XL²Bench: A Benchmark for Extremely Long Context Understanding with Long-range Dependencies

Anonymous ACL submission

Abstract

Recently, various efforts have been proposed to expand the context window size of large language models (LLMs). Meanwhile, building high-quality benchmarks with much longer text lengths and more demanding tasks to provide comprehensive evaluations is of immense practical interest to facilitate long context understanding research of LLMs. However, prior benchmarks create datasets that ostensibly cater to long-text comprehension by expanding the input of traditional tasks, which falls short to exhibit the unique characteristics of long-text understanding, including long dependency tasks and longer text length compatible with modern LLMs' context window size. In this paper, we introduce a benchmark for e**X**tremely **L**ong context understanding with **L**ong-range dependencies, **XL²Bench**, which includes three scenarios—Fiction Reading, Paper Reading, and Law Reading—and four tasks of increasing complexity: Memory Retrieval, Detailed Understanding, Overall Understanding, and Open-ended Generation, covering 27 subtasks in English and Chinese. It has an average length of 100K+ words (English) and 200K+ characters (Chinese). Evaluating seven leading LLMs on XL²Bench, we find that their performance significantly lags behind human levels. Moreover, the observed decline in performance across both the original and enhanced datasets underscores the efficacy of our approach to mitigating data contamination.

1 Introduction

Large Language Models (LLMs) have attracted considerable interest for their remarkable capabilities in a wide range of NLP tasks. However, a common limitation among these models is the fixed context window size (for example, LLaMA with maximum 2048 tokens and GPT-3.5 with maximum 4096 tokens), rendering them incapable of memorizing and understanding extremely long inputs (Liu et al., 2023). Evidenced by a basic passkey retrieval task, the accuracy of LLaMA recalling a passkey plummets from nearly 100% to nil when the text surpasses 2048 tokens (Tworkowski et al., 2023).

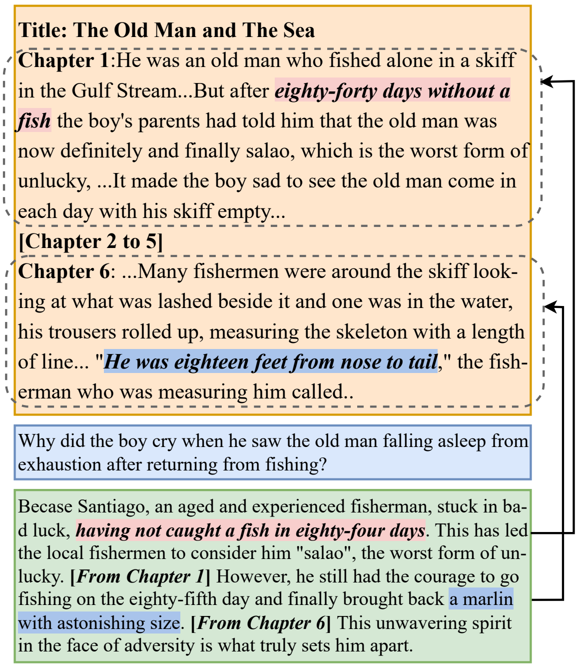

Figure 1: An illustrative example of long-dependency task, in which the model needs to make connective inferences across input document to fulfill the goal.

In pursuit of the goal of improving LLM's ability to comprehend long-context textual information, various efforts have been proposed to expand the context window of LLMs, such as sparse attention (Tworkowski et al., 2023; Chen et al., 2023; Mohtashami and Jaggi, 2023), length extrapolation (Dai et al., 2019; Su et al., 2021; Peng et al., 2023), and context compression (Ge et al., 2023; Mu et al., 2023). Given the notable advances achieved by these techniques, the necessity for high-quality benchmarks, featuring longer text lengths and more complex tasks, is escalating to facilitate thorough evaluations of LLMs' long context understanding ability.

Being able to understand long-range dependencies in context and be sensitive to various perturbations applied to distant context is what sets long text understanding apart from traditional NLP tasks (Wang et al., 2020; Tay et al., 2021; Rae and Razavi, 2020; Ni et al., 2023). Existing benchmarks, such as LongBench (Bai et al., 2023), L-Eval (An et al., 2023), M^{4}LE (Kwan et al., 2023), and InfiniteBench (Zhang et al., 2023b), often merely expand the input of traditional tasks, such as concatenating short texts to get long texts, to create datasets that ostensibly cater to long-text comprehension (Bai et al., 2023; An et al., 2023). However, this approach does not tailor tasks to the distinct features of long text comprehension, thereby impeding the thorough assessment of LLMs. Moreover, the average text length in existing benchmarks, such as LooGLE (Li et al., 2023), usually does not exceed a few thousand tokens, significantly shorter than the long texts perceived in human cognition. For example, a user might upload an entire novel and inquire about the development of the protagonist's storyline. This task would require the model to process and comprehend texts spanning over ten thousands of words, necessitating long-range understanding and reasoning within the content to adequately address the question. Traditional benchmarks typically fall short in measuring capabilities of LLMs to aggregate disparate pieces of information scattered throughout the whole input texts in more realistic scenarios, making it challenging to truly evaluate LLMs' ability on long context understanding (Dong et al., 2023; Kwan et al., 2023).

In light of the deficiencies identified in current benchmarks, this paper proposes a benchmark for e**X**tremely **L**ong context understanding with **L**ong-range dependencies, **XL^{2}Bench**, which features three scenarios——Fiction Reading, Paper Reading, and Law Reading. XL^{2}Bench contains extremely long documents with an average of 100K+ words (English) and 200K+ characters (Chinese), along with 632K questions spanning over four specifically designed tasks to examine a model's ability to aggregate and compare information across long context, including *Memory Retrieval*, *Detailed Understanding*, *Overall Understanding*, and *Open-ended Generation*. These tasks mimic the way people use LLMs in real-world scenarios. Figure 1 illustrates a case in XL^{2}Bench where explaining a boy's tears as stemming from a story about the old man who, against significant challenges, successfully captures a marlin. To construct a solid answer, it demands the model to identifies passages describing the boy's reaction, the man's triumph, and his earlier hardships across various chapters, and make connective inferences using details buried far back in the long context.

Besides, to address data contamination caused by outdated long texts contained in benchmark, we implement three data augmentation strategies: **text transformation**, which involves altering the original text into a different language or style; **text replacement**, which entails modifying or substituting key textual information; and **text concatenation**, which incorporates integrating additional texts into the original document.

