Leave No Document Behind: Benchmarking Long-Context LLMs with Extended Multi-Doc QA

Anonymous ACL submission

Abstract

Long-context modeling capabilities of Large Language Models (LLMs) have garnered widespread attention, leading to the emergence of LLMs with ultra-context windows. Meanwhile, benchmarks for evaluating long-context language models are gradually catching up. However, existing benchmarks employ irrelevant noise texts to artificially extend the length of test cases, diverging from the real-world scenarios of long-context applications. To bridge this gap, we propose a novel long-context benchmark, **Loong**, aligning with realistic scenarios through extended multi-document question answering (QA). Unlike typical document QA, in Loong's test cases, each document is relevant to the final answer, ignoring any document will lead to the failure of the answer. Furthermore, Loong introduces four types of tasks with a range of context lengths: *Spotlight Locating*, *Comparison*, *Clustering*, and *Chain of Reasoning*, to facilitate a more realistic and comprehensive evaluation of long-context understanding. Extensive experiments indicate that existing long-context language models still exhibit considerable potential for enhancement. Retrieval augmented generation (RAG) achieves poor performance, demonstrating that Loong can reliably assess the model's long-context modeling capabilities.

1 Introduction

Large Language Models (LLMs) have exhibited remarkable proficiency in diverse downstream applications (OpenAI, 2023). Recent works focus on scaling up the context window of LLMs (Xiong et al., 2023; Peng et al., 2023; Chen et al., 2024b), which is crucial for LLMs in handling complex tasks that require delving deeply into long texts. A few of LLM (e.g. GPT4o, Gemini-Pro) websites have been equipped with the intelligent document analysis function, allowing users to upload documents for answering queries. Meanwhile, retrieval-augmented generation (RAG) have been a commonly used framework that prompts LLMs with multiple relevant retrieved contents and can significantly improve model performance (Wu et al., 2024; Chen et al., 2024a). These demand the model leverage its long-context capability to conduct an in-depth analysis of multiple long documents.

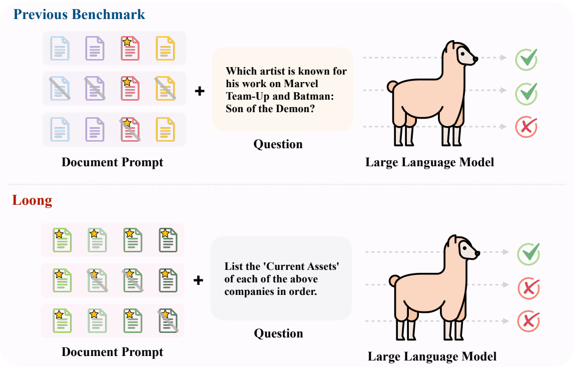

Previous Benchmark
Document Prompt
+
Which artist is known for his work on Marvel Team-Up and Batman: Son of the Demon?
Question
Large Language Model
Loong
Document Prompt
+
List the 'Current Assets' of each of the above companies in order.
Question
Large Language Model

Figure 1: Previous benchmarks vs. Loong. ☆ marks the existence of evidences related to the answer in that document. Compared to ***centralized*** distribution in previous ones, evidences in Loong are ***scattered*** in different parts across multi-document long contexts, necessitating that no document can be ignored for success.

However, there remains a lack of appropriate benchmarks for evaluating long-context understanding in real-world multi-document scenarios. Multi-document input as long-context modeling possesses extensive application scenarios of LLMs, such as analysis of financial reports over the years. Nevertheless, most existing benchmarks only place emphasis on single-document long contexts (An et al., 2023; Li et al., 2023; Kamradt, 2023) or involve multi-document question answering settings by adding distracting information to the input of existing short-context QA datasets (Hsieh et al., 2024). As shown in Figure 1, evidences supporting the answer in previous benchmarks are relatively centralized, such as being contained within a single document. Yet, such a centralized distribution of

Benchmark	Multi-doc Tasks	Broad Length Sets	Avoidance of Contamination	Realistic Scenarios	High Evidence Dispersion	Multilingual
L-Eval (An et al., 2023)	✗	✗	✗	✓	✗	✓
LongBench (Bai et al., 2023b)	✓	✗	✗	✓	✗	✓
Marathon (Zhang et al., 2023)	✓	✓	✗	✓	✗	✓
LooGLE (Li et al., 2023)	✗	✗	✗	✓	✗	✓
InfiniteBench (Zhang et al., 2024)	✓	✗	✗	✓	✗	✓
RULER (Hsieh et al., 2024)	✓	✓	✗	✗	✗	✗
NIAH (Kamradt, 2023)	✗	✓	✓	✗	✗	✗
Loong (Ours)	✓	✓	✓	✓	✓	✓

Table 1: Characteristics of Loong, where the evidences are scattered across multi-document long contexts.

evidences may cause the model to overlook certain documents and take shortcuts to formulate an answer, complicating the modeling of the real context length. Moreover, the prevalent evaluation tasks, such as "*needle in a haystack*" (NIAH) (Kamradt, 2023), only scratch the surface of long-context understanding by searching from context, far from real-world demands.

We commence with "`leave no document behind`" and scatter the evidences across multi-document long contexts. In this context, bypassing any document will lead to an erroneous answer, which better tests the long-context modeling ability. To this end, this paper develops Loong, an innovative benchmark crafted to evaluate the long-context ability of LLMs across multiple documents in real-world scenarios. Loong typically consists of 11 documents per test instance on average, spanning across three real-world scenarios in English and Chinese: (1) Financial Reports, (2) Legal Cases, and (3) Academic Papers. Meanwhile, Loong introduces new evaluation tasks from the perspectives of spotlight locating, comparison, clustering, and chain of reasoning. Furthermore, Loong features inputs of varying lengths (e.g., 10K-50K, 50K-100K, 100K-200K, >200K) and evaluation tasks of different difficulty levels, enabling fine-grained assessment of LLMs across different context lengths and task complexities.

We conduct extensive experiments on Loong to test the long-context modeling capabilities of serveral advanced LLMs. The empirical results show that even the current most powerful LLMs still struggle with the tasks in Loong, suggesting significant room for improvement in current LLMs. Furthermore, this paper conducts in-depth analyses regarding the behavior of long-context LLMs, involving RAG and the scaling law of context size. Our main contributions are summarized as follows:

- Loong primarily focuses on testing long-context ability of LLMs across multiple documents by scattering the evidences to examine the real length of long contexts.

- Loong provides evaluation sets with varying lengths of input and different levels of task difficulties, covering new task categories and common application scenarios.

