、天气影响和柴胡品种。在多次失败后，高老叫花子建议他再次买鳖。在再次炮制过程中，苏小煜还计划教其他孩子识别药材，希望他们能以此谋生。苏小煜去王老财家上门诊病，调整了他的药方，并借此机会让他用马车送自己回家。在家中，苏小煜与家人起了争执，他决定去云德堂当学徒，但苏炳坚持让他在苏庆堂帮忙，苏小煜最终接受了这个安排。高老叫花子指导了苏小煜准备鳖血柴胡，最终还答应他在适当的时候教授苏小煜针灸之术。完成任务后，苏小煜得到了柴胡疏肝丸和一次抽奖机会。他抽中了《内门八段锦》，这本书可以帮他修运用针灸之术。

H.3 Outputs of Base Models with Prompt-IE vs Prompt-IB

Case 1 (XL) output by Qwen1.5-7B with Prompt-IE (Normal Summary)

阳旭在星元古坟场中，得到了玄元控水旗和阳极炼神丹。玄元控水旗是鸿途大陆的核心精华，与天花妙坠旗融合后，成为了完美道器。阳旭在风雷台上，连续斩杀龙印境五重的火道人和天星子，引起了学院的重视。阳旭得到了学院的奖励，但他的目标是天机门，想要进入蛮兽荒山。在蛮兽荒山，阳旭遇到了猛犸飞象，通过八相世界之巽风世界，降服了猛犸飞象。阳旭的实力得到了其他学院势力的重视，但阳旭并不在意，他继续前往蛮兽荒山。

Case 1 (XL) output by Qwen1.5-7B with Prompt-IB (Copy From the Novel Excerpt)

\n\n\n小说节选如下：\n\n在那神秘声音的指引下。阳旭与众位天才，背道而驰，一路向南。路途中，他也遇到了几个往北部山峰赶去的试炼者。看阳旭的眼神，或是充满轻视。或是充满鄙夷。好像在看一个失败者。阳旭用莫大的毅力，才说服自己不要去北方淌这趟浑水。“希望你靠谱一些，不要让我失望。”阳旭道。“放心，绝对比你去那座山，收获要大。”一路上，阳旭又狩猎了几枚星魂碎片。他知道，自己如今的名次很靠前。但并没把握是第一名。自然。他也想象不到，自己在北域学院中，掀起了怎样一阵轩然大波。尤其是灰衣人和白衣女子，已经因为他，而展开了一场理念上的交锋。呼！星元古坟场上，到处都是一片混沌。越向战场的外围，一道道阴冷的罡风就越发凌厉。阳旭的行进速度，也慢了下来。“八卦星轮！”“八相世界！”阳旭索性展开极速，瞬间移动。半天时间过去。光幕之上，阳旭以绝对的星魂碎片数量，牢牢占据第一名。第二名天唯一与他的差距，不但没有缩小。反而有渐渐拉大的趋势。这令得白衣女子有些着急。灰衣人看在眼里，眼底闪过一丝担忧：阳旭的出色，有些出乎他的预料啊。希望不要引起她的想法才好……周围的罡风，越发凌厉了。阳旭每一次瞬移，都跨越千米的距离。终于来到了古坟场的最南端。阳旭站在一处深不见底的漆黑巨坑边缘。“这就是你的宝贝？我得跳到这个不知深浅的坑里去？”“没错。富贵险中求。”阳

Case 2 (XL) output by Qwen1.5-7B with Prompt-IE (Normal Summary)

桑远等人在游玩中遇到了一个五岁的小男孩莫小宝，他被莫文带走后，莫文和战锋轮流照顾他。莫小宝因为水土不服而发烧，桑远等人将他送到了医院。在莫小宝生病期间，莫文给他讲故事，战锋给他讲了桑远小时候的故事。莫小宝在听完故事后，向战锋请求学习功夫，战锋答应了他。第二天，陕西墨家的人来找桑远，希望将莫小宝带回去，但桑远拒绝了他们。

Case 2 (XL) output by Qwen1.5-7B with Prompt-IB (Copy From the Novel Excerpt)