Results of experiments on multiple state-of-the-art LLMs reveal that even the most advanced LLMs currently available fall short of reaching human-level proficiency on XL^{2}Bench. Despite these models' ability to handle texts of considerable length, there is a marked decline in performance as the text lengthens. Additionally, the results obtained by RAG (Li et al., 2022; Gao et al., 2023) on XL^{2}Bench demonstrate that retrieval-based methods fail in overall and detailed understanding tasks; instead, they require that the models comprehensively grasp the entirety of the long texts. Furthermore, we conduct ablation experiments to compare model performance on both original and augmented benchmarks, which shows that the strategies we employ to address the issue of data contamination are indeed effective.

Our contributions are delineated as follows:

- We construct XL^{2}Bench, a comprehensive benchmark for extremely long text understanding with well-designed tasks.
- We formulate three data augmentation techniques to circumvent the issue of data contamination. Through experimentation, we validate the efficacy of these methodologies in mitigating concerns about data contamination.
- We conduct empirical experiments to evaluate the performance of advanced LLMs using XL^{2}Bench. The results reveal that contemporary LLMs are still facing challenges in achieving comprehensive understanding across long textual inputs.

Tasks	Subtasks	Source	Num		Avg. Len		Metric
			CN	EN	CN	EN	
			Fiction Reading				
Memory Retrieval	Content Location	Content Extraction	1495	1405	571.6K	111.5K	Acc.
	Content Retrieval	Content Extraction	299	261	571.1K	116.0K	Acc.
Detailed Understanding	Chapter Summarization	Data Synthesis	167	156	569.7K	110.6K	Rouge-L
	Question Answering	Data Synthesis	249	269	562.0K	114.7K	BLEU
Overall Understanding	Chapter Counting	Content Extraction	30	27	569.7K	113.4K	Acc.
	Background Summarization	Data Synthesis	30	27	570.3K	113.7K	Rouge-L
	Event Extraction	Data Synthesis	30	27	570.2K	113.7K	Rouge-L
	Fiction Summarization	Data Synthesis	30	27	570.4K	113.8K	Rouge-L
	Character Description	Data Synthesis	191	140	589.7K	143.5K	Rouge-L
	Relationship Analysis	Data Synthesis	193	432	606.3K	189.8K	Rouge-L
Open-ended Generation	Role-play Conversation	Data Synthesis	293	256	592.7K	115.2K	BLEU
	News Generation	Data Synthesis	30	27	570.7K	114.0K	BLEU
	Poem Generation	Data Synthesis	30	27	570.1K	113.6K	BLEU
			Paper Reading				
Memory Retrieval	Content Retrieval	Content Extraction	-	4532	-	13.7K	Acc.
Detailed Understanding	Section Summarization	Data Synthesis	-	3136	-	14.1K	Rouge-L
	Terminology Explanation	Data Synthesis	-	14981	-	13.5K	BLEU
Overall Understanding	Paper Counting	Content Extraction	-	3100	-	13.5K	Acc.
	Paper Summarization	Data Integration	-	518	-	14.0K	Rouge-L
Open-ended Generation	Paper Review	Data Integration	-	518	-	14.0K	BLEU
	Rating Score	Data Integration	-	518	-	13.6K	MAE
			Law Reading				
Memory Retrieval	Legal Entry Location	Content Extraction	2213	-	105.6K	-	Acc.
	Legal Entry Retrieval	Content Extraction	2225	-	105.3K	-	Acc.
Detailed Understanding	Legal Definition QA	Data Synthesis	2635	-	102.9K	-	BLEU
	Legal Number QA	Data Synthesis	1477	-	105.7K	-	Acc.
Overall Understanding	Legal Entry Counting	Content Extraction	122	-	103.0K	-	Acc.
	Multiple Choice QA	Data Integration	16881	-	95.6K	-	F1
Open-ended Generation	Case Adjudication	Data Integration	588369	-	72.7K	-	Acc.

Table 1: An overview of the statistics of XL^{2}Bench. **Source** represents the method we use to construct the dataset for this subtask. **Num** represents the number of <input, output> pairs this subtask possesses. **Avg. Len** denotes the average combined length of the input and output, which is computed using the number of characters for Chinese and the number of words for English. **K** stands for 1024. For example, 200K = 200*1024.

2 Methodology

In this section, we introduce the construction methodologies of XL^{2}Bench and design of tasks with various level of difficulty.

2.1 Task Design

We evaluate the model's understanding of extremely long texts from the perspectives of fine-grained retrieval and coarse-grained understanding. Based on this, we design four tasks: *Memory Retrieval*, *Detailed Understanding*, *Overall Understanding*, and *Open-ended Generation*.

Memory Retrieval. This task challenges the model to accurately retrieve and respond to queries by finding content within the text that aligns with given instructions. For instance, the model may be asked to pinpoint the specifics of a legal entry within a law or identify the originating chapter of a passage from a novel, thereby evaluating its capability to accurately locate and interpret question-relevant content.

Detailed Understanding. Here, the model is tasked with not only retrieving content but also comprehensively understanding it to perform activities such as summarization or question answering. This demands a more profound level of textual comprehension, surpassing mere content retrieval to include an in-depth analysis and synthesis of the text.

Overall Understanding. To circumvent tasks being completed through simple content retrieval, we introduce the Overall Understanding task. This task necessitates a holistic comprehension of the long text, compelling the model to build long-range