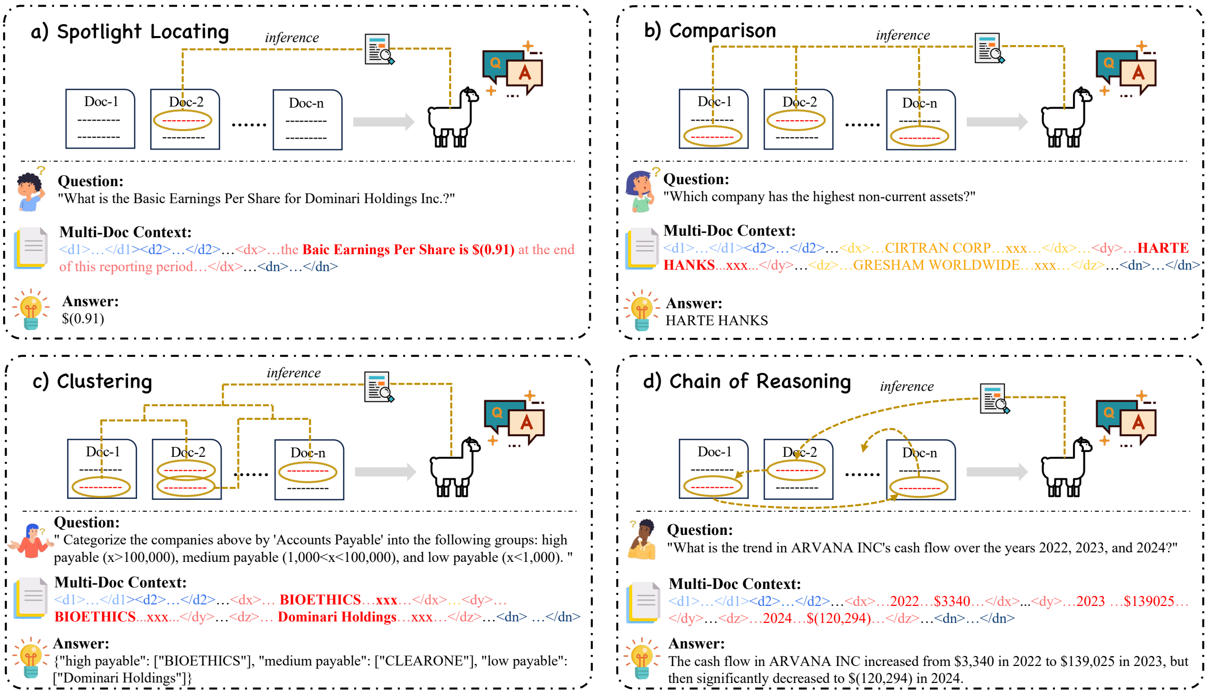
- All test instances are newly annotated and checked to guarantee the quality. Extensive experiments and analyses deeply unveil the long-context modeling abilities of LLMs.

2 Related Work

2.1 Long-Context Language Models

With support for increasingly larger context windows, closed-source LLMs have taken the lead in the field of long-context modeling. From 128k to 1000k, GPT4-Turbo-128k, Claude3-200k (Anthropic, 2024) and Gemini-1.5pro-1000k (Reid et al., 2024) are capable of modeling increasingly longer documents, expanding the new scenarios that LLMs can handle.

Considering the quadratic complexity of Transformer (Vaswani et al., 2017), training LLMs with extensive context windows from scratch necessitates substantial computational resources, exceeding the capabilities of the general researchers. Consequently, recent studies have explored ways to expand the context length of these models during the fine-tuning stage. For example, PI (Chen et al., 2023), NTK-aware (bloc97, 2023), YaRN (Peng et al., 2023), Giraffe (Pal et al., 2023), Code LLaMA (Roziere et al., 2023), and PoSE (Zhu et al., 2023) adapts position embedding based on the rotary position encoding (RoPE) (Su et al.,



a) Spotlight Locating

Question:
"What is the Basic Earnings Per Share for Dominari Holdings Inc.?"

Multi-Doc Context:
<d1>…</d1><d2>…</d2>…<dx>…the **Baic Earnings Per Share is $(0.91)** at the end of this reporting period…</dx>…<dn>…</dn>

Answer:
$(0.91)

b) Comparison

Question:
"Which company has the highest non-current assets?"

Multi-Doc Context:
<d1>…</d1><d2>…</d2>…<dx>…CIRTRAN CORP…xxx…</dx>…<dy>…**HARTE HANKS**...xxx...</dy>…<dz>…GRESHAM WORLDWIDE…xxx…</dz>…<dn>…</dn>

Answer:
HARTE HANKS

c) Clustering

Question:
" Categorize the companies above by 'Accounts Payable' into the following groups: high payable (x>100,000), medium payable (1,000<x<100,000), and low payable (x<1,000). "

Multi-Doc Context:
<d1>…</d1><d2>…</d2>…<dx>… **BIOETHICS**…**xxx**…</dx>…<dy>… **BIOETHICS**...xxx...</dy>…<dz>… **Dominari Holdings**…xxx…</dz>…<dn> …</dn>

Answer:
{"high payable": ["BIOETHICS"], "medium payable": ["CLEARONE"], "low payable": ["Dominari Holdings"]}

d) Chain of Reasoning

Question:
"What is the trend in ARVANA INC's cash flow over the years 2022, 2023, and 2024?"

Multi-Doc Context:
<d1>…</d1><d2>…</d2>…<dx>…2022…$3340…</dx>...<dy>…2023 …$139025…</dy>…<dz>…2024…$(120,294)…</dz>…<dn>…</dn>

Answer:
The cash flow in ARVANA INC increased from $3,340 in 2022 to $139,025 in 2023, but then significantly decreased to $(120,294) in 2024.

Figure 2: Showcase of four evaluation tasks in Loong (<di>...</di> marks the content of the i-th document). a) *Spotlight Locating*: Locate the evidences. b) *Comparison*: Locate and compare the evidences. c) *Clustering*: Locate and cluster the evidences into groups. d) *Chain of Reasoning*: Locate and organize the chain of the evidences.

2024), with only a few fine-tuning steps, the context length can be efficiently extended.

Another strong baseline for long-context modeling is the sliding window method. Various sliding window-based variants such as ALibi (Press et al., 2021), xPos (Sun et al., 2022), PCW (Ratner et al., 2022), LM-Infinit (Han et al., 2023), StreamingLLM (Xiao et al., 2023) are used to achieve efficient context scaling. Yet they diverge from the global perception characteristic of the Transformer, failing to exploit the entire context.

2.2 Long-Context Benchmarks

Long-context modeling methods are rapidly evolving, yet the quality of existing benchmarks does not align with this progress. Synthetic task such as Needle-in-a-Haystack (NIAH) (Kamradt, 2023) and Counting stars (Song et al., 2024) are initially utilized for evaluating long-context language models (LCLMs) due to their lower construction costs, but they are indicative of only a surface form of long-context understanding.

Longbench (Bai et al., 2023b), LooGLE (Li et al., 2023) and Marathon (Zhang et al., 2023) are earlier benchmarks for comprehensive assessment of long context. However, the average length for most tasks is between 5k and 25k, far less than the window size of LCLMs. L-Eval (An et al., 2023), BAMBOO (Dong et al., 2023), CLongEval (Qiu et al., 2024) and InfiniteBench (Zhang et al., 2024) contain sufficiently long evaluation data, and the wide variety of tasks makes the assessment more comprehensive. RULER (Hsieh et al., 2024) creates a comprehensive testing method with flexibly adjustable length and difficulty, yet they only add distracting information to the input of existing short-context QA datasets.

