LongIns: A Challenging Long-context Instruction-based Exam for Large Language Models

Anonymous ACL submission

Abstract

The long-context capabilities of large language models (LLMs) have been a hot topic in recent years. To evaluate LLMs' performance in different scenarios, various assessment benchmarks have emerged. However, as most of these benchmarks focus on identifying key information to answer questions, which mainly requires the retrieval ability of LLMs, these benchmarks can partially represent the reasoning performance of LLMs from large amounts of information. Meanwhile, although LLMs often claim to have context windows of 32k, 128k, 200k, or even longer, these benchmarks fail to reveal the actual supported length of these LLMs. To address these issues, we propose the ***LongIns*** benchmark dataset, a challenging long-context instruction-based exam for LLMs, which is built based on the existing instruction datasets. Specifically, in our *LongIns*, we introduce three evaluation settings: **Global Instruction & Single Task (GIST)**, **Local Instruction & Single Task (LIST)**, and **Local Instruction & Multiple Tasks (LIMT)**. Based on *LongIns*, we perform comprehensive evaluations on existing LLMs and have the following important findings: (1). The top-performing GPT-4 with 128k context length performs poorly on the evaluation context window of 16k in our *LongIns*. (2). For the multi-hop reasoning ability of many existing LLMs, significant efforts are still needed under short context windows (<4k).

1 Introduction

The topic of extending the context window length in large language models (LLMs) remains a focal point in current research. Existing LLMs can handle context lengths ranging from 32k to 200k tokens (Wang et al., 2024), while some LLMs achieving capacities up to 10M tokens (Liu et al., 2024a; Bai et al., 2023a). The capacity of length expansions is crucial for enhancing long document comprehension and forms the basis for a series of advanced applications, including repository-level code. Comprehension/Editing and the dependence of agent-based tasks on long-term memory. Despite these advancements, the predominant evaluation metrics continue to prioritize the retrieval performance of LLMs, overlooking the actual window length that the model can understand when faced with long texts.

The capabilities of LLMs in scenarios involving long-context , such as cross-document aggregation, localization, and context tracking, is paramount for ensuring a seamless interaction between users and LLMs. Users often expect LLMs to read and understand the entire documents. However, existing benchmarks measure the proficiency of LLMs based on their capacity to retrieve and understand only key information from the inputs, thereby bypassing the necessity for an exhaustive understanding of the entire text. This does not align well with the real-world expectations from users who seek in-depth processing of long texts. Datasets commonly used to evaluate the long-text capabilities of LLMs, such as L-eval (An et al., 2023) or longBench (Bai et al., 2023b), are insufficient for assessing the understanding of long sequences. *Therefore, how to truly evaluate the capabilities of LLMs in handling long-context tasks of the realistic scenarios still requires further exploration.*

To bridge this gap, we introduce *LongIns*, a benchmark tailored to critically assess the proficiency of LLMs in understanding extensive sequences. *LongIns* incorporates a contextual learning approach where the input involves more substantial key information segments, and is poised to ensure that correctly answering the culminating question necessitates a deep comprehension of the entire lengthy input sequence. This approach mandates that LLMs must truly excel in long-sequence understanding capabilities, beyond just the retrieval of key information from extended texts. We evaluate the performance of 20 different LLMs under

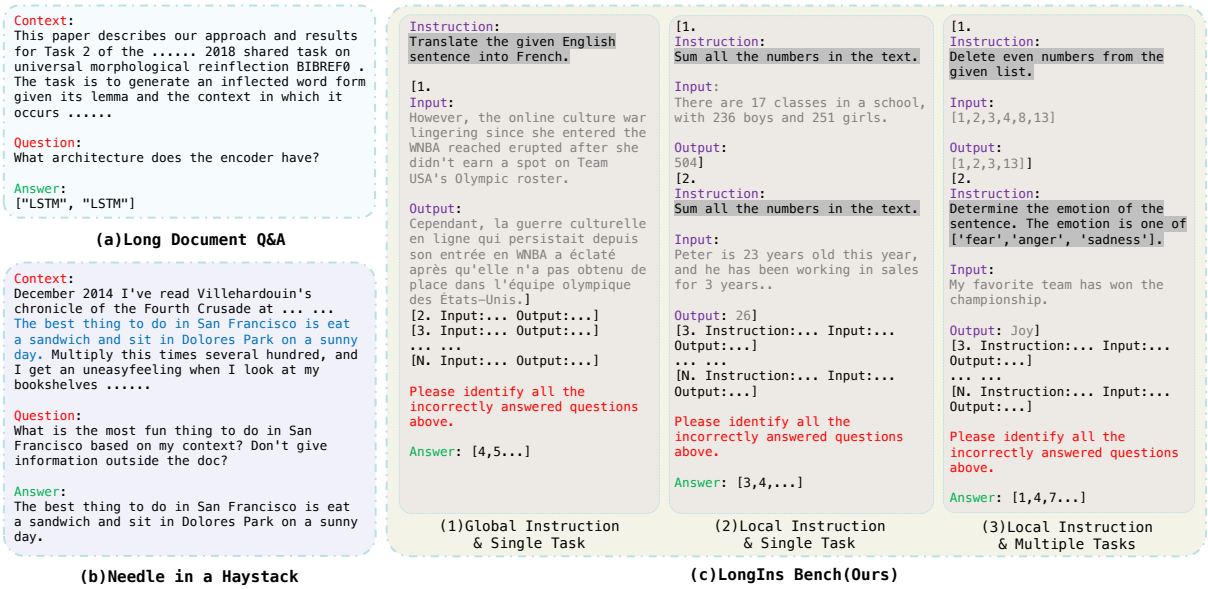


Figure 1: (a) Long Document Q&A is one of the common types of questions in mainstream long-context benchmarks. (b) Needle in a Haystack is an evaluation paradigm that tests the retrieval capabilities of LLMs in long contexts by inserting key information into long texts and asking questions based on that. (c) Our *LongIns* consists of (1) GIST mode, Each sample is composed of the same task type concatenated together, and the task instruction is given only once at the beginning globally. (2) LIST mode, which is similar to GIST in the composition of samples, but instructions are provided for each question within the sample; (3) consists of questions from different task types concatenated together.