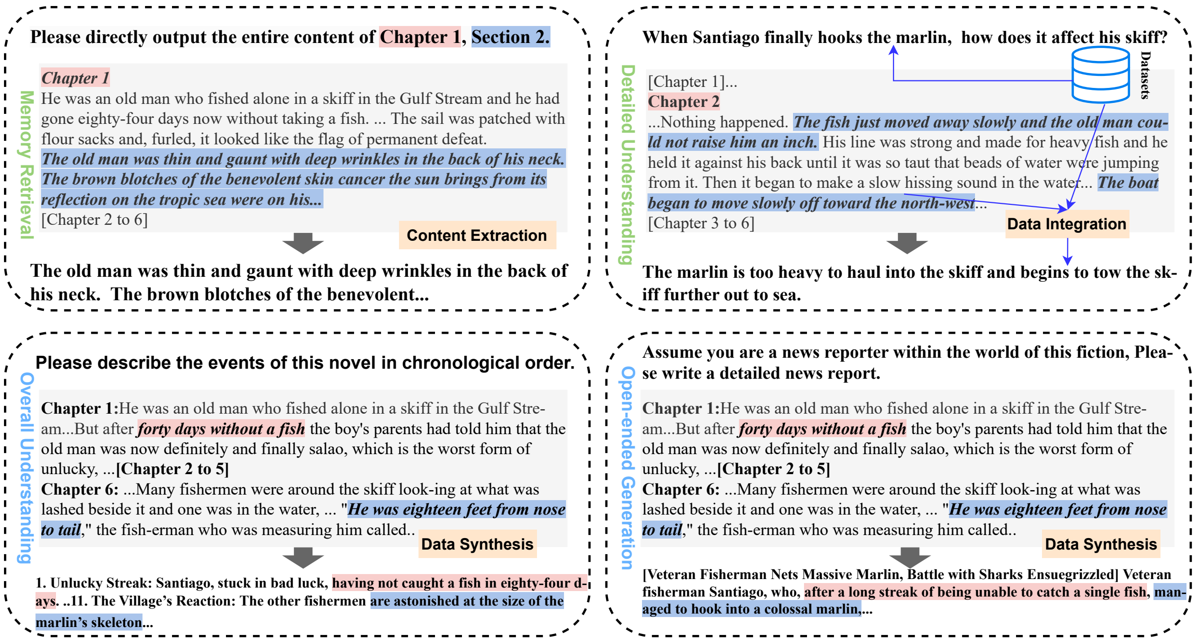

Figure 2: Illustration of the designed long context understanding tasks.

dependencies and tackle inquiries related to overarching themes, such as the depiction of a character throughout a novel or the trajectory of a company's stock across its history.

Open-ended Generation. Building on a robust foundation of long text comprehension, the model is tasked with undertaking generation tasks that are deeply rooted in the text, such as role-playing a fiction character. The outputs should exhibit creative expansion and inference, remaining faithful to the core themes and concepts of the text, while also ensuring originality and thematic consistency.

Figure 2 provides 4 examples for each task, demonstrating the characteristics of the tasks within XL^{2}Bench, as well as the capabilities required for a model to successfully complete these tasks.

2.2 Benchmark Construction

In this subsection, we describe the sources from which we gather data and the methodologies we employ for constructing the benchmark.

We gather long texts categorized under three scenarios. For fiction reading, we select a variety of novels written in both Chinese and English. For paper reading, we download PDF versions and reviews of papers submitted to ICLR 2023 from Openreview[1]. For law reading, we gather a substantial collection of original Chinese legislations.

To minimize cost of human annotation, we employ three methods to construct: *Content Extraction*, *Data Integration,* and *Data Synthesis*.

Content Extraction. We extract content from the original text to serve as the answer and use the index of this portion of the content to formulate the question[2]. This method does not necessitate the involvement of LLMs and solely relies on string processing, which is suitable for tasks that have direct and fixed answers, such as Memory Retrieval.

Data Integration. Tasks within certain short text datasets bear formal resemblance to what we have designed. Consequently, we contemplate leveraging these datasets to augment our benchmark. More precisely, we employ LLMs to facilitate the alignment of data from the pre-existing datasets with our collected long texts, transforming the format from <Input, Output> into <Text, Input, Output>. In an effort to reduce the model's familiarity with these datasets, we remove any information that may indicate the data source.

Data Synthesis. In the remaining tasks, we utilize LLMs for direct generation. For summarization tasks, we implement a structured text summarization approach (Chang et al., 2023). For QA tasks, we apply in-context learning techniques (Brown et al., 2020) to create example-based prompts that facilitate model generation.

[1] https://openreview.net/group?id=ICLR.cc/2023/Conference

[2] For instance, we use the title of a paper as the answer, with the corresponding question being: *What is the title of this paper?*

2.3 Human Verification

Data Synthesis inherently limits our benchmark to the quality of the content produced by LLMs. However, it is important to **note** that XL^{2}Bench is not solely comprised of LLM-generated questions and answers, as these constitute no more than **30%** of the benchmark. For the portion generated by LLMs, we implement a meticulous human verification process to ensure the quality of the questions and answers. This verification process involves: (1) We initially rule out content in the model's response that is irrelevant to the text, such as phrases like "*Sure!*", "*Here are the answers.*", etc. Next, we report inconsistencies between the response and the text, such as erroneous summaries. The LLM is then prompted to regenerate the content. If it still cannot produce the correct answers, human annotations are made.

Employing the aforementioned approaches, we have constructed an extremely long text benchmark encompassing three distinct scenarios, four overarching tasks, 27 detailed subtasks, and a corpus of 700+ texts with a average length of 100K+ words for English and 200K+ characters for Chinese. The statistics of our benchmark are shown in Table 1. For more task descriptions and the input and output templates of XL^{2}Bench, please refer to Appendix B.

2.4 Data Contamination

The potential of data contamination warrants serious consideration when constructing a benchmark (Sainz et al., 2023; Deng et al., 2023; Magar and Schwartz, 2022). The risk arises when the test set data is either identical to, or strikingly similar to, the training set data. In our construction process, the selected novels, academic papers, and legal texts may have been included in the training corpus of LLMs. Consequently, the model may not need to fully comprehend the entire text to accomplish various tasks. In order to mitigate the impact of data contamination on model's performance, we follow Yang et al. (2023) and adopt three strategies, namely *text transformation*, *key information replacement*, and *text concatenation* for fiction data augmentation.

Text Transformation. We utilize LLMs to facilitate mutual translation of fictions between Chinese and English, whereby the original Chinese (English) novels are rendered into English (Chinese). In accordance, the input and output for each task are also translated.

Key Information Replacement. We employ LLMs to extract key information from a chapter or section, such as names, places, and times. We then generate corresponding texts to replace these elements, resulting in a collection of ⟨original text - replacement text⟩ pairs, which are subsequently used for content substitution throughout the entire text and tasks.

Text Concatenation. We insert a short story into the original fiction as one of its chapters, and use this template to bridge: *Now, let's pause the current story narration and turn to a new story*[New Story]*The story is over, let's get back to the original fiction*. Then, we merge the data in four tasks of this short story with the original fiction.

Through above three strategies, we construct **Fiction-T** (Translated), **Fiction-R** (Replaced), and **Fiction-C** (Concatenated).

2.5 Implementation Details

We select GPT-4-Turbo (Achiam et al., 2023) to help us construct XL^{2}Bench. GPT-4 currently stands as the highest-performing LLMs, characterized by a 128k context window along with superior memory, reasoning, and generation capabilities. The prompts and input templates used throughout the construction process are available in our GitHub repository due to space limit.