While these long-context benchmarks have their own advantages, we still lack a benchmark that is sufficiently long, free from data contamination (Golchin and Surdeanu, 2023), and fully aligned with the real-world multi-document question answering scenario.

2.3 Retrieval Augmented Language Models

Leveraging long documents as external knowledge, Retrieval Augmented Language Models (RALMs) has achieved comparable or even better performance than LCLMs fine-tuned for specific tasks with long document. In previous study, RALMs could directly utilize the content retrieved during the inference phase. REPLUG (Shi et al., 2023) treats the language model as a black box and the retrieval component as an adjustable plug-and-play module. RETRO (Borgeaud et al., 2022) use a chunked cross-attention module to incorporate the retrieved text. Additionally, Xu et al. (2023) explored whether RALMs or LCLMs are more suitable for long-context tasks under a larger parameter setting. However, there is currently a lack of analysis on what tasks RALMs and LCLMs each excel at, thus making it difficult to determine which type a black box model belongs to.

3 Loong: A Long-Context Benchmark

3.1 Overview

The Loong benchmark comprises tasks across four categories: *Spotlight Locating*, *Comparison*, *Clustering*, and *Chain of reasoning*. To align with realistic scenarios, we collect documents from three domains: *financial reports*, *academic papers*, and *legal cases*. Furthermore, all tasks are presented in the question-answering format, which are all newly annotated by GPT-4o and humans. Totally, Loong includes 1600 test instances in both Chinese and English, featuring four sets with different intervals of context size: `Set1` (10-50K), `Set2` (50-100K), `Set3` (100-200K) and `Set4` (200-250K). We use `tiktoken`[1] tokenizer to tokenize the input and report the number of tokens. Table 2 and Appendix B show the details of data statistics. The following sections will provide a detailed description of evaluation task and benchmark construction.

3.2 Evaluation Task

Based on various multi-document semantic relationships and LLMs' handling of multi-document input, we propose new task categories for multi-document long-context modeling and closer alignment with real-world scenarios. Figure 2 illustrates the evaluation tasks of Loong benchmark. Appendix C shows the detailed test case and prompt of each task.

3.2.1 Spotlight Locating

Spotlight locating task is designed to assess the model's capability for knowledge localization, which constitutes the foundation ability of long-context processing. In this task, the evidences are contained in only one of multiple documents, which is the atomic setting of the key information locating. Spotlight locating task is aimed at examining the LLMs' ability to search the evidences within one document from multiple ones. The upper left of figure 2 provides an example of the spotlight locating task about *Financial Reports*.

Category	**Avg Token**	**Language**	**#Test Instance**
Task			
Spotlight Locating	119.3K	EN, ZH	250
Comparison	110.6K	EN, ZH	300
Clustering	109.8K	EN, ZH	641
Chain of Reasoning	103.9K	EN, ZH	409
Sub Task			
Sequential Enumeration	103K	EN, ZH	87
Extremum Acquisition	115K	EN, ZH	143
Range Awareness	111K	EN, ZH	70
Report Integration	117K	EN, ZH	250
Citation&Reference	105K	EN	270
Case Classification	106K	ZH	121
Temporal Analysis	112K	EN, ZH	100
Citation Chain	91K	EN	130
Link the Links	117K	ZH	113
Solitaire	94K	ZH	66
Domain			
Financial Report	117.5K	EN, ZH	700
Legal Case	107.2K	ZH	500
Academic Paper	100.9K	EN	400
Length Set			
`Set1` (10-50K)	37.8K	EN, ZH	323
`Set2` (50-100K)	75.6K	EN, ZH	564
`Set3` (100-200K)	138.9K	EN, ZH	481
`Set4` (200-250K)	233.9K	EN, ZH	232

Table 2: Data statistics of Loong benchmark.

3.2.2 Comparison

Comparison task is primarily aimed at evaluating the model's ability to compare multi-source information with long contexts. In this event, the evidences supporting the answer are distributed across multiple documents, testing the LLMs' ability to locate dispersed evidences, and to correlate and compare them.

Comparison task includes three sub-tasks: 1) *Sequential Enumeration*: Based on the concrete numerical value of a specific attribute, it requires the model to list all specific values corresponding to that attribute across multiple documents in a given order. 2) *Extremum Acquisition*: It requires the model to deduce the extremum of all values corresponding to the certain attribute in multiple documents. 3) *Range Awareness*: Given a specific numerical or conceptual range, the model should output all objects within multiple documents that meet the condition. The upper right of figure 2 gives an example of comparison task.

3.2.3 Clustering

Clustering task entails an assessment of the model's ability to cluster key information based on specific conditions across multi-document long contexts. This task claims that LLMs cluster relevant evidences scattered in multiple documents based on the specified criteria. Furthermore, it necessitates

[1] https://platform.openai.com/tokenizer

the extraction of pertinent information from documents and the integration of these information by grouping according to conditions.

Clustering task encompasses three sub-tasks: 1) *Report Integration*: This sub-task requires the model to group the evidences existing in the provided financial reports into corresponding sets based on textual or numerical criteria. 2) *Citation&Reference*: For a given paper, the model is tasked with identifying its citations and references from the candidate papers. 3) *Case Classification*: Given the causes of several legal cases, the model is required to accurately categorize judgment documents. The bottom left of figure 2 depicts an example of the clustering task.

3.2.4 Chain of Reasoning

Chain of reasoning task requires the model to engage in multi-document reasoning along a logical pathway. This task evaluates the model's proficiency in logical reasoning, which requires LLMs to locate the corresponding evidences within multiple documents and model the logical relationships among them for deducing the answer.

Chain of reasoning task contains four sub-tasks: 1) *Temporal Analysis*: This task requires the model to analyze the changes or trends of a particular attribute based on the temporal relationship, such as taking into account the financial reports of a certain company over consecutive years or multiple quarters. 2) *Citation Chain*: This task requires the model to accurately understand each paper's content and their interconnections, ultimately inferring the linear citation relationships among them. 3) *Link the Links*: This task involves presenting fact descriptions and trial results from different judgment documents separately. The model is tasked with accurately pairing each fact description with its corresponding trial result. 4) *Solitaire*: This task first requires the model to match causes of action with judgment documents correctly, and then to sequentially infer multiple judgment documents based on the given sequence of causes of action. The bottom right of figure 2 gives an example of the chain of reasoning task.

3.3 Benchmark Construction

3.3.1 Data Collection

We established six criteria for the manual collection of the required English and Chinese documents: (1) *Timeliness*: The majority of the documents are the latest ones from the year 2024; (2) *Accessibility*: The data is publicly available and permitted for download and collection; (3) *Appropriate Length*: Collecting longer documents as much as possible and ensure they fit within the four designated length sets; (4) *Parseability*: Chosen documents are easy to process and parse, facilitating conversion into natural language text; (5) *Categorizability*: Documents can be manually sorted based on certain attributes, such as case type, research theme, or company category, allowing for organized archival; (6) *Authoritativeness*: All documents are collected from scratch from official websites (e.g. China Judge Online[2], U.S. SEC[3], cninf[4], Arxiv[5], Semantic Scholar[6]), ensuring the quality and authority of the documents.