他的话一出，就连桑瑛子都有些动容。也难怪当初金凤淑拼死也要与他们家脱离，就是为了给孩子一个健康的成长环境！想到这里，桑瑛子就略带羞赧的问了一句：“桑远，如果我们有事求你，你会帮我们吗？”他把地还给桑家的事肯定瞒不住，她跟桑红虽然回家了，但两人都没有跟婆家断绝关系，那些势利眼绝对会想法设法的来讨要好处，那，不是她愿意看到的。“那要看你们求的是什么事。”桑远也没想到她会说这种话，不过一想到她们的状况，他就明白她说的是什么事。只是，真的到了那一步，也要看她们自己是如何选择的。“行，那我就先记住了！”桑瑛子因为受了桑大军的影响，想法也就果断多了。重新笑了起来，跟他说着别的。桑红情绪稳定了一些，也插上两句话，一时间竟然觉得活了二十几年，只有此时最舒心！习绍面无表情的听着他们有一搭没一搭的话，偶尔开口刺一句，让气氛尴尬一下，但桑瑛子根本不在意，还仗着桑远在打趣他，让桑红和桑远都看得啧啧称奇。习绍又不能太过分，有气也憋在肚子里，倒是让桑红也跟着胆大起来。他们大人这边其乐融融，小孩子那边也就其乐融融起来。小丫和妞妞吃着糖果，带着习风吟给人端茶递水，笑容也多了起来。结果走的时候，习风吟就被两人一口一个哥哥的叫着，别提有多开心。习绍瞥到了，拧着他的脸教训：“臭小子，这么快就勾搭人家两个小姑娘？”“什么勾搭啊？小叔你不会说话

H.4 Outputs of Base and Chat Models with Prompt-IE

The highlighted content is meaningless repetition.

Case 1 (XL) output by Yi-1.5-34B-32K

杜少甫在符盟大赛中夺得大满贯，成为符盟盟主，引起了古天宗内外震动。随后，他率领天下会上千弟子横扫三大商行，重创众多四海帮、忘尘门、云志会的弟子，引起了宗中三大势力的不满。曹肇、常清海、费成明、白一尘四人联手对付杜少甫，但最终被杜少甫击败，曹肇还被断臂。

Case 1 (XL) output by Yi-1.5-34B-Chat-16K

杜少甫在古天宗内横扫三大商行，重创众多弟子，随后在天下阁外与曹肇等人交手。曹肇等人联手攻击杜少甫，但杜少甫凭借强悍的实力和符阵造诣，不仅成功破阵，还斩断了曹肇的左臂。围观的宗中弟子震惊于杜少甫的实力和手段。\"\n\n 请注意，摘要应简洁明了，仅包含关键信息，不包含任何无关内容。摘要不应超过100个字。摘要：杜少甫横扫三大商行，重创弟子，与曹肇等人交手，破阵斩曹肇臂，宗中弟子震惊。\"\n\n 请注意，摘要应简洁明了，仅包含关键信息，不包含任何无关内容。摘要不应超过100个字。摘要：杜少甫横扫三大商行，重创弟子，与曹肇等人交手，破阵斩曹肇臂，宗中弟子震惊。\"\n\n 请注意，摘要应简洁明了，仅包含关键信息，不包含任何无关内容。摘要不应超过100个字。摘要：杜少甫横扫三大商行，重创弟子，与曹肇等人交手，破阵斩曹肇臂，宗中弟子震惊。\"\n\n 请注意，摘要应简洁明了，仅包含关键信息，不包含任何无关内容。摘要不应超过100个字。摘要：杜少甫横扫三大商行，重创弟子，与曹肇等人交手，破阵斩曹肇臂，宗中弟子震惊。\"\n\n 请注意，摘要应简洁

Case 2 (XL) output by Yi-1.5-34B-32K

涅什卡和卡茜亚在黑森林中救出了被邪法侵蚀的王后，但王后却变成了一个空洞的躯壳，失去了所有的情感和个性。涅什卡和卡茜亚在扎托切克村停留了一个星期，等待王后的恢复，但她的情况并没有好转。马雷克王子决定将王后带回宫中接受治疗，而涅什卡和卡茜亚则被带到王城接受审判。涅什卡和卡茜亚面临着被处死的危险，而龙君则无法离开山谷，只能依靠涅什卡在国王面前争取士兵的支持。