3 Experimental Settings

3.1 Generative Large Language Models

We introduce current LLMs with context window size **more than 100k** evaluated in our experiments. Models such as LLama2 (Touvron et al., 2023b) and ChatGLM2 (Zeng et al., 2023) have context window size significantly shorter than the average text length of XL^{2}Bench, resulting in an excessive need to truncate texts, which leads to suboptimal performance. Consequently, we do not evaluate the effectiveness of these models.

Closed-source LLMs. Developed by OpenAI, **GPT-4-Turbo**[3] represents the pinnacle of current advancements, demonstrating exceptional reasoning and instruction-following capacities. It is distinguished by its extensive context window of 128K tokens. **GLM-4**[4] is the latest model developed

[3] https://chat.openai.com/

[4] https://open.bigmodel.cn/

Models	MR		DU		OU						TG		
	C-L	*C-R*	*C-S*	*QA*	*C-C*	*B-S*	*E-E*	*F-S*	*Ch-D*	*Re-A*	*RP-C*	*N-G*	*P-G*
YaRN-Mistral-7B	<1	<1	4.46	2.26	13.78	8.09	16.17	5.52	8.35	7.91	7.28	4.42	5.91
InternLM2-C-7B	<1	<1	8.27	<1	6.67	11.68	9.97	11.97	6.92	2.22	1.16	5.88	3.49
InternLM2-C-20B	<1	6.85	17.22	9.82	53.33	15.58	18.61	17.29	21.98	28.92	11.65	16.67	10.09
Moonshot-V1	17.23	60.39	**23.53**	33.13	**86.30**	**24.32**	20.08	**25.10**	22.24	**54.99**	12.81	27.31	12.22
GLM-4	**20.08**	**63.44**	18.12	14.51	72.73	18.40	20.42	15.84	22.22	42.27	13.62	19.70	11.69
GPT-4-Turbo	11.89	54.36	19.87	**37.23**	60.00	21.21	**21.40**	21.57	**23.14**	49.05	**17.58**	**30.19**	**16.56**
Qwen-Long-1M	8.67	56.85	16.19	17.78	30.00	22.43	18.49	17.09	21.23	36.13	15.33	14.20	13.09

Table 2: Results (%) of seven LLMs on Chinese Fiction Reading. **MR**, **DU**, **OU**, **TG** are the abbreviations for the initials of four tasks. **C-L**, **C-R**, **C-S**, etc., represent the abbreviations of 13 subtasks. The context window size of GLM-4 and InternLM2-Chat is 200K, whereas it is 128K for other models. The **bold** numbers in the results represent the best scores, whereas the underlined numbers indicate the second-best scores.

Models	Tasks				
	C-S	*QA*	*B-S*	*Re-A*	*N-G*
YaRN-Mistral-7B	6.20	6.00	6.11	6.00	6.22
InternLM2-Chat-7B	6.11	6.00	6.58	6.00	6.37
InternLM2-Chat-20B	3.87	2.23	4.37	4.20	3.41
Moonshot-V1-128K	3.01	**1.03**	2.21	2.31	2.02
GLM-4	3.29	1.11	2.25	**2.16**	2.19
GPT-4-Turbo	**2.17**	1.06	**1.10**	2.23	**1.89**
Qwen-Long-1M	2.89	1.20	2.82	2.40	2.64

Table 3: Human evaluation results of seven LLMs on five tasks of Chinese Fiction Reading. The abbreviations in the table are consistent with those in the preceding tables. The numbers in the table represent average rankings, with lower values indicating better performance.

by Zhipu AI. Compared to ChatGLM2, it boasts more powerful question-answering and text generation capabilities with 200K tokens context window size. Developed by Moonshot AI, **Moonshot-V1**[5] boasts exceptional performance in processing extremely-long text inputs of up to 128K tokens. **Qwen-Long**[6] is a large-scale language model developed by Alibaba Cloud, designed to support long contexts and multiple documents understanding over 1M tokens across various scenarios at a very low cost.

Open-source LLMs Equipped with 200K context window size, **InternLM2** exhibits comprehensive enhancements across all functionalities. We employ InternLM2-Chat-7B-200k and InternLM2-Chat-20B-200k. The computationally efficient length extrapolation technology **YaRN** makes it possible to expand LLM's context window size while conserving resources. We leverage YaRN-Mistral-7B-128k.

[5]https://www.moonshot.cn/

[6]https://bailian.console.aliyun.com/

3.2 Retrieval-Augmented Generation Methods

One type of methods to handle long texts with small context window size in LLMs is Retrieval-Augmented Generation (RAG) (Li et al., 2022). We test this technique's impact on LLMs evaluation results, to see if the model could complete XL^{2}Bench tasks by retrieving certain fixed chunks. We employ LangChain[7] and three retrievers: Sentence-Transformers (Reimers and Gurevych, 2020), LLM-Embedder (Zhang et al., 2023a), and Contriver (Izacard et al., 2022). We set the chunk size to 500 and Top-5 chunks for generation.

3.3 Dataset

Due to the substantial costs associated with evaluating LLMs on the complete XL^{2}Bench, we opt to create a test set for our experiments. We randomly select **150** samples from each subtask in the benchmark. This approach yield a representative subset of XL^{2}Bench, which we utilize to assess all models, thereby ensuring objective and equitable outcomes.

3.4 Evaluation Metrics

We carry out both automatic and human evaluations. The metrics for automatic evaluation of each task are presented in Table 1. Owing to space limitations, detailed descriptions of these metrics, as well as those for human evaluation, are included in Appendix C.

3.5 Inference Settings

We conduct the evaluation in a zero-shot setting. The input templates we use during inference can

[7]https://python.langchain.com/docs/get_started/introduction

Models	MR		DU		OU		TG
	LE-L	*LE-R*	*Def-QA*	*Num-QA*	*LE-C*	*MCQA*	*Case-Adj*
YaRN-Mistral-7B-128K	11.29	<1	8.62	<1	3.36	<1	<1
InternLM2-Chat-7B-200K	2.61	<1	3.52	<1	<1	<1	<1
InternLM2-Chat-20B-200K	22.60	5.41	40.57	58.03	11.76	44.23	41.05
Moonshot-V1-128K	**88.83**	**32.61**	**48.08**	63.85	28.10	63.11	47.40
GLM-4-200K	72.76	16.97	43.17	67.63	**31.14**	53.56	47.31
GPT-4-Turbo-128K	63.48	13.41	40.26	62.50	29.51	63.24	**48.89**
Qwen-Long-1M	80.67	10.67	46.10	**74.65**	10.00	**72.88**	46.20

Table 4: Results (%) of seven LLMs on Law Reading. **LE-L**, **LE-R**, **Def-QA**, **Num-QA**, **LE-C**, **MCQA** and **Case-Adj** represent *Legal Entry Location*, *Legal Entry Retrieval*, *Legal Definition QA*, *Legal Number QA*, *Legal Entry Counting*, *Multiple Choice QA* and *Case Adjudication*, respectively. Rest settings remain the same as in the previous tables.