Specifically, regarding financial reports, we primarily collect the latest quarterly and annual reports for the year 2024, totaling 574 documents. For legal documents, our collection consists exclusively of cases adjudicated by the higher and intermediate courts in 2024, amounting to 629 documents. As for academic papers, our focus is on procuring the latest articles from arXiv in 2024, with a total of 764 papers. Additionally, to meet requirements of the chain of reasoning task, we gather a small portion of financial reports and academic papers from before 2024. Upon the collection of documents, we first parse these documents, converting them uniformly into TXT format. Subsequently, we carry out further data cleansing, removing any portions that contain personal information.

3.3.2 Annotation Process

Compared to annotating short texts, annotating long texts is more challenging. To address this issue, we designed innovative annotation workflows to reduce the cost of annotation while ensuring the quality.

For *financial reports*, we compress the information contained within the long context, breaking down the annotation process into numerous simple tasks. We initially manually identify hundreds of key attributes which cover the important information in the long context. Subsequently, we employ GPT-4o to execute the relatively simple task of information extraction, pulling the values corresponding to these key attributes. After obtaining

[2] https://wenshu.court.gov.cn/
[3] https://www.sec.gov/
[4] http://www.cninfo.com.cn/
[5] https://arxiv.org/
[6] https://www.semanticscholar.org/

Model	Spotlight Locating		Comparison		Clustering		Chain of Reasoning		Overall	
GPT4o (128K)	73.95	0.62	50.50	0.28	44.29	0.09	57.95	0.28	53.47	0.26
Gemini-Pro1.5 (1000K)	75.02	0.56	49.94	0.27	44.10	0.09	64.97	0.37	55.37	0.27
Qwen2-72B-Instruct (128K)	54.17	0.36	42.38	0.20	36.71	0.04	47.76	0.18	43.29	0.15
Claude3-Haiku (200K)	68.68	0.59	42.10	0.21	35.04	0.02	47.59	0.17	44.88	0.19
Kimi-Chat (200k)	60.98	0.50	34.74	0.13	28.76	0.04	38.52	0.15	37.49	0.16
GLM4-9B-Chat (1000K)	57.35	0.47	40.38	0.20	28.52	0.02	39.94	0.16	38.31	0.16

Table 3: Overall results (%) on four evaluation tasks. For each task, the indicator on the left represents the ***Avg scores*** **(0~100)**, while the right one represents the ***perfect rate*** **(0~1)**.

the key attributes and their corresponding values, we can proceed to annotate only the compressed information, eliminating the need to refer back to the original lengthy texts. For *legal cases*, we follow the classification provided by China Judge Online, manually downloading judgment documents sorted by different causes of action and case types. Additionally, we use a rule-based method to segment each judgment document into its factual statement and verdict sections. For *academic papers*, we leverage the Semantic Scholar website's API to access the target paper's citations and references. Moreover, by utilizing the bbl files of each arXiv paper, we write scripts to recursively collect articles that meet the requirements of the linear citation chain task.

During the question-and-answer annotation phase, we adopt two approaches: (1) *Template-based*: We design question types and templates, and based on pre-classified documents, we construct Q&A pairs using rules. (2) *Free annotation*: Referring to the compressed information of multiple documents, we design prompts with four different task descriptions. We employ GPT-4o to generate Q&A pairs for each task.

3.3.3 Quality Control

Throughout the annotation process, we employ several methods to ensure accuracy: (1) *Evidence Recall*: By designing prompts that not only prompt GPT-4o to generate labels but also to recall evidence supporting the labels from the text, significantly enhancing the accuracy in practical applications. (2) *Self-Check*: GPT-4o reviews the original text to re-evaluate and correct any mistakes in the generated labels. (3) *Manual Check*: We manually review and confirm the quality of annotations, eliminating any unreasonable or low-quality questions. Additionally, we also take into account the distribution and number of different length intervals, sub-questions, and language types. From a pool of 2,814 entries, we conduct a secondary selection process, ultimately choosing 1,600 entries for our final benchmark.

4 Experiments

4.1 Experimental Setup

Models We evaluate six advanced long-context LLMs, with their context window sizes ranging from 128K to 1000K, including API-based LLMs: GPT-4o-128K (OpenAI, 2023), Gemini-Pro1.5-1000K (Reid et al., 2024), Claude3-Haiku-200K (Anthropic, 2024), Kimi-Chat-200K[7] and Open-sourced LLMs: Qwen2-72B-Instruct-131K (Bai et al., 2023a), GLM4-9B-Chat-1000K (Du et al., 2022).

Evaluation Metric In the long-context question answering scenarios, traditional evaluation metrics F1 and Rouge-L may lead to inaccurate responses. Recent research (Liu et al., 2024; Wang et al., 2024) indicates that the GPT-4 evaluator demonstrates high consistency with human evaluations, making it a reasonably reliable annotator. Building on these considerations, we prompt GPT-4 as a judge to evaluate the model's output based on the golden answer and the question's requirements from three aspects: *Accuracy*, *Hallucinations*, and *Completeness*, scoring from 0 to 100. For a detailed prompt, please refer to the appendix A. We also design two indicators: (1) *Avg Scores*: the average value of scores given by GPT-4 for all questions; (2) *Perfect Rate*: the proportion of cases scoring 100 out of the total cases. The latter is a more stringent evaluation metric compared to the former.

Prompt Templates For different sub-tasks, we require the model to follow the given instructions and output the answer according to the specific prompts shown in appendix C.

Input Truncation Due to input length limits, we assess whether adding a document would exceed