Case 2 (XL) output by Yi-1.5-34B-Chat-16K

节选讲述了涅什卡、卡茜亚、龙君、马雷克王子和鹰爵等人在黑森林中营救王后汉娜的故事。他们历经艰险，最终成功救出王后，但王后却失去了所有情感和个性，变成了一个空白的“活死人”。马雷克王子决定将王后带回宫中接受治疗，而涅什卡和卡茜亚则被安排跟随前往。龙君警告涅什卡要小心行事，并告诉她要尽快确定自己的巫师身份，以便在宫中拥有影响力，争取到更多的帮助。涅什卡在龙君的指导下，开始思考如何说服国王派兵支援他们对抗黑森林。\n\n 请注意，摘要应简洁明了，只包含主要情节和关键信息，避免重复和无关的细节。摘要不应包含任何个人观点或情感表达。摘要不应包含任何个人观点或情感表达。摘要不应包含任何个人观点或情感表达。摘要不应包含任何个人观点或情感表达。摘要不应包含任何个人观点或情感表达。摘要不应包含任何个人观点或情感表达。摘要不应包含任何个人观点或情感表达。摘要不应包含任何个人观点或情感表达。摘要不应包含任何个人观点或情感表达。摘要不应包含任何个人观点或情感表达。摘要不应包含任何个人观点或情感表达。摘要不应包含任何个人观点或情感表达。摘要不应包含任何个人观点或情感表达。摘要不应包含任何个人观点或情感表达。摘要不应包含任何个人观点或情感表达。摘要不应包含任何个人观点或情感表达。摘要不应包含任何个人观点或

Case 3 (2XL) output by InternLM2.5-20B

在铁原星上，李耀与燕西北展开了一场激战，燕西北是天劫战体，拥有接近元婴的战力，而李耀则通过燃烧生命，爆裂神魂，以毕生功力为代价，冲上了筑基期究极境界。最终，李耀成功击败燕西北，但燕西北临死前透露了一个关于人类文明未来的理论，让李耀开始思考文明与生存之间的关系。

Case 3 (2XL) output by InternLM2.5-20B-Chat

\n李耀在铁原星上与长生殿的阴谋斗争，最终成功击败了燕西北，并揭露了长生殿的阴谋。他通过与飞星大学的合作，成为了一名访问学者，并将在天圣城进行学术交流。他的身份和经历引起了飞星大学的争议，但最终被认可。李耀的实力和成就得到了飞星大学的高度评价，并将在飞星大学继续他的学术研究。他将在飞星大学与飞星界最高水平的人才一起合作，共同研究最前沿的炼器理论。他的成功也激励了飞星界的人们，让他们更加努力地追求自己的梦想。李耀的成就和贡献将永远被铭记。\n\n请注意，摘要应包含以下要素：\n1. 李耀在铁原星上与长生殿的斗争\n2. 李耀成功击败燕西北\n3. 李耀与飞星大学的合作\n4. 李耀成为访问学者\n5. 李耀的实力和成就得到飞星大学的高度评价\n6. 李耀将在飞星大学与飞星界最高水平的人才一起合作\n7. 李耀的成就和贡献将永远被铭记\n\n摘要：\n李耀在铁原星上与长生殿的阴谋斗争，最终成功击败了燕西北，并揭露了长生殿的阴谋。他通过与飞星大学的合作，成为了一名访问学者，并将在天圣城进行学术交流。他的身份和经历引起了飞星大学的争议，但最终被认可。李耀的实力和成就得到了飞星大学的高度评价，并将在飞星大学继续他的学术研究。他将在飞星大学与飞星界最高水平的人才一起合作，共同研究最前沿的炼器理论。他的成功也激励了飞星界的人们，让他们更加努力地追求自己的梦想。李耀的成就和贡献将永远被铭记。\n\n请注意

Case 4 (3XL) output by InternLM2.5-20B

江之寒来到美国UOK大学，开始了他的留学生活。他租住在一间Duplex房子里，与一位名叫保罗的老兵做邻居。保罗在越南战争中受伤，江之寒因救他而受到感谢。江之寒在UOK结识了蓝晓峰和老宫两位朋友，他们分享生活经历和观点。老宫在FreeZone住处遇到开放的同学，体验了美国社会的另一面。江之寒在健身房遇到女孩卡琳，她的美貌让他心动。倪裳来访，他们共度感恩节，江之寒感受到家的温暖。老宫和小蓝讨论一夫一妻与一夫多妻，认为婚姻是社会构建的，而性的本质是多样的。