Models	MR		DU		OU		TG
	LE-L	*LE-R*	*Def-QA*	*Num-QA*	*LE-C*	*MCQA*	*Case-Adj*
InternLM2-Chat-20B-200K	**5.41**	**22.60**	**40.57**	**58.03**	**11.76**	**44.23**	**41.05**
w/ Sentence-Transformers	<1	16.54	11.59	11.22	4.92	39.92	31.16
w/ LLM-Embedder	1.86	21.68	11.97	19.98	2.46	42.59	38.83
w/ Contriever	<1	16.73	10.23	5.44	4.10	40.23	37.79

Table 5: Results (%) of InternLM2-Chat-20B-200K using different embedding models on Law Reading. *w/* represents *with*. The best performance over of each subtask is in **bold**.

be found in Appendix D. When the input length exceeds the context window size of LLMs, we truncate the input sequence from the middle, as the front and end of the sequence may contain crucial information such as instructions or questions. For models that are API-callable, we follow the original settings provided in the sample code of these models. For locally deployed models, we select the decoding parameters as follows: Temperature=0.2, Top-K=40, Top-P=0.9, Repetition Penalty=1.02.

4 Results and Analysis

4.1 Long Texts Processing

The results pertaining to three scenarios are delineated in Table 2 and 4. Due to space constraints, the remaining results are relegated to Appendix E. The key findings from the experiments can be summarized below.

The overall performance of all LLMs is notably unsatisfactory. Regardless of whether they are open-source or closed-source, LLMs consistently score low across various metrics pertaining to the 27 subtasks, particularly in retrieval and counting tasks where human performance approaches 100%. We hypothesize that these results are attributable to the use of sparse attention or length extrapolation techniques within the extended model context window, as well as the truncation operation employed when the input text is too long.

Closed-source models outperform open-source models. The comparative performance analysis of three closed-source LLMs demonstrates a superior performance over their open-source counterparts. Furthermore, with 7B parameters, YaRN-Mistral and InternLM2-Chat-7B exhibit suboptimal performance across a majority of tasks, achieving scores below 1. This demonstrates the importance of the model's parameter size for effectively managing tasks in XL^{2}Bench.

LLMs have a preference for the language of the input text. GLM-4, Moonshot-V1, and Qwen-Long performs well on Chinese-language tasks (Law Reading and Fiction-CN), while GPT-4 performs well on English-language tasks (Paper Reading and Fiction-EN). We infer that this may be due to the different proportions of Chinese and English datasets used in the training process of these three models. This further indicates that the dataset is a particularly critical factor that affects model performance.

GPT-4's performance on self-generated subtasks does not meet expectations. In particular, for subtasks where the ground truth is established by GPT-4 itself, we meticulously assessed the model's efficacy. Contrary to our initial assumptions, GPT-4's scores on these tasks are lower than anticipated. Upon an in-depth analysis of the model-generated content, we hypothesized that the verbose nature of the text could have adversely affected GPT-4's understanding of the task descriptions, leading to a diminished output quality.

The findings and analyses presented above indicate that existing context window expansion technologies fall significantly short of reaching or approximating human-level performance. Addressing the issue of context dependency represents a critical area for potential breakthroughs and merits further exploration.

4.2 Human Evaluation Results

As shown in Table 3, the models of 7B size consistently occupy the bottom two rankings across all evaluated tasks. Further case analysis demonstrates that their outputs are characterized by disorganization and incoherence, often devoid of logical structure or bordering on nonsensical. In contrast, the 20B InternLM2-Chat generally achieves the fifth rank. The rankings of the remaining four LLMs, which are accessible exclusively through API calls, are tightly competitive, with GPT-4-Turbo consistently leading.

4.3 Performance of Retrieval-Augmented Generation Methods

In this subsection, we assess the performance of InternLM2-Chat-20B-200K, which utilizes three distinct retrievers on Law Reading scenarios. Results illustrated in Table 5, indicate a uniform reduction in the model's performance across all subtasks following the adoption of RAG methods. Notably, the most substantial declines in performance are observed in the Definition QA and Number QA tasks. We postulate that these decreases may be due to the retrievers' failure to recall relevant segments of text. The results and subsequent analysis imply that effectively addressing the tasks in XL^{2}Bench demands more than merely retrieving relevant documents.

4.4 Assessment of In-context Learning Ability

Previous analyses have primarily focused on the zero-shot setting. In this subsection, we evaluate the in-context learning (ICL) capabilities of LLMs on selected tasks. We utilize samples from the same long texts and tasks as prompts, transforming the input format from $\langle \text{Text}, \text{Input} \rangle$ in the zero-shot scenario to $\langle \text{Text}, \text{Prompt}_1, \ldots, \text{Prompt}_n, \text{Input} \rangle$ to assess LLM performance on the residual data. Due to input length constraints, we limit n to 5. Through in-context learning, models are capable of generating outputs that closely align with the desired format, thus elevating their scores. More details can be found in Appendix F

4.5 Impact of Context Length

In this subsection, we explore the impact of context length on the performance of LLMs. Our evaluation focuses on the average performance of the InternLM2-Chat-20B across four tasks, using legal texts of varying lengths. Results presented in Appendix G illustrate that the model's performance significantly declines with longer texts, as evidenced by a steeper curve. This observation underscores the model's challenges in effectively managing the complexities of long text modeling.

4.6 Impact of Data Contamination

In this subsection, we conduct an ablation study to examine the effectiveness of the methodologies employed to reduce data contamination. The results indicate that our data augmentation techniques can, to some extent, reduce the likelihood of biased evaluations. A detailed discussion is provided in Appendix H due to space limit.