[7] https://kimi.moonshot.cn/

Model	Spotlight Locating		Comparison		Clustering		Chain of Reasoning		Overall	
					Set1 **(10K-50K)**					
GPT4o (128K)	85.67	0.81	64.27	0.33	57.01	0.24	81.58	0.55	70.40	0.44
Gemini-Pro1.5 (1000K)	75.00	0.60	54.88	0.28	56.15	0.23	70.64	0.37	63.36	0.34
Qwen2-72B-Instruct (128K)	68.49	0.55	60.60	0.37	47.08	0.08	70.39	0.36	60.11	0.29
Claude3-Haiku (200K)	60.94	0.55	59.97	0.40	45.53	0.04	66.85	0.34	57.14	0.28
Kimi-Chat (200k)	81.11	0.74	46.70	0.20	47,84	0.07	53.77	0.17	55.02	0.24
GLM4-9B-Chat (1000K)	63.11	0.53	54.10	0.27	39.50	0.08	56.32	0.28	51.43	0.25
					Set2 **(50K-100K)**					
GPT4o (128K)	86.76	0.72	59.81	0.40	47.83	0.11	62.09	0.34	58.38	0.29
Gemini-Pro1.5 (1000K)	76.50	0.57	54.51	0.34	44.58	0.09	64.87	0.34	55.56	0.26
Qwen2-72B-Instruct (128K)	64.53	0.43	42.60	0.21	38.52	0.05	51.18	0.20	45.71	0.17
Claude3-Haiku (200K)	73.71	0.66	41.90	0.22	36.18	0.02	50.20	0.15	45.45	0.17
Kimi-Chat (200k)	72.82	0.52	46.77	0.21	33.46	0.06	40.51	0.15	42.40	0.16
GLM4-9B-Chat (1000K)	65.04	0.54	41.80	0.23	30.72	0.02	42.34	0.17	40.19	0.17
					Set3 **(100K-200K)**					
GPT4o (128K)	74.84	0.65	42.40	0.21	38.70	0.04	45.06	0.09	46.95	0.19
Gemini-Pro1.5 (1000K)	81.25	0.56	44.66	0.20	39.90	0.05	58.38	0.36	52.05	0.24
Qwen2-72B-Instruct (128K)	46.99	0.27	37.06	0.13	31.50	0.02	35.01	0.07	35.94	0.09
Claude3-Haiku (200K)	77.81	0.67	37.07	0.17	30.94	0.01	36.87	0.12	41.41	0.18
Kimi-Chat (200k)	62.13	0.54	24.20	0.05	21.98	0.01	31.02	0.14	31.37	0.14
GLM4-9B-Chat (1000K)	69.19	0.56	37.99	0.18	26.63	0.01	32.30	0.09	37.36	0.16
					Set4 **(200K-250K)**					
GPT4o (128K)	36.79	0.19	23.97	0.08	30.40	0.00	32.89	0.07	31.11	0.07
Gemini-Pro1.5 (1000K)	62.23	0.49	43.08	0.20	36.48	0.00	68.51	0.49	50.70	0.25
Qwen2-72B-Instruct (128K)	33.18	0.16	26.59	0.08	29.84	0.01	25.81	0.04	28.92	0.06
Claude3-Haiku (200K)	53.26	0.40	27.00	0.03	25.36	0.00	28.11	0.05	32.15	0.10
Kimi-Chat (200k)	20.17	0.12	9.17	0.00	5.65	0.00	22.61	0.11	13.50	0.05
GLM4-9B-Chat (1000K)	15.67	0.12	21.33	0.05	12.35	0.00	21.04	0.05	16.84	0.05

Table 4: The performance of LLMs on four evaluation tasks with different length sets. For each task, the indicator on the left represents the ***Avg scores*** **(0~100)**, while the right one represents the ***perfect rate*** **(0~1)**.

the model's processing length when concatenating multiple documents. If appending the document would surpass the model's capacity, we discard it from the concatenation process. The evaluation and selection process continues until we have reviewed all documents that need concatenation.

Implement Details We set 'temperature = 0' to eliminate randomness and keep other hyperparameters default. For API-Based LLMs, we directly utilize the official API for testing. Since the Kimi-Chat-200k currently does not provide an interface, we manually input content on the web. As for open-source models, we conduct experiments on a server with 8×A100 80GB.

4.2 Main Results

We assess six advanced LLMs on the Loong benchmark. The main results are shown in table 3 and 4. We can see that Gemini-Pro-1.5 shows the best overall performance, especially excelling in the processing of ultra-long context within Set3 and Set4. Its comprehensive score reached 55.37 with the perfect rate of 27%, followed by GPT-4o. Besides, the long-context modeling capacity of open-source models still falls short when compared to that of the most powerful closed-source models in the Loong. Additionally, larger-parameter models outperform their smaller counterparts within the same window size, indicating the advantages of scaling up model sizes for improved long-context modeling. The overall assessment results highlight that even the most advanced long-context LLMs currently fail to achieve passing marks, particularly in terms of the perfect rate. This suggests that there exists significant room for improvement in the long-context modeling capabilities of LLMs.

4.3 Scaling Law of Context Window

It's observed that the general performance of all models deteriorates with the increase in context size. As observed from table 4, it is apparent that for the same task, models perform well within small length sets but exhibit a notable performance decline as the length increases. This indicates that

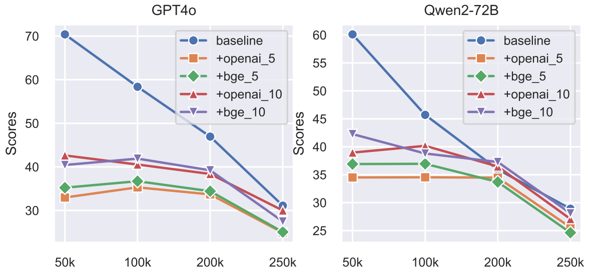


Figure 3: The experimental results of adding RAG module on GPT4o and Qwen2-72B-Instruct. The baseline means the setting without adding RAG.

the models possess a certain capability to process the task, yet their performance is constrained by the context window. Moreover, despite being trained on 128K data, the GPT4o and Qwen2-72b models begin to show performance degradation within the 50-100K interval, revealing that their actual capability boundary is significantly lower than the claimed window size. This suggests the presence of an ineffective zone within the claimed window. There exists a Scaling Law for model window sizes: to truly equip an LLM with the ability to handle 128K long texts, it should be trained on data exceeding 128K, meaning the training length should be greater than the actual processable length. Among numerous models, only the Gemini is less affected by changes in context length, which was training on the ultra-long context of 1000K. To ensure your model genuinely possesses the desired context window size, train it on longer data!

4.4 RAG or Not

We have also incorporated the Embedding RAG module into the GPT4o and Qwen2-72B-Instruct to explore whether RAG can enhance the model's performance on Loong. For the Embedding choice, We employ two distinct models: the OpenAI Embedding model and the BGE Embedding model. Regarding the configuration of default parameters, we set the top-k value of 5 and 10 for each model respectively, and the chunk size is 1024. The experimental result is shown in the Figure 3 and the detailed results can be seen in Table 7, 8, and 9. It is evident that the inclusion of RAG does not enhance the model's performance on the Loong, and there is a noticeable decline in assessment. This is because the evidence in the Loong is distributed relatively evenly across multiple documents, requiring comprehensive understanding of long texts by the model. RAG, being more limited, only shows some effectiveness in the task with sparse evidence, such as spotlight locating. However, for tasks that require a high level of comprehensiveness, RAG's negative impact is significant. The conclusion drawn from Loong suggests that, in order to truly improve a model's long-context modeling capabilities, training on longer texts is more effective than employing RAG.

4.5 Task Analysis

Analyzing performance across different tasks, models exhibit their best performance in the spotlight locating task. This can be attributed to the task's relative simplicity, which tests the foundational capabilities of long-context modeling. Moreover, the evidence is only distributed within a single document, making it easier to locate and less prone to confusion. In contrast, due to the requirements of multi-source information inference, the comparison and cluster tasks present greater challenges, leading to model underperformance. These tasks necessitate not only the collection of evidence across documents but also involve complex reasoning processes such as matching, contrasting, and classification. Thus, they more rigorously test the higher-order capabilities of long-context modeling, revealing significant gaps in the current models' abilities. Regarding the chain of reasoning task, models perform well within `Set1`. However, as the context length increases, their performance drastically declines. This suggests that within the scope of long-context modeling capabilities, LLMs possess adequate skills in temporal analysis, logical sequencing, and linking multiple concepts. Nevertheless, an overflow in context length leads to the loss of key evidence, severely impacting the accuracy of chain reasoning tasks.