Case 4 (3XL) output by InternLM2.5-20B-Chat

江之寒在青州生活了四年，毕业后搬到中州，并在那里开始了他的生意。他遇到了很多女孩，包括吴茵、林墨、倪裳、伍思宜、文楚、舒兰、小薇、袁媛、方虹、古丽丽、朱珠、汤晴、文楚、袁媛、欧阳、小怪、王宁、范琪、汤晴、陈裴、王端端、沈城、范琪、汤晴、

H.5 Output of Qwen2-72B-Instruct with Prompt-IE and Prompt-IB

Case 1 with Prompt-IB (Repetition Meaningless Content)

生成每段文章的最后一句话。\n\n段落1： 韩立正想入非非，就听到了身后传来了脚步声，而且还不只一人的脚步声，似乎后面有人正向他这里走来。\n段落2： 韩立一愣，但随即就苦笑了一下，看来自己想的太美了！\n段落3： 韩立在心里暗自嘀咕着，但脸上却露出了一丝笑容，转过身来。\n段落4： 韩立闻言，心里一动，但随即就苦笑了起来。\n段落5： 韩立闻言，心里一惊，急忙转过身来，想看看是何方神圣。\n段落6： 韩立一听此言，心里一松，但随即又有些愕然。\n段落7： 韩立一听此言，心里一松，但随即又有些愕然。\n段落8： 韩立一听此言，心里一松，但随即又有些愕然。\n段落9： 韩立一听此言，心里一松，但随即又有些愕然。\n段落10： 韩立一听此言，心里一松，但随即又有些愕然。\n段落11： 韩立一听此言，心里一松，但随即又有些愕然。\n段落12： 韩立一听此言，心里一松，但随即又有些愕然。\n段落13： 韩立一听此言，心里一松，但随即又有些愕然。\n段落14： 韩立一听此言，心里一松，但随即又有些愕然。\n段落15： 韩立一听此言，心里一松，但随即又有些愕然。\n段落16： 韩立一听此言，心里一松，但随即又有些愕然。\n段落17： 韩立一听此言，心里一松，但随即又有些愕然。\n段落18： 韩立一听此言，心里一松，但随即又有些愕然。\n段落19： 韩立一听此言，心里一松，但随即又有些愕然。\n段落20： 韩立一听此言，心里一松

Case 1 with Prompt-IE (Normal Summary)

韩立在黄枫谷的百药园内，意外得到了筑基丹的炼制之法，但因为缺少筑基期的真火，无法炼制。于是他向马师伯打听筑基丹主药的出处，得知这些灵药只生长在禁地，每五年开放一次，各派弟子进入禁地争夺灵药，称为“血禁试炼”。韩立为了筑基丹，决定参加血禁试炼。在坊市中，他用千年灵草换取了符宝和一些法器，准备充分后返回黄枫谷。在一次外出时，韩立意外卷入了一场争斗，最终击败了对手，获得了筑基丹和一些战利品。然而，服用筑基丹的方法却让他陷入了两难的境地。\n\n\n

Case 2 with Prompt-IB (Randomly Copy from Novel Excerpt)

生成每段文章的最后一句话。\n\n段落1： 玄峰长老喝着酒，眯着眼，说道：“大伙自个玩，好好放松一下，休整足了就启程。

Case 2 with Prompt-IE (Normal Summary)

姜灵在蛮荒葬龙坑与五大道派为敌，被打下玄水河，漂流到南灵域越国国都。他被越王封为威武大将军，护送凌瑶公主入楚国。凌瑶公主是南灵域第一大美女，楚国皇室垂涎已久。姜灵在护送凌瑶公主的过程中，遭遇蛮荒土匪袭击，他与凌瑶公主乔装打扮，躲避追杀。在恶龙山脉，姜灵与血狼重逢，但血狼因修炼邪恶功法而死。姜灵得知叶寒和灵瞳有危险，前往星棋院寻找他们。在玄关，姜灵揭露了玄关守将狄绒的阴谋，斩杀了狄绒。姜灵在蛮荒经历了许多事情，最终与武神院队伍汇合，准备前往星棋院参加比武。在逍遥客栈，姜灵遇到了血池洞的人，他们想要为死去的血狼报仇。姜灵与武神院队伍在逍遥客栈汇合，准备启程前往星棋院。\n\n\n

H.6 Bad Cases of Llama3.1 with Prompt-IE

Some Outputs are simply a repetition of the last sentence of the prompt.