LongIns, including GPT-4o. We observe that LLMs generally perform worse on tasks requiring understanding of complete long sequences compared to retrieval tasks of the same length. As the total length increases, the performance gap becomes more pronounced. Additionally, the models' performance is largely independent of their advertised context window length, leading us to believe that the advertised context window length for most models should be understood as the "maximum acceptable sequence length" rather than the "maximum comprehensible sequence length."

Moreover, by controlling the distribution of incorrect answers and response situations, we analyze the distribution of attention changes at different text positions. Further analysis on the density of key information within the same total length shows that, except for GPT-4 and GPT-4o, the accuracy of most models rapidly declines as the density of key information increases. In summary, the primary contributions can be summarized as:

- We provide the *LongIns* benchmark test, specifically designed to evaluate the actual understanding and processing capabilities of LLMs on long sequences. Unlike other benchmarks based on retrieval tasks, we focus more on assessing the actual comprehensible window length of the models.
- In our *LongIns*, we introduce three evaluation settings: Global Instruction & Single Task (GIST), Local Instruction & Single Task (LIST), and Local Instruction & Multiple Tasks (LIMT) to comprehensively evaluate the long-context abilities of existing LLMs.
- We evaluate a series of long-context LLMs using this benchmark and find that most models fail to achieve high scores when the critical information length is only 8k. Even GPT-4 and GPT-4o score poorly at 16k length. This result is significantly different from the commonly recognized long context lengths (128k or longer), indicating that current LLMs still have considerable shortcomings in performing such tasks. We hope these results can provide a reference for research in the field of long-context LLMs.

2 Related Work

Long-context LLM The computational cost of processing sequences in Transformer-based models increases quadratically with sequence length,

resulting in higher resource consumption and performance issues when handling long context inputs. Many studies explore various strategies to address this challenge, including the use of new positional encodings to achieve position interpolation or extrapolation (Peng et al., 2023; Chen et al., 2023; Liu et al., 2024b). For example, Rope (Su et al., 2021) extends the positional knowledge learned during the pre-training phase to longer sequence lengths through extrapolation, including various variants such as NTK-RoPE (NormXU, 2021). Alibi (Press et al., 2021), on the other hand, maps long sequences to within recognizable lengths through interpolation. Some studies attempt to fine-tune LLMs to give the model a longer context window. LongLoRA (Chen et al., 2024)is an efficient fine-tuning approach that significantly extends the context sizes of pre-trained LLMs with limited computational cost by using sparse local attention and improved parameter-efficient fine-tuning techniques. RingAttention (Liu et al., 2023a) can reduce the memory requirements of the Transformer model, allowing it to train sequences over 500 times longer than previous memory-efficient methods, without needing to approximate the attention mechanism. Additionally, traditional engineering techniques such as sliding windows or RAG (Lewis et al., 2020) are also solutions for addressing long context scenarios in LLMs. These methods improve the performance of LLMs in handling long contexts in certain aspects.

Long-context Benchmark To evaluate the performance of different LLMs on long texts, various benchmarks are often used to test different aspects of LLMs' capabilities. For instance, Zero-scrolls (Shaham et al., 2023) is a zero-shot benchmark for natural language understanding over long texts, which contains only test and small validation sets, without training data. LongBench (Bai et al., 2023b) is a bilingual, multi-task benchmark for long context understanding, which also provides a subset with uniformly distributed data called LongBench-E. L-Eval (An et al., 2023) constructs a long context evaluation benchmark containing multiple tasks and domains, including multiple-choice questions, with data that has undergone strict manual screening to ensure high quality. LooGLE (Li et al., 2023), as a benchmark with longer samples, effective for evaluating the ability of LLMs to understand short-term and long-term dependencies in the content. M4LE (Kwan et al., 2023) provides additional tasks and datasets in both Chinese and English, covering a wider range of domains. ∞BENCH is the first benchmark with more than 100K tokens, with the longest test items reaching up to 2000K tokens. Needle in a Haystack (Doe and Smith, 2022) employs Paul Graham's 218 essays as the "haystack" inserting a single sentence at various depths within the documents to statistically analyze the model's accuracy in identifying the locations of various "needles" and thereby evaluate its performance.

Model	Size	Strategy	Support
GPT-4o	🔒	🔒	128k
GPT-4-Turbo	🔒	🔒	128k
GPT-3.5-Turbo	🔒	🔒	16K
ERNIE-Speed	🔒	🔒	8k
Qwen-Long	🔒	NTK+LNS+WAttn	10M
Deepseek-Chat	🔒	YaRN	32k
Moonshot-v1	🔒	🔒	128k
Yi-Large-Turbo	🔒	SeqPara+DistrAttn	200k
Yi-Spark	🔒	SeqPara+DistrAttn	200k
GLM-4	🔒	🔒	128k
ChatGLM2-6B	6B	PI	32k
Deepseek-LLM-Chat	7B	-	4k
Lwm-Text-Chat	7B	RingAttn	128k
LongChat-7B	7B	Rope	16k
MAP-Neo-Ins-v0.1	7B	-	8k
Mistral-7B-Ins-v0.2	7B	Rope	32k
Longpalca-7B	7B	LongLoRA	32k
Qwen1.5-7B-Chat	7B	Rope+Sliding Window	32k
Internlm2-Chat-7B	7B	Dynamic NTK	200k
Llama3-8B-Ins	8B	RoPE	8K
LongChat-13B	13B	Rope	16k
Baichuan2-13B-Chat	13B	-	4k