5 Conclusion

In this study, we propose Loong, a benchmark for evaluating long-context understanding in real-world multi-document scenarios. We compare six advanced LLMs, including variations in their parameter sizes and context windows, along with GPT4o and Gemini-Pro1.5. Moreover, we conduct deeply analyses regarding how to improve long-context modeling capability by comparing the RAG and the scaling law of context size.

Limitations

Here we list some of the limitations that are not considered when designing Loong: (1) Limited Domains. The purpose of Loong is to evaluate the long-context understanding capabilities in real-world multi-document scenarios. However, a sea of multi-document domains exist in the real world. Considering annotation costs and model evaluation efficiency, we only cover the most representative part of them: financial, legal, and academic. (2) High Annotation Cost. To enhance the reliability of Loong in assessing the LLM's long-context understanding capabilities, we recruite a group of experts for each of the three domains to proofread the data, and they are proficient in both English and Chinese. They need to understand the question and search for relevant evidences in multiple documents with an average length of up to 110k to judge the consistency between the question and the answer, which requires a significant amount of time and effort.

References

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. *arXiv preprint arXiv:2307.11088*.

AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*.

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

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, et al. 2023b. Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*.

bloc97. 2023. Ntk-aware scaled rope. https://www.reddit.com/r/LocalLLaMA/comments/14lz7j5/ntkaware_scaled_rope_allows_llama_models_to_have/.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *Proceedings of ICML*, pages 2206–2240. PMLR.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024a. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 17754–17762.

Longze Chen, Ziqiang Liu, Wanwei He, Yunshui Li, Run Luo, and Min Yang. 2024b. Long context is not long at all: A prospector of long-dependency data for large language models. *arXiv preprint arXiv:2405.17915*.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. 2023. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*.

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. *arXiv preprint arXiv:2309.13345*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of ACL*, pages 320–335.

Shahriar Golchin and Mihai Surdeanu. 2023. Time travel in llms: Tracing data contamination in large language models. *arXiv preprint arXiv:2308.08493*.

Chi Han, Qifan Wang, Wenhan Xiong, Yu Chen, Heng Ji, and Sinong Wang. 2023. Lm-infinite: Simple on-the-fly length generalization for large language models. *arXiv preprint arXiv:2308.16137*.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. 2024. Ruler: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*.

Greg Kamradt. 2023. Needle in a haystack - pressure testing llms. https://github.com/gkamradt/LLMTest_NeedleInAHaystack.

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

Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. 2024. Calibrating LLM-based evaluator. In *Proceedings of LREC-COLING*, pages 2638–2656.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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

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

Ofir Press, Noah A Smith, and Mike Lewis. 2021. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*.

Zexuan Qiu, Jingjing Li, Shijue Huang, Wanjun Zhong, and Irwin King. 2024. Clongeval: A chinese benchmark for evaluating long-context large language models. *arXiv preprint arXiv:2403.03514*.

Nir Ratner, Yoav Levine, Yonatan Belinkov, Ori Ram, Inbal Magar, Omri Abend, Ehud Karpas, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2022. Parallel context windows for large language models. *arXiv preprint arXiv:2212.10947*.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, et al. 2023. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*.

Mingyang Song, Mao Zheng, and Xuan Luo. 2024. Counting-stars: A simple, efficient, and reasonable strategy for evaluating long-context large language models. *arXiv preprint arXiv:2403.11802*.

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

Yutao Sun, Li Dong, Barun Patra, Shuming Ma, Shaohan Huang, Alon Benhaim, Vishrav Chaudhary, Xia Song, and Furu Wei. 2022. A length-extrapolatable transformer. *arXiv preprint arXiv:2212.10554*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of NeurIPs*, page 30.

Cunxiang Wang, Sirui Cheng, Qipeng Guo, Yuanhao Yue, Bowen Ding, Zhikun Xu, Yidong Wang, Xiangkun Hu, Zheng Zhang, and Yue Zhang. 2024. Evaluating open-qa evaluation. In *Proceedings of NeurIPs*, page 36.

Kevin Wu, Eric Wu, and James Zou. 2024. How faithful are rag models? quantifying the tug-of-war between rag and llms' internal prior. *arXiv preprint arXiv:2404.10198*.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2023. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*.

Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, et al. 2023. Effective long-context scaling of foundation models. *arXiv preprint arXiv:2309.16039*.

Peng Xu, Wei Ping, Xianchao Wu, Lawrence McAfee, Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina, Mohammad Shoeybi, and Bryan Catanzaro. 2023. Retrieval meets long context large language models. *arXiv preprint arXiv:2310.03025*.

Lei Zhang, Yunshui Li, Ziqiang Liu, Junhao Liu, Min Yang, et al. 2023. Marathon: A race through the realm of long context with large language models. *arXiv preprint arXiv:2312.09542*.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Khai Hao, Xu Han, Zhen Leng Thai, Shuo Wang, Zhiyuan Liu, et al. 2024. ∞bench: Extending long context evaluation beyond 100k tokens. *arXiv preprint arXiv:2402.13718*.

Dawei Zhu, Nan Yang, Liang Wang, Yifan Song, Wenhao Wu, Furu Wei, and Sujian Li. 2023. Pose: Efficient context window extension of llms via positional skip-wise training. *arXiv preprint arXiv:2309.10400*.

A GPT4-as-the-Judge Prompt

In Loong, GPT is used as a Judger to evaluate the correctness of the model-generated content, with the prompt used shown in Text A. With this evaluation method, we expect the Judger model to output a percentage score along with its corresponding explanation.

[Gold Answer] <answer>

[The Start of Assistant's Predicted Answer]
<LLM's response>

[The End of Assistant's Predicted Answer]

[System]

We would like to request your feedback on the performance of the AI assistant in response to the user question displayed above according to the gold answer. Please use the following listed aspects and their descriptions as evaluation criteria:

- Accuracy and Hallucinations: The assistant's answer is semantically consistent with the gold answer; The numerical value and order need to be accurate, and there should be no hallucinations.
- Completeness: Referring to the reference answers, the assistant's answer should contain all the key points needed to answer the user's question; further elaboration on these key points can be omitted.

Please rate whether this answer is suitable for the question. Please note that the gold answer can be considered as a correct answer to the question.

The assistant receives an overall score on a scale of 1 to 100, where a higher score indicates better overall performance.Please note that if the assistant's answer and the gold answer fully meet the above criteria, its overall rating should be the full marks (100). Please first provide a comprehensive explanation of your evaluation, avoiding any potential bias.Then, output a line indicating the score of the Assistant.

PLEASE OUTPUT WITH THE FOLLOWING FORMAT, WHERE THE SCORE IS A SCALE OF 1 TO 100 BY STRICTLY FOLLOWING THIS FORMAT: "[[score]]", FOR EXAMPLE "Rating: [[100]]":

<start output>
Evaluation evidence: your evluation explanation here, no more than 100 words Rating: [[score]]
<end output>

Now, start your evaluation:

B Length Distribution

As shown in figure 4 and table 5, we present the distribution of data lengths in Loong. It can be observed that the data is primarily distributed around 30-120k. Moreover, we have sufficient data in both shorter and longer ranges, allowing us to assess the model's capabilities across each length interval.