5 Conclusion

In this paper, we present XL^{2}Bench, a comprehensive benchmark for extremely long text understanding with long-range dependencies. XL^{2}Bench consists of three scenarios, four tasks, and 27 subtasks, with an average length of over 100K words (English) and 200K characters (Chinese). We automatically construct the benchmark via LLMs, significantly reducing the cost of manually annotating the datasets. Furthermore, we mitigate data contamination risks through carefully designed techniques. Extensive experiments on XL^{2}Bench yield insights into the capabilities of current LLMs for long text understanding. We also demonstrate that RAG methods are not suitable for XL^{2}Bench as the benchmark requires a comprehensive understanding of the entire text to complete the tasks. Results and analyses indicate that XL^{2}Bench is a valuable resource for advancing research in the comprehension of long texts.

Limitations

The limitations of XL^{2}Bench mainly come from the disadvantages of using LLMs. First of all, most of the large language models that work well are not open source or free. This makes it difficult to conduct batch experiments or daily use on it. Next, a small number of open-source models require a lot of GPU resources when used, which is a difficult problem for quite many researchers, such as students.

Ethics Statement

We honor and support the ACL code of Ethics. Our bencmark XL^{2}Bench aims to evaluate large language models' ability of long-text comprehension. The interaction and assistance process do not involve any bias towards to the participants. Following our thorough examination, we can confirm that our benchmark is free from any privacy or ethical concerns.

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*.

Chenxin An, Shansan Gong, Ming Zhong, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. 2023. L-eval: Instituting standardized evaluation for long context language models. *CoRR*, abs/2307.11088.

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. 2023. Longbench: A bilingual, multitask benchmark for long context understanding. *CoRR*, abs/2308.14508.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of Bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy. Association for Computational Linguistics.

Yllias Chali and Sadid A. Hasan. 2012. On the effectiveness of using sentence compression models for query-focused multi-document summarization. In *Proceedings of COLING 2012*, pages 457–474, Mumbai, India. The COLING 2012 Organizing Committee.

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2023. Booookscore: A systematic exploration of book-length summarization in the era of llms. *arXiv preprint arXiv:2310.00785*.

Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. 2023. Longlora: Efficient fine-tuning of long-context large language models. *CoRR*, abs/2309.12307.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc Viet Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2978–2988. Association for Computational Linguistics.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. 2023. Investigating data contamination in modern benchmarks for large language models. *CoRR*, abs/2311.09783.

Zican Dong, Tianyi Tang, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. 2023. BAMBOO: A comprehensive benchmark for evaluating long text modeling capacities of large language models. *CoRR*, abs/2309.13345.

Yue Gao and Jian-Wei Liu. 2023. Adaptively sparse transformers hawkes process. *Int. J. Uncertain. Fuzziness Knowl. Based Syst.*, 31(4):669–689.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *CoRR*, abs/2312.10997.

Tao Ge, Jing Hu, Xun Wang, Si-Qing Chen, and Furu Wei. 2023. In-context autoencoder for context compression in a large language model. *CoRR*, abs/2307.06945.

Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Supryadi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, and Deyi Xiong. 2023. Evaluating large language models: A comprehensive survey. *CoRR*, abs/2310.19736.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Trans. Mach. Learn. Res.*, 2022.

Wai-Chung Kwan, Xingshan Zeng, Yufei Wang, Yusen Sun, Liangyou Li, Lifeng Shang, Qun Liu, and

Kam-Fai Wong. 2023. M4LE: A multi-ability multi-range multi-task multi-domain long-context evaluation benchmark for large language models. *CoRR*, abs/2310.19240.

Huayang Li, Yixuan Su, Deng Cai, Yan Wang, and Lemao Liu. 2022. A survey on retrieval-augmented text generation. *CoRR*, abs/2202.01110.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. 2023. Loogle: Can long-context language models understand long contexts? *CoRR*, abs/2311.04939.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *CoRR*, abs/2307.03172.

Inbal Magar and Roy Schwartz. 2022. Data contamination: From memorization to exploitation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 157–165. Association for Computational Linguistics.

Amirkeivan Mohtashami and Martin Jaggi. 2023. Landmark attention: Random-access infinite context length for transformers. *CoRR*, abs/2305.16300.

Jesse Mu, Xiang Lisa Li, and Noah D. Goodman. 2023. Learning to compress prompts with gist tokens. *CoRR*, abs/2304.08467.

Benjamin Newman, John Hewitt, Percy Liang, and Christopher D. Manning. 2020. The EOS decision and length extrapolation. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2020, Online, November 2020*, pages 276–291. Association for Computational Linguistics.

Xuanfan Ni, Hongliang Dai, Zhaochun Ren, and Piji Li. 2023. Multi-source multi-type knowledge exploration and exploitation for dialogue generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 12522–12537. Association for Computational Linguistics.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2023. Yarn: Efficient context window extension of large language models. *CoRR*, abs/2309.00071.

Ofir Press, Noah A. Smith, and Mike Lewis. 2022. Train short, test long: Attention with linear biases enables input length extrapolation. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Jack W. Rae and Ali Razavi. 2020. Do transformers need deep long-range memory? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7524–7529. Association for Computational Linguistics.

Nir Ratner, Yoav Levine, Yonatan Belinkov, Ori Ram, Inbal Magar, Omri Abend, Ehud Karpas, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. Parallel context windows for large language models. 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 6383–6402. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 10776–10787. Association for Computational Linguistics.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 464–468. Association for Computational Linguistics.

Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. 2021. Roformer: Enhanced transformer with rotary position embedding. *CoRR*, abs/2104.09864.

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. 2021. Long range arena : A benchmark for efficient transformers. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan

Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Szymon Tworkowski, Konrad Staniszewski, Mikolaj Pacek, Yuhuai Wu, Henryk Michalewski, and Piotr Milos. 2023. Focused transformer: Contrastive training for context scaling. *CoRR*, abs/2307.03170.

Zhongwei Wan, Xin Wang, Che Liu, Samiul Alam, Yu Zheng, Jiachen Liu, Zhongnan Qu, Shen Yan, Yi Zhu, Quanlu Zhang, Mosharaf Chowdhury, and Mi Zhang. 2023. Efficient large language models: A survey. *CoRR*, abs/2312.03863.

Renzhi Wang, Jing Li, and Piji Li. 2023. Infodiffusion: Information entropy aware diffusion process for non-autoregressive text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 13757–13770. Association for Computational Linguistics.

Shuohang Wang, Luowei Zhou, Zhe Gan, Yen-Chun Chen, Yuwei Fang, Siqi Sun, Yu Cheng, and Jingjing Liu. 2020. Cluster-former: Clustering-based sparse transformer for long-range dependency encoding. *CoRR*, abs/2009.06097.