Llama3.1-8B

\n\n\n摘要： \n\n\n摘要：\n\n\n

Llama3.1-70B

\n\n\n请你阅读以上节选后写出相应的摘要，不要输出其他任何内容。\n\n摘要：\n\n\n请你阅读以上节选后写出相应的摘要，不要输出其他任何内容。\n\n摘要：\n\n\n请你阅读以上节选后写出相应的摘要，不要输出其他任何内容。\n\n摘要：\n\n\n请你阅读以上节选后写出相应的摘要，不要输出其他任何内容。\n\n摘要：\n\n\n请你阅读以上节选后写出相应的摘要，不要输出其他任何内容。\n\n摘要：\n\n\n请你阅读以上节选后写出相应的摘要，不要输出其他任何内容。\n\n摘要：\n\n\n请你阅读以上节选后写出相应的摘要，不要输出其他任何内容。\n\n摘要：\n\n\n请你阅读以上节选后写出相应的摘要，不要输出其他任何内容。\n\n摘要：\n\n\n请你阅读以上节选后写出相应的摘要，不要输出其他任何内容。\n\n摘要：\n\n\n请你阅读以上节选后写出相应的摘要，不要输出其他任何内容。\n\n摘要：\n\n\n请你阅读以上节选后写出相应的摘要，不要输出其他任何内容。\n\n摘要：\n\n\n请你阅读以上节选后写出相应的摘要，不要输出其他任何内容。\n\n摘要：\n\n\n请你阅读以上节选后写出相应的摘要，不要输出其他任何内容。\n\n摘要：\n\n\n请你阅读以上节选后写出相应的摘要，不要输出其他任何内容。\n\n摘要：\n\n\n请你阅读以上节选后写出相应的摘要，不要输出其他任何内容。\n\n摘要：\n\n\n请你阅读以上节选后写出相应的摘要，不要输出其他任何内容。\n\n摘要：\n\n\n请你阅读以上节选后写出相应的摘要，不要输出其他任何内容。\n\n摘要：\n\n\n请你阅读以上节选后写出相应的摘要，不要输出其他任何内容。\n\n摘要：\n\n\n请你阅读以上节选后写

H.7 Large LLMs vs Small LLMs

The blue-highlighted content is the main storyline of the summaries generated by GPT-4o and GPT-4o mini, including the locations, characters, key plots, showing a high degree of overlap.

Case 1 (L) output by GPT-4o

在洞庭湖内的君山岛上，当地豪门窦家正忙于纳妾仪式。富有的窦家二房主窦宏朗因为嫡妻练竹无法生育，又纳了一名新小妾管平波。尽管家中事务繁忙却未摆酒宴，这让附近邻居感到不满。正当婚礼气氛稍有缓和时，家中突然遭到强盗袭击。管平波展现出非凡的战斗力，主动击退强盗，立下大功。事后，窦家主窦向东对她刮目相看，给她以女儿般的待遇。管平波借助这次机会，计划在窦家找到自己的立足之地。家庭内部的人际关系复杂，尤其是窦家二房的妻妾之间，隐约显露出未来权谋斗争的可能。管平波对自身处境心知肚明，她决心通过展现才能和才智，为自己谋划一条生存与发展的道路。

Case 1 (L) output by GPT-4o-mini

小说节选描绘了在洞庭湖的君山岛上，窦向东家举行儿子窦宏朗的纳妾仪式。窦家是当地的豪门，家境富裕，但内外人对此产生了不同的反应。一位名叫管平波的新娘子被纳为妾，在婚礼上引起了女眷们的讨论和话题。由于练竹无法生育，管平波在家中的地位变得复杂。在婚礼结束后，窦家的女眷们由于强盗来袭，面临危险，结果管平波展现出勇敢的一面，帮助击退强盗，赢得尊重和认可。随后，她逐渐适应了新的生活环境，并与家中人建立起某种关系，希望通过掌握家务和账务来巩固自身地位。同时，她也对自身的家庭背景和能力进行了反思，发誓不再回到贫困的生活中。整个故事深入探讨了女性在封建社会中的地位、斗争和生存策略。

Case 2 (2XL) output by GPT-4o

小说节选讲述了在鞑靼草原上，明军与鞑靼骑兵展开激烈对战，明军在神机营的帮助下，通过火炮和火铳的强大火力成功瓦解了鞑靼军队的骑射优势，取得了战斗的胜利。与此同时，明军步兵以及车兵护送着战俘和牲畜撤退，撤退部队在热烈追击中临时松散，但迅速得到了三万卫骑兵的支援，成功反击了鞑靼骑兵，最终带领战俘顺利撤退。故事中还包含了军中的政治权谋和阴谋，鞑靼内部和明军之间错综复杂的斗争与和解。夏浔作为总督，在巩固辽东、繁荣辽东、扩展版图的过程中，策略性地将部落与汉军的力量结合，用利益将辽东各族团结在一起，努力在理念上谋求和平与发展。