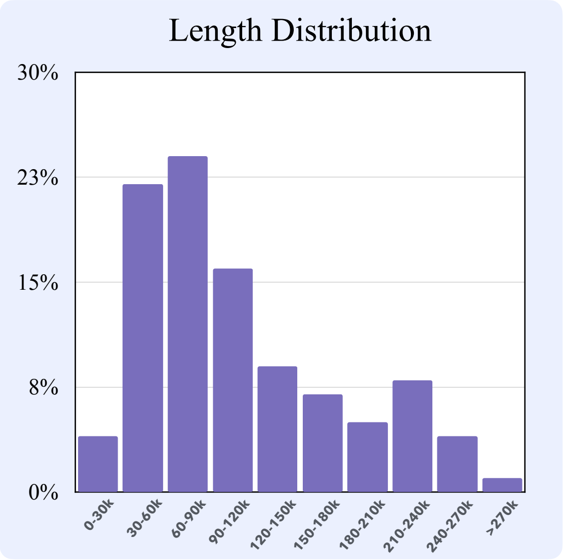


Figure 4: Test Case Length Distribution in Loong.

C Case

To facilitate understanding of Loong's data examples, we present examples of 11 sub-tasks in Table 6, showing the format we input to the model as well as the prompts we used.

D RAG Detailed Results

We conducted experiments on GPT-4o and Qwen2-72b with the addition of a RAG module. As shown in Table 7, Table 8, and Table 9, we have published detailed experimental results. It can be seen that RAG achieved subpar results on our Loong, indicating that Loong requires the model to have genuine long-context understanding capabilities.

Dataset	#data in 10-50k	#data in 50-100K	#data in 100K-200K	#data in 200-250K
Spotlight Locating	*53*	*70*	*80*	*47*
Comparison	*60*	*105*	*95*	*40*
Sequential Enumeration	24	29	20	14
Extremum Acquisition	16	55	59	13
Range Awareness	20	21	16	13
Clustering	*113*	*246*	*194*	*88*
Report Integration	40	90	90	30
Citation&Reference	37	120	79	34
Case Classification	36	36	25	24
Chain of Reasoning	*97*	*143*	*112*	*57*
Temporal Analysis	10	40	35	15
Citation Chain	33	50	41	6
Link the Links	35	25	28	25
Solitaire	28	19	8	11
Overall	*323*	*564*	*281*	*232*
Chinese	240	284	251	130
English	83	280	230	102

Table 5: Data length distributions in Loong benchmark.

Sub-Task	Test Case
Spotlight Locating	<multi_documents> <requirments> <question: What is the name of the company with accounts payable of $11,864,561?> <answer: CPI AEROSTRUCTURES INC>
Sequential Enumeration	<multi_documents> <requirments> <question: Please list the 'Changes in Undistributed Profits' of each of the aforementioned companies in descending order.> <answer: $2,095,166, $(5,441), $(415,325) in thousands compared to $(409,508) in thousands>
Extremum Acquisition	<multi_documents> <requirments> <question: Which company has the highest 'Total Current Liabilities'?> <answer: BROADWAY FINANCIAL CORP \DE\>
Range Awareness	<multi_documents> <requirments> <question: How many companies have 'Total Shares Outstanding' exceeding 10,000,000 shares?> <answer: 4 companies>
Report Integration	<multi_documents> <requirments> <instruction: Please categorize the companies listed above by 'Total Shares Outstanding' into the following groups: below 10,000,000 shares and 10,000,000 shares or more. Place companies into the same collection for the same category and into different collections for different categories.> <answer: {"below 10,000,000 shares": ["GSE SYSTEMS INC", "CROSS TIMBERS ROYALTY TRUST"], "10,000,000 shares or more": ["HUGOTON ROYALTY TRUST"]}>
Citation Reference	#Papers: <Break the Sequential Dependency of LLM Inference Using Lookahead Decoding> <Kangaroo: Lossless Self-Speculative Decoding via Double Early Exiting> <Mistral 7B> <instruction: We hope you will carefully study the provided papers and determine the citation relationships between them.> <requirments> #The paper you need to analyze: <Break the Sequential Dependency of LLM Inference Using Lookahead Decoding> <answer: {"Reference": ["# Mistral 7B"], "Citation": ["# Kangaroo"]}>
Case Classification	<multi_documents> <instruction: After reading the above judgments, please classify all the judgments according to the following three types of cases: 'Civil Cases', 'Enforcement Cases', and 'Administrative Cases'.> <requirments> <answer: {"Civil Cases": ["Judgment Document 2"], "Enforcement Cases": ["Judgment Document 4"], "Administrative Cases": ["Judgment Document 1", "Judgment Document 3"]}>
Temporal Analysis	<multi_documents> <requirments> <question: What is the trend in ARVANA INC's share capital from 2021 to 2024?> <answer: ARVANA INC's share capital has consistently increased from $4,611 in 2021 to $34,149 in 2022, $35,949 in 2023, and $107,847 in 2024.>
Citation Chain	<instruction: Given several papers, you are required to identify and list the longest citation chain, which demonstrates the citation relationship among the provided papers.> <requirments> #Paper Provided: <Understanding the Difficulty of Training Transformers> <Very Deep Transformers for Neural Machine Translation> <MonaCoBERT: Monotonic attention based ConvBERT for Knowledge Tracing> <answer: ["# Very Deep Transformers for Neural Machine Translation ", "# Understanding the Difficulty of Training Transformers ", "# MonaCoBERT: Monotonic attention based ConvBERT for Knowledge Tracing"]>
Link the Links	<multi_documents> <instruction: After reading the above judgment document, I will give you several judgment results: <a list of judgment result> You need to determine the most likely judgment result for each of the above judgment documents.> <answer: {"Judgment Document 1": "Judgment Result 1", "Judgment Document 2": "Judgment Result 6", "Judgment Document 3": "Judgment Result 2", "Judgment Document 4": "Judgment Result 5"}>
Solitaire	<multi_documents> <instruction: After reading the above judgment documents, I will provide a list of several Legal Basis arranged in order from left to right: ["Legal Basis 1", "Legal Basis 2", ..., "Legal Basis 6"]. You need to arrange all the judgment documents according to the order of the Legal Basis given above.> <answer: {"Legal Basis 1": "Judgment Document 2", "Legal Basis 2": "Judgment Document 6", "Legal Basis 3": "Judgment Document 4", "Legal Basis 4": "Judgment Document 1"}>

Table 6: Test case and prompts of each sub-task in Loong benchmark.