Table 1: Details on the evaluated models. "Ins" indicates "Instruct". "LNS" indicates "LogNScaling". "WAttn" indicates "WindowAttention"

3 *LongIns* Benchmark

To support the evaluation of various types and lengths of tasks, we collect a diverse range of questions from the Super-NaturalInstructions (SNI) and BIG-bench datasets, covering different areas of natural language processing. This ensures comprehensive coverage of task types and difficulty levels. To address the issue of insufficient samples in some categories, we generate additional questions. As detailed in Table 2, our dataset, named *LongIns*, includes seven different context lengths: 256, 512, 1024, 2048, 4096, 8192, and 16384 tokens, with 1409 questions for each length. These questions divide into seven task types, including question answering, classification, reading comprehension, natural language inference, translation, named entity recognition, and commonsense reasoning.

Specifically, we construct our dataset by concate-

	GIST		LIST		LIMT	
	Number of Questions	**Q-Density**	**Number of Questions**	**Q-Density**	**Number of Questions**	**Q-Density**
QA	265 × 7	1.40	265 × 7	1.39	-	-
Classif	229 × 7	1.67	229 × 7	1.65	-	-
RC	142 × 7	1.73	142 × 7	1.71	-	-
NLI	140 × 7	1.72	140 × 7	1.70	-	-
MT	432 × 7	1.10	432 × 7	1.09	-	-
NER	96 × 7	2.32	96 × 7	2.27	-	-
CSR	105 × 7	2.69	105 × 7	2.64	-	-
Total	1409 × 7	1.41	1409 × 7	1.39	50 × 7	1.40

Table 2: The distribution of task types in the LongBBH dataset. "QA", "Classif", "RC", "NLI", "MT", "NER", "CSR" represent Question Answering, Classification, Reading Comprehension, Natural Language Inference, Machine Translation, Named Entity Recognition, and Common Sense Reasoning, respectively. "Q-Density" represents the number of questions per 100 tokens, reflecting question density in each paper. Samples in LIMT consist of a mix of various task types, so the information is not presented by category.

nating questions to the specified context lengths and modifying some of the questions' answers to incorrect ones (we control the overall error rate at around 10% but ensure at least one incorrect answer for shorter lengths). The aim is to reflect the long-context performance of LLMs based on their efficiency in identifying incorrect answers after reading multiple questions simultaneously. Based on the above concept, we construct a dataset using the Global Instruction & Single Task (GIST) mode. Additionally, to explore the impact of different factors on the evaluation results, we create two other datasets: Local Instruction & Single Task (LIST) and Local Instruction & Multiple Tasks (LIMT).

Global Instruction & Single Task The GIST dataset serves as our core benchmark. GIST evaluates by concatenating the same type of questions to the specified lengths, constructing prompts for each sample according to the template shown in Figure 6. To assess the model's attention capabilities across different lengths and positions, we ensure an even distribution of incorrect answers to maintain statistical significance. During testing, for open-source models, we use the official example system prompt; if no official example is provided, we default to standard configurations. For closed-source models, we utilize examples from the official documentation.

Local Instruction & Multiple Tasks Each piece of sample in GIST is composed of questions of the same task type, sharing a common task description. Therefore, the task description (i.e., Instruction) is placed at the beginning globally. However, this positioning might result in the subsequent questions being far from the Instruction, potentially affecting accuracy. To explore this issue, we construct a dataset in LIST mode for ablation experiments, as illustrated in Figure 7.

Local Instruction & Multiple Tasks Both GIST and LIST benchmarks are composed of questions of the same task type. When the context content is similar, the model's in-context learning ability becomes more significant. To explore the role of the model's in-context learning ability in the evaluation results, we design a subset called LIMT. LIMT includes 7 lengths, with each length containing 200 test items formed by mixing various task types. The construction template prompt for LIMT is similar to Figure 7, with the only difference being that each question within the sample corresponds to a different instruction.

4 Experiment

In this chapter, we present the experimental methods as well as the main findings and results obtained through these experiments.

4.1 Experimental Setup

In this paper, we conduct a comprehensive evaluation of recent open-source and closed-source LLMs (Bai et al., 2023a; AI et al., 2024; Zeng et al., 2023; DeepSeek-AI et al., 2024; Du et al., 2022; Liu et al., 2024a; Li* et al., 2023; Zhang et al., 2024; Team, 2023; AI@Meta, 2024; Baichuan, 2023; Achiam et al., 2023; Moo; ERN). Since the core idea of this benchmark is "long key information" rather than the total length of the text, we include some models that only retain the original 4k length window. Table 1 provides basic information about the models we include.