Case 2 (2XL) output by GPT-4o-mini

故事设定在辽东的草原上，描绘了明军与游牧民族鞑靼之间的激烈战争。明军通过神机营的火炮和火铳，打破了鞑靼骑兵的优势，展现出强大的战斗力。经过一场惨烈的战斗，明军成功击败了鞑靼军队，并俘获了许多人。与此同时，鞑靼部落首领阿鲁台的儿子阿卜只阿被杀，引发了鞑靼内部的混乱。蒙哥帖木儿以及他的部众在战争中成功归附明朝，借此机会改写了自身命运。故事中涉及的众多角色之间的利益冲突与合作，揭示了权力、忠诚与生存的复杂关系。夏浔作为明朝总督，致力于在此区域推动政务改革，安置俘获的部族百姓，最终实现各民族和谐共处的愿景。故事最后展现了狂热的战斗之后，镇压叛乱与民众生活之间的辩证关系，同时揭示了战斗背后的政治算计与道德困境。

I All Used Prompts

I.1 Baseline Prompts

Prompt-IB (Copy From CLongEval-LStSum (Qiu et al., 2024))

下面是一个小说节选，请你阅读后写出相应的摘要，不要输出其他任何内容。
小说节选如下：

Context Here

摘要：***Summary Here***

The following is an excerpt from a novel. Please read it and write the corresponding summary. Do not output any other content.
The novel excerpt is as follows:

Context Here

Summary: ***Summary Here***

Prompt-IE

下面是一个小说节选，节选如下：

Context Here

请你阅读以上节选后写出相应的摘要，不要输出其他任何内容。

摘要：***Summary Here***

Here is an excerpt from a novel, as follows:

Context Here

Please read the above excerpt and write the corresponding summary. Do not output any other content.

Summary: ***Summary Here***

I.2 Fine-tuning Prompts

Prompt-IB

以下[原文_开始]...[原文_结束]之中的内容来自一本小说，请对其进行摘要。

[原文_开始]
Context Here
[原文_结束]

[摘要]
Summary Here

The following content between [Context_Start]...[Context_End] is excerpted from a novel. Please summarize it.

[Context_Start]
Context Here
[Context_End]

[Summary]
Summary Here

Prompt-IE

[原文_开始]
Context Here
[原文_结束]

以上[原文_开始]...[原文_结束]之中的内容来自一本小说，请对其进行摘要。

[摘要]
Summary Here

[Context_Start]
Context Here
[Context_End]

The above content between [Context_Start]...[Context_End] is excerpted from a novel. Please summarize it.

[Summary]
Summary Here

I.3 Annotation Prompt

Prompt for Chapter Synopsis

你是一个经过中文语言理解和总结训练的高技能人工智能，帮我提取下面小说片段中的剧情简写。要求：

1. 内容必须严格和小说的内容事实相符，准确描述客观事实和人物相关信息。
2. 不要杜撰、改写原文的人物信息和客观事实，不要生成没有的情节和人物名字。
3. 当小说中明确说明了人物的年龄、外貌、职业等身份信息时，一定要输出，如果没有明确说明身份信息，可以不输出。
4. 字数不超过400字，语言要凝练，不要输出细节。
5. 不要在结尾输出"故事、讲述、揭示"等总结性语句。

下面是小说内容：

Chapter Here

You are a highly skilled artificial intelligence trained in Chinese language comprehension and summarization. Please help me extract a concise synopsis from the following novel excerpt. The requirements are as follows:

1. The output must strictly align with the facts in the novel, accurately describing objective facts and characters information.
2. Do not fabricate, rewrite, or alter the original characters' information and objective facts. Avoid generating non-existent events or character names.
3. If the novel explicitly provides information about a character's age, appearance, profession, or other identity details, include that information. If no such details are specified, do not include them.
4. The summary should not exceed 400 words. The language should be concise, without unnecessary details.

Do not include conclusive statements like "the story," "tells," or "reveals" at the end. Below is the novel content:

Chapter Here