Model	Spotlight Locating		Comparison		Clustering		Chain of Reasoning		Overall	
Set1 **(10K-50K)**										
GPT4o (128K)	85.67	0.81	64.27	0.33	57.01	0.24	81.58	0.55	70.40	0.44
w/ Openai Embedding, Top k=5	47.60	0.31	29.75	0.10	29.10	0.06	31.46	0.08	32.98	0.11
w/ BGE Embedding, Top k=5	57.17	0.43	34.15	0.12	30.71	0.07	28.77	0.08	35.23	0.14
w/ Openai Embedding, Top k=10	61.25	0.44	38.33	0.17	37.00	0.08	41.67	0.16	42.63	0.18
w/ BGE Embedding, Top k=10	61.00	0.44	39.74	0.19	36.14	0.08	34.90	0.11	40.44	0.17
Set2 **(50K-100K)**										
GPT4o (128K)	86.76	0.72	59.81	0.40	47.83	0.11	62.09	0.34	58.38	0.29
w/ Openai Embedding, Top k=5	56.01	0.35	39.56	0.22	31.84	0.04	27.01	0.03	35.31	0.11
w/ BGE Embedding, Top k=5	67.33	0.43	43.90	0.28	29.37	0.04	27.84	0.04	36.72	0.14
w/ Openai Embedding, Top k=10	64.77	0.45	45.44	0.31	36.07	0.05	32.29	0.05	40.54	0.15
w/ BGE Embedding, Top k=10	72.07	0.52	50.15	0.32	34.35	0.05	33.49	0.07	41.90	0.17
Set3 **(100K-200K)**										
GPT4o (128K)	74.84	0.65	42.40	0.21	38.70	0.04	45.06	0.09	46.95	0.19
w/ Openai Embedding, Top k=5	67.45	0.49	29.00	0.13	25.09	0.01	27.22	0.02	33.69	0.12
w/ BGE Embedding, Top k=5	71.12	0.56	31.36	0.14	25.32	0.00	25.78	0.04	34.43	0.13
w/ Openai Embedding, Top k=10	72.37	0.55	31.41	0.13	30.59	0.01	33.14	0.08	38.38	0.14
w/ BGE Embedding, Top k=10	79.04	0.67	34.29	0.18	30.59	0.02	29.69	0.06	39.22	0.17
Set4 **(200K-250K)**										
GPT4o (128K)	36.79	0.19	23.97	0.08	30.40	0.00	32.89	0.07	31.11	0.07
w/ Openai Embedding, Top k=5	50.76	0.22	17.25	0.00	19.53	0.00	16.61	0.00	24.91	0.05
w/ BGE Embedding, Top k=5	51.02	0.26	18.75	0.03	17.83	0.00	18.77	0.02	25.07	0.06
w/ Openai Embedding, Top k=10	57.98	0.31	23.00	0.03	25.08	0.00	21.29	0.02	30.00	0.07
w/ BGE Embedding, Top k=10	51.48	0.25	23.36	0.05	22.55	0.00	18.95	0.02	27.48	0.06

Table 7: The result of adding RAG module on GPT4o with different length sets.

Model	Spotlight Locating		Comparison		Clustering		Chain of Reasoning		Overall	
Set1 **(10K-50K)**										
Qwen2-72B-Instruct (128K)	68.49	0.55	60.60	0.37	47.08	0.08	70.39	0.36	60.11	0.29
w/ Openai Embedding, Top k=5	54.62	0.45	26.17	0.08	29.60	0.03	34.41	0.08	34.51	0.12
w/ BGE Embedding, Top k=5	62.92	0.53	30.92	0.08	31.28	0.03	32.95	0.11	36.91	0.15
w/ Openai Embedding, Top k=10	59.81	0.43	34.93	0.15	29.33	0.02	41.27	0.15	38.96	0.15
w/ BGE Embedding, Top k=10	72.13	0.62	32.42	0.12	31.90	0.05	44.12	0.20	42.27	0.20
Set2 **(50K-100K)**										
Qwen2-72B-Instruct (128K)	64.53	0.43	42.60	0.21	38.52	0.05	51.18	0.20	45.71	0.17
w/ Openai Embedding, Top k=5	56.64	0.40	36.68	0.19	30.91	0.03	28.38	0.01	34.54	0.10
w/ BGE Embedding, Top k=5	67.29	0.47	43.39	0.28	28.31	0.03	32.22	0.07	36.95	0.14
w/ Openai Embedding, Top k=10	67.07	0.53	44.30	0.27	34.31	0.05	34.03	0.06	40.17	0.15
w/ BGE Embedding, Top k=10	71.74	0.54	47.68	0.30	30.55	0.03	30.57	0.03	38.80	0.14
Set3 **(100K-200K)**										
Qwen2-72B-Instruct (128K)	46.99	0.27	37.06	0.13	31.50	0.02	35.01	0.07	35.94	0.09
w/ Openai Embedding, Top k=5	63.91	0.44	33.56	0.17	25.98	0.01	28.98	0.04	34.48	0.12
w/ BGE Embedding, Top k=5	64.81	0.47	30.27	0.14	25.88	0.01	27.86	0.05	33.70	0.12
w/ Openai Embedding, Top k=10	67.50	0.46	33.44	0.16	27.94	0.02	31.62	0.06	36.47	0.13
w/ BGE Embedding, Top k=10	75.88	0.56	33.76	0.15	27.20	0.01	30.17	0.04	37.28	0.14
Set4 **(200K-250K)**										
Qwen2-72B-Instruct (128K)	33.18	0.16	26.59	0.08	29.84	0.01	25.81	0.04	28.92	0.06
w/ Openai Embedding, Top k=5	51.49	0.26	17.12	0.03	21.59	0.00	16.37	0.00	25.59	0.06
w/ BGE Embedding, Top k=5	48.40	0.26	14.55	0.00	20.69	0.00	18.07	0.00	24.63	0.05
w/ Openai Embedding, Top k=10	50.32	0.28	20.30	0.03	24.56	0.00	16.38	0.00	27.08	0.06
w/ BGE Embedding, Top k=10	51.02	0.28	21.88	0.03	25.45	0.00	17.29	0.00	28.10	0.06

Table 8: The result of adding RAG module on Qwen2-72B-Instruct with different length sets.

Model	**Spotlight Locating**		**Comparison**		**Clustering**		**Chain of Reasoning**		**Overall**	
GPT4o (128K)	73.95	0.62	50.50	0.28	44.29	0.09	57.95	0.28	53.47	0.26
w/ Openai Embedding, Top k=5	56.97	0.36	31.28	0.14	27.71	0.03	26.65	0.04	32.85	0.11
w/ BGE Embedding, Top k=5	63.32	0.44	34.63	0.17	26.74	0.03	26.21	0.04	34.01	0.13
w/ Openai Embedding, Top k=10	65.20	0.46	36.80	0.19	33.06	0.04	33.26	0.08	38.80	0.14
w/ BGE Embedding, Top k=10	68.27	0.50	39.51	0.22	31.91	0.04	30.71	0.07	38.71	0.15
Qwen2-72B-Instruct (128K)	54.17	0.36	42.38	0.20	36.71	0.04	47.76	0.18	43.29	0.15
w/ Openai Embedding, Top k=5	57.57	0.40	30.98	0.14	27.91	0.02	28.30	0.04	33.22	0.10
w/ BGE Embedding, Top k=5	62.02	0.44	32.90	0.16	27.05	0.02	29.26	0.06	34.18	0.12
w/ Openai Embedding, Top k=10	62.52	0.44	35.79	0.18	30.16	0.03	32.67	0.08	36.92	0.13
w/ BGE Embedding, Top k=10	69.24	0.51	36.78	0.18	29.07	0.02	31.90	0.07	37.50	0.14

Table 9: Overall results (%) of adding RAG module on GPT4o and Qwen2-72B-Instruct.