Model	Size	Support	256	512	1024	2048	4096	8192	16384
GPT-4o	🔒	128k	70.94	59.14	60.58	55.43	52.92	43.81	31.53
GPT-4-Turbo	🔒	128k	69.59	63.13	64.21	59.08	57.52	50.73	40.91
GPT-3.5-Turbo	🔒	16K	54.61	45.38	41.68	34.81	26.27	18.81	12.23
ERNIE-Speed	🔒	8k	44.16	35.99	24.88	17.29	12.22	1.13	-
Qwen-Long	🔒	10M	61.58	54.60	42.07	34.24	25.99	18.71	10.33
Deepseek-Chat	🔒	32k	45.71	40.86	32.41	19.32	11.33	5.42	3.10
Moonshot-v1	🔒	128k	69.45	61.22	52.35	52.94	34.89	24.29	15.10
Yi-Large-Turbo	🔒	200k	57.05	42.22	33.81	28.60	19.82	11.69	6.01
Yi-Spark	🔒	200k	51.61	39.04	35.53	24.44	15.97	5.08	0.88
GLM-4	🔒	128k	69.20	64.66	56.62	56.11	33.50	23.72	12.01
ChatGLM2-6B	6B	32k	20.75	17.00	14.08	10.81	6.38	2.93	0.87
Deepseek-LLM-Chat	7B	4k	29.08	24.01	16.75	12.99	3.61	-	-
Lwm-Text-Chat	7B	128k	22.19	15.20	10.77	8.80	3.18	0.73	-
LongChat-7B	7B	16k	24.58	20.85	17.48	14.35	9.62	6.31	2.54
MAP-Neo-Ins-v0.1	7B	8k	41.22	34.17	28.09	21.96	14.72	3.28	-
Mistral-7B-Ins-v0.2	7B	32k	40.53	37.00	29.15	20.37	0	0	0
Longalpaca-7B	7B	32k	31.75	25.37	20.08	14.94	9.68	4.74	1.95
Qwen1.5-7B-Chat	7B	32k	33.48	29.88	24.32	19.50	16.72	11.57	5.22
Internlm2-Chat-7B	7B	200k	52.14	45.23	36.84	26.88	18.19	11.98	6.06
Llama3-8B-Ins	8B	8k	52.93	46.35	40.63	32.06	22.48	10.09	-
LongChat-13B	13B	16k	27.96	25.52	22.19	18.06	14.30	8.21	3.15
Baichuan2-13B-Chat	13B	4k	27.99	23.96	21.37	17.19	0	-	-
Yi-34B	34B	200k	26.98	22.14	17.00	13.24	8.13	4.02	1.19

Table 3: The evaluation results on GIST Subset. We report scores of multiple open-source and closed-source models across different lengths. “Ins” indicates “Instruct”.

The goal of this work is to construct a benchmark where LLMs must read and understand the entire context word by word to answer correctly. To achieve this, we employ an examination method requiring LLMs to carefully read all questions in a long text to correctly identify the “numbers of questions with incorrect answers.” A specific example can be found in Table 7 in the appendix.

We also conduct experiments on the accuracy distribution of LLMs when incorrect questions are located at different depths for different text lengths. This approach reflects focus ability of LLMs at various positions and lengths. Additionally, since the length of questions affects LLM performance for the same total length, we conduct further analysis on question density under the same total length.

4.2 Results on GIST Subset

The overall evaluation results of our study are presented in Table 3. Our inputs are sample created by concatenating multiple questions of the same type, with some answers intentionally altered to be incorrect. We generate prompt words using appropriate templates to guide the model, which serves as a grading tool, to identify all incorrectly answered questions. We use the F1 score between the list of actual incorrect question numbers and predicted incorrect question numbers predicted by LLMs. The results in the table indicate that closed-source models generally outperform open-source models, especially in fields requiring longer key information. Notably, GPT-4-turbo and GPT-4o maintain scores of 40.91 and 31.53, respectively, even with an exam length of 16,000 tokens, whereas other models are practically unusable at this length. Although a 256-token exam length is considered relatively short in the context of LLM interactions, certain LLMs still perform poorly at this length. This suggests that LLM performance declines when facing slightly longer texts that require sustained attention to key information. These findings highlight certain shortcomings in current LLMs, particularly in their multi-hop compositional reasoning abilities.

4.3 Results on LIST Subset

To investigate the effect of LLMs’ dependence on instruction positioning, we conduct an ablation experiment by changing the instructions from global instructions to local instructions. The specific prompts are shown in Figure 7. For the GIST subset, the instructions are given only at the beginning, but for the LIST subset, instructions are provided before each question to explore the model’s dependency performance at different distances between the instructions and inputs. The results, as shown in Table 5, indicate that compared to the evaluation results on the GIST subset, models generally achieve higher scores on the LIST sub-

Model	Size	Support	256	512	1024	2048	4096	8192	16384
GPT-4o	🔒	128k	76.31	74.65	71.16	66.55	58.83	56.23	51.15
GPT-4-Turbo	🔒	128k	73.89	69.01	65.16	60.22	59.63	51.49	44.22
GPT-3.5-Turbo	🔒	16K	51.60	50.12	42.04	32.07	18.58	19.67	7.64
ERNIE-Speed	🔒	8k	42.34	36.77	33.52	23.67	16.86	4.18	-
Qwen-Long	🔒	10M	59.00	54.45	53.87	47.34	38.29	35.22	23.97
Deepseek-Chat	🔒	32k	69.08	67.56	59.19	47.28	44.67	41.56	35.23
Yi-Large-Turbo	🔒	200k	50.03	44.43	38.53	33.89	27.43	26.19	17.38
Yi-Spark	🔒	200k	43.93	38.19	34.39	29.86	26.66	21.26	15.92
GLM-4	🔒	128k	58.52	53.00	48.58	44.64	42.15	41.02	38.74
ChatGLM2-6B	6B	32k	20.75	17.00	14.08	10.81	6.38	2.93	0.87
Deepseek-LLM-Chat	7B	4k	29.08	24.01	16.75	12.99	3.61	-	-
Mistral-7B-Ins-v0.2	7B	32k	40.53	37.00	29.15	20.37	0	0	0
Qwen1.5-7B-Chat	7B	32k	33.48	29.88	24.32	19.50	16.72	11.57	5.22
Internlm2-Chat-7B	7B	200k	52.14	45.23	36.84	26.88	18.19	11.98	6.06
Llama3-8B-Ins	8B	8k	52.93	46.35	40.63	32.06	22.48	10.09	-
Baichuan2-13B-Chat	13B	4k	27.99	23.96	21.37	17.19	-	-	-

Table 4: The result of Local Instruction & Multiple Tasks. “Ins” indicates “Instruct”

set. Changing the position of instructions leads to higher model scores, indicating that most models are sensitive to the distance of instruction dependency. When instructions and inputs are separated by a greater distance, the performance of most models rapidly degrades. This is common across various models such as GPT-4-Turbo, including those that perform well in other tasks. And this issue generally affects the model's reliance on the early memories in multi-turn conversations. But as a trivial solution, repeating the important instructions would significantly increase inference costs.