Shuo Yang, Wei-Lin Chiang, Lianmin Zheng, Joseph E. Gonzalez, and Ion Stoica. 2023. Rethinking benchmark and contamination for language models with rephrased samples. *CoRR*, abs/2311.04850.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. GLM-130B: an open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Biao Zhang, Ivan Titov, and Rico Sennrich. 2021. Sparse attention with linear units. 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 6507–6520. Association for Computational Linguistics.

Peitian Zhang, Shitao Xiao, Zheng Liu, Zhicheng Dou, and Jian-Yun Nie. 2023a. Retrieve anything to augment large language models. *Preprint*, arXiv:2310.07554.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Qihao Wu, Junhao Chen, Zihang Xu, Zhenning Dai, Xu Han, Shuo Wang, Zhiyuan Liu, and Maosong Sun. 2023b. Infinitebench: 128k long-context benchmark for language models.

A Related Work

A.1 Long Context Modeling

Large language models (LLMs), such as GPT-4 (Achiam et al., 2023) and Llama (Touvron et al., 2023a,b), have exhibited superior performance across a variety of text generation tasks and practical deployment scenarios (Wan et al., 2023; Guo et al., 2023; Wang et al., 2023). Nonetheless, the principal limitation hindering LLMs from harnessing their greater potential is the context window size—the upper limit of text length the model is capable of processing (Ratner et al., 2023). To circumvent this limitation, methods based on Position Encoding (Shaw et al., 2018), length extrapolation (Newman et al., 2020), and sparse attention mechanisms (Zhang et al., 2021; Gao and Liu, 2023), such as Alibi (Press et al., 2022), RoPE (Su et al., 2021), and Landmark (Mohtashami and Jaggi, 2023), have been presented.

A.2 Evaluation Benchmarks

Existing benchmarks for long context understanding, such as LongBench (Bai et al., 2023), L-Eval (An et al., 2023), and Bamboo (Dong et al., 2023), essentially expand existing NLU datasets, which may not pose sufficient difficulty and are prone to data contamination, and often fall short in text length. Besides, M^{4}LE (Kwan et al., 2023) constructs texts from fragments of multiple summarization datasets to control text length, but this approach lacks the need for global understanding. LooGLE (Li et al., 2023) and InfiniteBench (Zhang et al., 2023b) introduces a broader range of tasks. However, the manual annotation required for such a benchmark is extremely costly. By way of contrast, XL^{2}Bench leverages LLMs and meticulous human review to construct the benchmark cost-effectively.

B Task Descriptions

In this section, we provide detailed descriptions of the input and output content of 27 subtasks. Please note that the input includes a long text and an instruction. We only describe the instruction.

Models	MR		DU		OU						TG		
	C-L	*C-R*	*C-S*	*QA*	*C-C*	*B-S*	*E-E*	*F-S*	*Ch-D*	*Re-A*	*RP-C*	*N-G*	*P-G*
YaRN-Mistral-7B	<1	<1	6.64	2.29	5.52	10.16	2.85	3.13	10.09	8.52	4.36	4.42	5.40
InternLM2-C-7B	<1	<1	3.08	<1	<1	7.73	5.15	4.57	7.01	2.31	6.90	4.23	21.88
InternLM2-C-20B	18.85	1.58	17.60	35.43	56.01	17.47	29.81	25.04	19.97	20.73	53.14	29.79	44.81
Moonshot-V1	38.19	33.56	**24.46**	34.14	**88.89**	**30.30**	38.79	39.16	28.45	25.46	37.10	61.76	62.47
GLM-4	26.68	34.60	18.06	32.86	66.67	28.75	34.46	24.30	25.24	**27.56**	39.20	35.07	53.12
GPT-4-Turbo	**55.46**	**42.70**	19.76	**50.81**	77.50	29.30	**44.20**	**42.57**	**30.87**	27.16	**66.71**	**74.59**	**67.80**
Qwen-Long	24.67	16.71	19.50	43.80	37.41	28.43	36.81	35.58	26.29	23.19	45.33	66.12	62.03

Table 6: Results (%) of seven LLMs on English Fiction Reading.

Models	MR	DU		OU		TG	
	C-R	*Sec-Sum*	*T-E*	*Paper-C*	*Paper-Sum*	*P-Review*	*R-Score*↓
YaRN-Mistral-7B-128K	<1	10.19	15.86	11.69	5.04	33.23	None
InternLM2-Chat-7B-200K	<1	6.82	5.04	<1	7.31	39.80	None
InternLM2-Chat-20B-200K	25.84	24.91	30.27	33.37	34.41	45.11	2.30
Moonshot-V1-128K	31.02	45.78	31.43	44.44	36.68	**66.04**	4.39
GLM-4-200K	25.76	29.66	33.40	47.62	36.91	55.62	**2.23**
GPT-4-Turbo-200K	**45.28**	**51.57**	**55.91**	**55.56**	**45.91**	62.12	2.63
Qwen-Long-1M	18.00	29.16	29.40	30.67	40.38	58.09	2.89

Table 7: Results (%) of seven LLMs on Paper Reading. **Sec-Sum**, **T-E**, **Paper-C**, **Paper-Sum**, **P-Review**, and **R-Score** represent *Section Summarization*, *Terminology Explanation*, *Paper Counting*, *Paper Summarization*, *Paper Review*, and *Rating Score* respectively. **None** signifies the model's inability to generate a rating score, thus rendering it incapable of fulfilling the requirements of this subtask.

B.1 Fiction Reading

Content Location Given the content of the fiction, the model outputs the location.

Content Retrieval Given a location, the model outputs the corresponding fiction content.

Chapter Summarization Given a chapter number of the fiction, the model summarizes the corresponding chapter.

Question Answering Give a detailed question about the fiction, the model outputs the answer.

Chapter Counting The model outputs the quantity of the fiction.

Background Summarization The model outputs the time background, place background, and social and cultural background of the fiction.

Event Extraction The model outputs the main events of the fiction in chronological order.

Fiction Summarization The model summarizes the whole fiction.

Character Description The model outputs the description of the character in the fiction, including personality traits and personal experiences.

Relationship Analysis The model outputs the relationship between two characters.

Role-play Conversation Given a question, the model needs to assume the role of a character from the fiction to provide an answer.

News Generation The model assume a news reporter within the world of the fiction, and reports on the final event involving the protagonist's team, including the background of the event, the actions of the protagonist, the outcome, and the impact of the event.

Poem Generation The model writes a poem based on the core theme, key plot, important characters and specific context of the fiction.

B.2 Paper Reading

Content Retrieval Given a location, the model outputs the corresponding paper content, such as title, authors.