4.4 Results on LIMT Subset

To explore the long-context comprehension ability of LLMs outside the in-context learning paradigm, we also concatenate the questions of different task types to the same length and evaluate on them. We extract a mini-mix dataset comprising 50 samples for each of the 7 task types, totaling 350 samples. The evaluation results are shown in Table 4.

We observe that models generally score higher under the LIMT setting compared to the GIST setting. This is because questions from different types of tasks have different instructions, and therefore each output needs to be given separately. As a result, the distance between the question and the instructions is closer, and the dependence on the instructions becomes more apparent, leading to higher scores. In contrast, when compared with the evaluation results in Table 5, it is evident that the LIMT setting are more challenging than the LIST setting under the same input format. This indicates that most models still fall short in tasks outside of the in-context learning paradigm.

5 Discussion

During the analysis, we discover several noteworthy phenomena:

- The Yi-spark model, small-parameter model, achieves results close to those of larger parameter models, suggesting that the ability to focus on long key information is not strictly dependent on the number of parameters.
- The ERNIE-Speed model has a relatively low score because its strict review mechanism results in many refusals to answer.
- Some open-source models, such as baichuan2-13B-Chat and mistral-7b-32k, exhibit a sharp degradation when the length of key information approaches but does not exceed their context window limit. For instance, when the test length jumps from 2k to 4k tokens, these models suddenly fail to follow instructions correctly and start producing nonsensical outputs. We believe this implies the true length that some models can handle effectively.

5.1 Effect of the Position of Key Information

“Needle in a Haystack” (Doe and Smith, 2022) and “Lost In the Middle” (Liu et al., 2023b) indicate LLMs exhibit varying levels of attention when processing different positions within long texts. Therefore, we analyze the LLMs' ability to handle key information located at different positions by controlling the distribution of error question.

Specifically, We analyze the model's sensitivity to position when dealing with lengthy key information by controlling the depth of incorrect questions

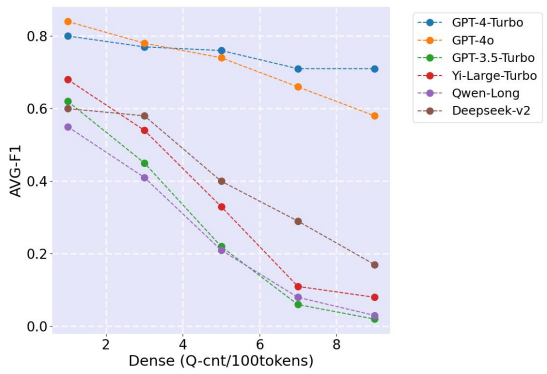


Figure 2: The impact of question density on model performance. The horizontal axis represents the number of questions per 100 tokens in the paper; the higher this value, the more questions are contained within the same total length. The vertical axis represents the model's score at that question density.

within papers of various lengths and recording the model's accuracy at those positions. The results are shown in Figure 3.

Through the above experiment, we can clearly observe the performance of models when faced with incorrect answers at different depths and lengths of test papers. Using the GPT series and Yi series as examples, both exhibit the same characteristics: the models demonstrate better recognition ability at the beginning of the test paper for the same length, with performance gradually declining as depth increases. The longer the overall length of the test paper at the same depth, the worse the performance. This is generally in line with our expectations. However, it is noteworthy that, unlike needle-in-a-haystack or other long-text benchmarks, *LongIns* only requires length of 16k to challenge the current state-of-the-art GPT-4 series models, while conventional closed-source models (such as the Yi series models) only require 8k to basically reach their performance limit. More depth-length two-dimensional diagrams for additional models can be found in Figure 5 in the appendix.

5.2 Effect of the Density of Key Information

In our current setup, we concatenate questions of the same type to form a test paper, maintaining a constant length. However, due to variations in the lengths of the questions, the number of questions in papers of the same length varies. Therefore, we analyze the accuracy of multiple models with varying numbers of questions (i.e., question density) under the same test length, aiming to explore the impact of key information density on LLMs. We track the scoring results of test papers with different question densities and plot a graph of scores varying with question density.

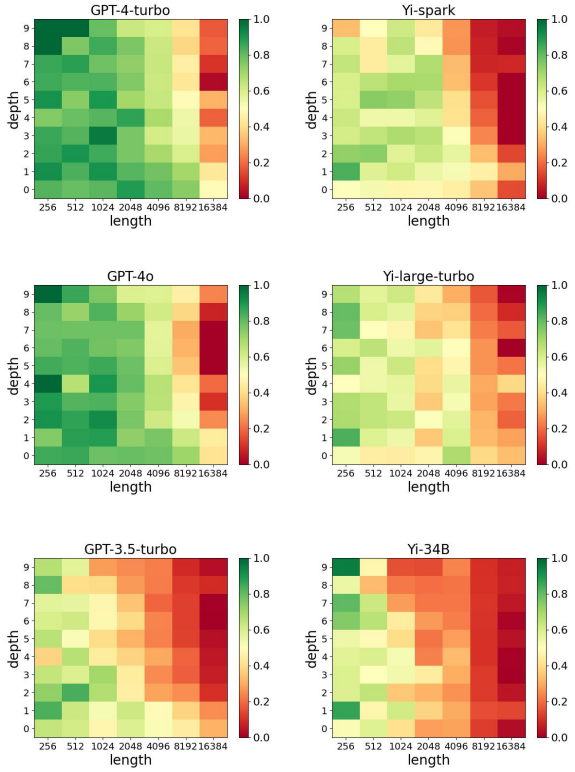


Figure 3: The impact of positional distribution and overall length of the questions on model performance is illustrated, with GPT series models on the left and Yi series models on the right. The horizontal axis represents the total length of the questions, while the vertical axis indicates the position of incorrect answers within the questions, where 0 denotes the beginning and 9 denotes the end. The greener the color, the higher the model's recognition accuracy at that position.

As illustrated in Figure 2, it is evident that most models exhibit a performance degradation as the number of questions increases under the same test paper length. This includes high-scoring models such as Qwen. This demonstrates most models are more sensitive to the number of questions when the total length is the same.

This also indicates that besides the length of the sample, the density of key information is also an important factor affecting performance. However, GPT-4-turbo and GPT-4o maintain good performance even at higher question densities, which to some extent indicates their robustness to changes in key information density. They can still maintain high-density information processing capabilities in long contexts.

Model	Size	Support	256	512	1024	2048	4096	8192	16384
GPT-4o	🔒	128k	74.89	74.62	70.43	66.98	61.82	59.07	55.30
Deepseek-Chat	🔒	32k	67.71	63.13	56.99	51.04	46.78	42.00	37.25
GLM-4	🔒	128K	60.25	54.11	49.79	46.05	41.32	37.88	34.15
MAP-Neo-Ins-v0.1	7B	8k	39.34	33.71	28.45	23.82	16.04	6.83	-
Deepseek-LLM-Chat	7B	4k	17.70	12.93	8.42	7.47	1.16	-	-
Qwen1.5-7B-Chat	7B	32k	25.93	24.71	20.81	14.71	12.35	11.05	10.69
Llama3-8B-Ins	8B	8k	46.84	40.97	37.41	27.46	21.11	16.08	-

Table 5: The result of Local Instruction & Single Task. "Ins" indicates "Instruct"

5.3 Ablation Experiment of Single Question

LongIns aims to assess the ability of LLMs to maintain focus over longer key information by identifying incorrect question numbers through a full-text review. However, LLMs might make recognition errors not due to its lack of contextual capabilities but because it inherently lacks the ability to answer questions. Therefore, we set up this ablation experiment to ensure a true evaluation of the long-context abilities of LLMs. We conduct experiments using single questions as test papers for several models. The results, as shown in Table 6, indicate that the accuracy of the models can generally exceed 80%, with some, like GPT-4-turbo, reaching up to 95.7%. These results suggest that *LongIns* provides a reliable evaluation of the actual contextual length performance of LLMs.

Model	Accuracy
GPT-4-Turbo	95.7
GPT-3.5-Turbo	87.0
Deepseek-v1	88.2
Yi-Large-Turbo	87.8
Llama3-8B-Instruct	81.0
ERNIE-Speed	81.5
GLM-4	90.1
Qwen-Long	83.6

Table 6: Directly present the QA of a single question to the model and have it determine whether the answer is correct, controlling the ratio of true labels to be 1:1. The accuracy of the LLMs is then evaluated.

5.4 Effect of Different Task Types

We also explore the performance of LLMs when using test papers composed of different task types., as shown in Figure 4.

A notable trend is that the model generally performs poorly in the common sense category but achieves higher scores in tasks such as NER and Reading Comprehension. By analyzing the actual responses and the questions, we find that despite our efforts to standardize the format of questions within the same test paper, there are still significant distribution disparities among different task types. The questions in the common sense category tend to vary widely within the same test paper, whereas tasks like NER and Reading Comprehension have almost consistent formats, which can stimulate the model's in-context learning ability. When the variation is large, the advantage of in-context learning is almost nonexistent, leading to the model's generally poor performance.

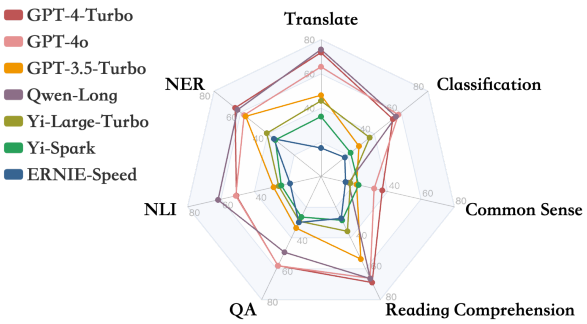


Figure 4: The average performance of the model across different task types.

6 Conclusion

In this paper, we introduce *LongIns* Bench, a benchmark designed to evaluate the long-context processing capabilities of LLMs by assessing their true reading window length. *LongIns* consists of over 1400 task types with varying lengths and controlled distributions of incorrect answers. Our findings highlight that models perform better with closer instruction distances, degrade with increasing length and density of key information, and exhibit varied performance across different task types. Notably, GPT-4-turbo and GPT-4o demonstrate robustness to high-density information. *LongIns* provides a reliable framework for assessing LLMs' proficiency in understanding extensive sequences, offering valuable insights for future research and development.

Limitations

Despite the comprehensive evaluation framework proposed by *LongIns*, there are several limitations to our study. The test set in *LongIns* is constructed using synthetic data, which, while controlled and consistent, may not entirely reflect the complexity and variability of natural language inputs. Additionally, our evaluation methodology emphasizes identifying incorrect answers within long texts, which may not cover other critical aspects of LLM performance such as creativity, user engagement, and adaptability to unforeseen inputs.