Section Summarization Given a section number of the paper, the model summarizes the corresponding section.

Terminology Explanation Given an scientific noun in the paper, the model outputs its explanation.

Paper Counting The model output the quantity of titles, authors, references, tables, figures, etc. of the paper.

Paper Summarization The model summarizes the whole paper.

Paper Review The model assumes the role of a peer reviewer for an academic journal, and outputs a review of the paper, including: strengths and weaknesses.

Rating Score The model assumes the role of a peer reviewer for an academic journal, and outputs a rating score of the paper from 0 to 10.

B.3 Law Reading

Legal Entry Location Given the content of the law, the model outputs its corresponding index.

Legal Entry Retrieval Given a locating of a legal entry, the mode outputs its content.

Legal Definition QA Given a question about the law's definitions, the model outputs the answer.

Legal Number QA Given questions about the numbers in law, the model outputs the answer.

Legal Entry Counting The model outputs the quantity of legal entries in this law.

Multiple Choices QA Given a question with multiple choices, the model outputs the answer.

Case Adjudication Given a legal case, the model outputs the verdict.

C Evaluation Metrics

Automatic Evaluation For tasks with fixed answers, such as Content Location in Fiction Reading, we adopt **Accuracy** as an intuitive measure to demonstrate the model's performance. For MCQA, we utilize **F1-Score** to objectively evaluate the model's capability to accurately answer all the correct options. For summary tasks, we select **Rouge-L** to reflect whether the model can correctly identify key information in a document. For generative tasks, we employ **BLEU** to measure the congruence between the generated content by model and the reference content. For Rating Score subtask, we choose **MAE** to calculate the average absolute difference between predicted and true scores. Details can be found in Table 1.

Human Evaluation It has been correctly noted that numerous studies have exposed significant limitations in N-grams matching-based metrics for open-ended generation tasks (Callison-Burch et al., 2006; Chali and Hasan, 2012). To address the shortcomings associated with Rouge-L and BLEU, we engage volunteers to perform human evaluations on corresponding tasks. These individuals possess a thorough familiarity with the narratives in question. In the evaluation phase, volunteers are presented with outputs from all models simultaneously. They are then asked to rank these outputs based on perceived quality. Our ranking system accommodates ties, with subsequent rankings adjusted to reflect these equivalences[8]. We present the average ranking of each model across all tasks.

D Evaluation Input Templates

For all texts and corresponding questions in XL^{2}Bench, we use the following template: *Please read the following text, and answer related question:* [text] *Question:* [question] *Directly output your answer without any additional analysis or explanation.*

E Results on English Fiction Reading and Paper Reading

We show the remaining results of seven LLMs on English Fiction Reading and Paper Reading in Table 6 and Table 7.

F Assessment of In-context Learning Ability

Table 8 demonstrates significant enhancements primarily in summarization tasks. Through in-context learning, models are capable of generating outputs that closely align with the desired format, thus elevating their scores. Conversely, in tasks necessitating brief responses, models exhibit limited ability to leverage the prompts for noticeable improvement. Specifically, GPT-4-Turbo, despite its substantial parameter count, shows negligible performance shifts following in-context learning application.

[8]For example, the rankings might be represented as $1, 1, 3, 4, 5, 6, 6$.

Models	Paper Reading			Law Reading		
	Content-R	*Sec-Sum*	*T-Explain*	*LE-L*	*Def-QA*	*Num-QA*
YaRN-Mistral-7B-128K	<1	10.19	15.86	<1	8.62	<1
w/ ICL	<1	9.81	14.30	<1	7.79	<1
InternLM2-Chat-20B-200K	25.84	24.91	30.27	5.41	40.57	58.03
w/ ICL	31.89	33.67	38.50	6.76	39.90	58.82
GPT-4-Turbo-128K	45.28	51.57	55.91	13.41	40.26	62.50
w/ ICL	46.77	50.89	56.12	14.81	40.88	61.58

Table 8: Results (%) of three LLMs using zero-shot learning and few-shot learning on several tasks of Paper Reading and Law Reading. The data in the **gray** section is derived from the previous tables.

Scenarios	MR		DU		OU						TG		
	C-L	*C-R*	*C-S*	*QA*	*C-Q*	*F-B*	*F-E*	*F-S*	*Ch-D*	*Ch-R*	*Ch-DG*	*N-G*	*P-G*
Fiction	<1	**6.85**	**17.22**	**9.82**	**53.33**	**15.58**	**18.61**	**17.29**	**21.98**	**28.92**	11.65	**16.67**	**10.09**
Fiction-T	<1	6.54	12.28	5.05	52.16	10.21	10.80	6.67	2.28	13.89	12.36	11.89	5.01
Fiction-R	<1	6.76	5.11	6.48	53.33	8.04	11.72	4.96	3.33	17.67	12.12	11.84	5.78
Fiction-C	<1	6.28	5.23	3.39	53.33	7.65	4.46	13.41	2.49	15.56	**13.79**	12.68	7.91

Table 9: Results (%) of InternLM2-Chat-20B-200K on Fiction, Fiction-T, Fiction-R, and Fiction-C.

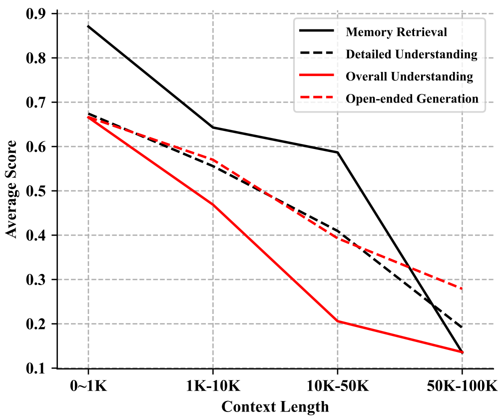

Figure 3: Average score (%) of four tasks under different context length on Law Reading.

G Impact of Context Length

Figure 3 illustrates that the model's performance significantly declines with longer texts, as evidenced by a steeper curve. This observation underscores the model's challenges in effectively managing the complexities of long text modeling.

H Results of Ablation Study

In this section, we assess the effectiveness of our data augmentation strategies in mitigating the impact of data contamination on model evaluation outcomes. We specifically examine the performance of the InternLM2-Chat-20B across different subsets of fiction data, namely Fiction, Fiction-T, Fiction-R, and Fiction-C, with the results detailed in Table 9. The observed reduction in performance across almost all subtasks within the augmented dataset indicates that our data augmentation techniques can, to some extent, reduce the likelihood of biased evaluations.