Ethics Statement

This research adheres to ethical guidelines for AI development. We aim to enhance the capabilities of LLMs while acknowledging potential risks such as bias, misuse, and privacy concerns. To mitigate these, we advocate for transparency, rigorous bias testing, robust security measures, and human oversight in AI applications. Our goal is to contribute positively to the field and to encourage responsible AI development and deployment.

References

Ernie-speed. https://qianfan.cloud.baidu.com/.

Moonshot. https://www.moonshot.cn/.

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*.

01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. Yi: Open foundation models by 01.ai. *Preprint*, arXiv:2403.04652.

AI@Meta. 2024. Llama 3 model card.

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

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023a. Qwen technical report. *Preprint*, arXiv:2309.16609.

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. 2023b. Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*.

Baichuan. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. 2023. Extending context window of large language models via positional interpolation. *CoRR*, abs/2306.15595.

Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. 2024. Longlora: Efficient fine-tuning of long-context large language models. *Preprint*, arXiv:2309.12307.

DeepSeek-AI, Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Hao Yang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jin Chen, Jingyang Yuan, Junjie Qiu, Junxiao Song, Kai Dong, Kaige Gao, Kang Guan, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruizhe Pan, Runxin Xu, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Size Zheng, T. Wang, Tian Pei, Tian Yuan, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Liu, Xin Xie, Xingkai Yu, Xinnan Song, Xinyi Zhou, Xinyu Yang, Xuan Lu, Xuecheng Su, Y. Wu, Y. K. Li, Y. X. Wei, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Zheng, Yichao Zhang, Yiliang Xiong, Yilong Zhao, Ying He, Ying Tang, Yishi Piao, Yixin Dong, Yixuan Tan, Yiyuan Liu, Yongji Wang, Yongqiang Guo, Yuchen Zhu, Yuduan Wang, Yuheng

Zou, Yukun Zha, Yunxian Ma, Yuting Yan, Yuxiang You, Yuxuan Liu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhewen Hao, Zhihong Shao, Zhiniu Wen, Zhipeng Xu, Zhongyu Zhang, Zhuoshu Li, Zihan Wang, Zihui Gu, Zilin Li, and Ziwei Xie. 2024. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *Preprint*, arXiv:2405.04434.

John Doe and Jane Smith. 2022. Needle in a haystack: Benchmark for extremely imbalanced data. In *Proceedings of the International Conference on Machine Learning (ICML)*.

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. *Preprint*, arXiv:2103.10360.

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.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Dacheng Li*, Rulin Shao*, Anze Xie, Ying Sheng, Lianming Zheng, Joseph E. Gonzalez, Ion Stoica, Xuezhe Ma, and Hao Zhang. 2023. How long can open-source llms truly promise on context length?

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

Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. 2024a. World model on million-length video and language with blockwise ringattention. *Preprint*, arXiv:2402.08268.

Hao Liu, Matei Zaharia, and Pieter Abbeel. 2023a. Ring attention with blockwise transformers for near-infinite context. *Preprint*, arXiv:2310.01889.

Jiaheng Liu, Zhiqi Bai, Yuanxing Zhang, Chenchen Zhang, Yu Zhang, Ge Zhang, Jiakai Wang, Haoran Que, Yukang Chen, Wenbo Su, et al. 2024b. E2-llm: Efficient and extreme length extension of large language models. *arXiv preprint arXiv:2401.06951*.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023b. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

NormXU. 2021. An experiment on dynamic ntk scaling rope. https://github.com/NormXU/Consistent-DynamicNTKRoPE.

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 Omer Levy. 2021. Train short, test long: Attention with linear biases enables input length extrapolation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3672–3683.

Uri Shaham, Maor Ivgi, Avia Efrat, Jonathan Berant, and Omer Levy. 2023. Zeroscrolls: A zero-shot benchmark for long text understanding. *Preprint*, arXiv:2305.14196.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. 2021. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*.

InternLM Team. 2023. Internlm: A multilingual language model with progressively enhanced capabilities. https://github.com/InternLM/InternLM-techreport.

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. *arXiv preprint arXiv:2402.02244*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 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*.

Ge Zhang, Scott Qu, Jiaheng Liu, Chenchen Zhang, Chenghua Lin, Chou Leuang Yu, Danny Pan, Esther Cheng, Jie Liu, Qunshu Lin, Raven Yuan, Tuney Zheng, Wei Pang, Xinrun Du, Yiming Liang, Yinghao Ma, Yizhi Li, Ziyang Ma, Bill Lin, Emmanouil Benetos, Huan Yang, Junting Zhou, Kaijing Ma, Minghao Liu, Morry Niu, Noah Wang, Quehry Que, Ruibo Liu, Sine Liu, Shawn Guo, Soren Gao, Wangchunshu Zhou, Xinyue Zhang, Yizhi Zhou, Yubo Wang, Yuelin Bai, Yuhan Zhang, Yuxiang Zhang, Zenith Wang, Zhenzhu Yang, Zijian Zhao, Jiajun Zhang, Wanli Ouyang, Wenhao Huang, and Wenhu Chen. 2024. Map-neo: Highly capable and transparent bilingual large language model series. *arXiv preprint arXiv: 2405.19327*.

A Example Prompt for GIST with a Length of 256 Tokens

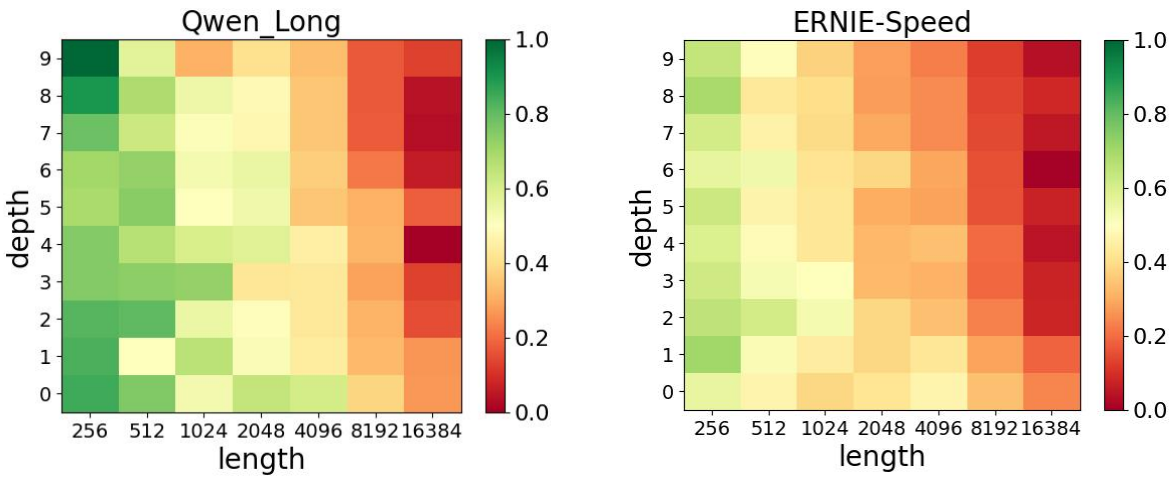
Prompt Example (256)

You are an excellent grading teacher, and you will be given many completed questions. The requirements for these questions are as follows:
[In this task, you need to answer the given multiple-choice question on the physics. Classify your answers into 'a', 'b', 'c', 'd', and 'e'.]
Please identify all the questions that have been answered incorrectly according to the requirements.
[0:Problem: how many seconds does sandy take to cover a distance of 700 meters , if sandy runs at a speed of 18 km / hr? Options: a.100 , b.120 , c.140 , d.160 , e.180 Answer: d]
[1:Problem: a train 300 m long takes 9 sec to cross a man walking at 3 kmph in a direction opposite to that of the train . find the speed of the train? Options: a.100 kmph , b.90 kmph , c.120 kmph , d.117 kmph , e.25 kmph Answer: d]
[2:Problem: a 300 m long train crosses a platform in 39 sec while it crosses a signal pole in 9 sec . what is the length of the platform ? Options: a.389m , b.350m , c.289m , d.799m , e.1000m Answer: e]
The above text contains a description of the task for this group, as well as many question-answer pairs. Each enclosed in brackets. You are requested to identify which answer(s) are incorrect and to directly provide the question number(s) of the incorrect answer(s) enclosed in brackets. Please note, I need the numbers of the questions that are incorrect, do not provide the numbers of the questions with correct answers. Additionally, ensure to output the question numbers in the specified standard format.
Answer:

Table 7: An example of a complete prompt under the length of 256 tokens.

B HotMap of More Models

The 2D length-depth distribution map for more models. The horizontal axis represents the total length of the samples, while the vertical axis indicates the position of incorrect answers within the questions, where 0 denotes the beginning and 9 denotes the end. The greener the color, the higher the model's recognition accuracy at that position.



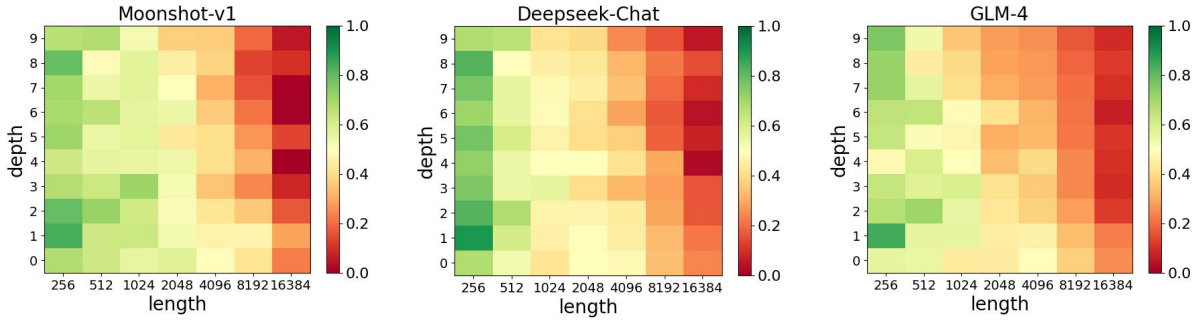


Figure 5: Expanding the Model's Heatmap.

C Prompt Template

C.1 Template of GIST

Prompt Template

You are an excellent grading teacher, and you will be given many completed questions. The requirements for these questions are as follows:
[The Task Description]
Please identify all the questions that have been answered incorrectly according to the requirements.
[Test Paper]
The above text contains a description of the task for this group, as well as many question-answer pairs. Each enclosed in brackets. You are requested to identify which answer(s) are incorrect and to directly provide the question number(s) of the incorrect answer(s) enclosed in brackets. Please note, I need the numbers of the questions that are incorrect, do not provide the numbers of the questions with correct answers. Additionally, ensure to output the question numbers in the specified standard format.
Answer:

Figure 6: The prompt template used for GIST in *LongIns*.

C.2 Template of LIST

Prompt Template

You are an excellent grading teacher, and you will be given many completed questions.
[The Task Description,Question 1]
[The Task Description,Question 2]
[The Task Description,Question 3]
[The Task Description, ...]
[The Task Description,Question N]
The above text contains a description of the task for this group, as well as many question-answer pairs. Each enclosed in brackets. You are requested to identify which answer(s) are incorrect and to directly provide the question number(s) of the incorrect answer(s) enclosed in brackets. Please note, I need the numbers of the questions that are incorrect, do not provide the numbers of the questions with correct answers. Additionally, ensure to output the question numbers in the specified standard format.
Answer:

Figure 7: The prompt template used for LIST in *LongIns*.