LooGLE: Can Long-Context Language Models Understand Long Contexts?

Jiaqi Li[1,*] , **Mengmeng Wang**[1,*], **Zilong Zheng**[1,†] , **Muhan Zhang**[1,2,†]
[1] National Key Laboratory of General Artificial Intelligence, BIGAI
[2] Institute for Artificial Intelligence, Peking University

Abstract

Large language models (LLMs) are typically limited to processing texts within context-window size, which has spurred significant research efforts into enhancing LLMs' long-context understanding as well as developing high-quality benchmarks to evaluate the ability. However, prior datasets suffer from shortcomings like short length compared to the context window of modern LLMs; outdated documents that might have data leakage problems; and an emphasis on short dependency tasks only. In this paper, we present LooGLE, a **Lo**ng C**o**ntext **G**eneric **L**anguage **E**valuation benchmark. It features documents post-2022, with over 24,000 tokens per document and 6,000 newly generated questions spanning varying dependency ranges in diverse domains. Human annotators meticulously crafted over 1,100 high-quality question-answer (QA) pairs with thorough cross-validation for a most precise assessment of LLMs' **long dependency** capabilities. We conduct a comprehensive evaluation of representative LLMs on LooGLE. The results indicate that most LLMs have shockingly bad long context ability and fail to capture long dependencies in the context, even when their context window size is enough to fit the entire document. Our results shed light on enhancing the "true long-context understanding" ability of LLMs instead of merely enlarging their context window.

1 Introduction

The pursuit of enabling large language models (LLMs), such as ChatGPT (Brown et al., 2020; Zeng et al., 2023), to go beyond their limited context window size to process, comprehend, or even learn from long-context textual information (Ding et al., 2023; Dao et al., 2022; Bulatov et al., 2023) is inevitable for next-generation of language intelligence attributed to its wide applications on real-world scenarios, such as domain-specific knowledge understanding, long-context conversational generation, long story or code generation, *etc*.

Meanwhile, there is an increasing need for high-quality benchmarks with much longer text lengths and more challenging tasks to provide comprehensive evaluations. However, traditional benchmarks (Sharma et al., 2019; Huang et al., 2021) often fall short in text length with an average number of thousand words (Yang et al., 2018). Besides, existing benchmarks automatically collect possibly outdated documents from existing datasets published a few years ago (Shaham et al., 2022; Trivedi et al., 2022; Wang et al., 2022), which might lead to data leakage (Liu et al., 2023c; Zhou et al., 2023) in pre-trained LLMs and make the evaluation inaccurate. Further, the long texts are often restricted to domain-specific articles, making it hard to evaluate LLMs' ability on generic tasks and domains. Finally, it is important to note that tasks in existing benchmarks are primarily short dependency tasks, which only require LLMs to retrieve answers from one specific sentence or paragraph, without really testing LLMs' ability to address long dependency questions.

To mitigate the shortcomings of existing datasets, in this paper, we introduce a novel benchmark **LooGLE** to evaluate the long context understanding abilities of LLMs illustrated in Fig. 1. Our benchmark has the following advantages:

- **Extra-long documents**. It contains 776 latest gathered and extremely long documents with an average of 19.3k words. There are over 6,448 test instances without distribution bias for a more generalized assessment. On one hand, they can better evaluate LLMs' capability on memorizing and understanding longer text that is far beyond their context window size. On the other hand, the excessive length is well suited to the common

---

[*] Equal contributions.

[†] Correspondence to Zilong Zheng <zlzheng@bigai.ai> and Muhan Zhang <muhan@pku.edu.cn>.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 16304–16333
August 11-16, 2024 ©2024 Association for Computational Linguistics

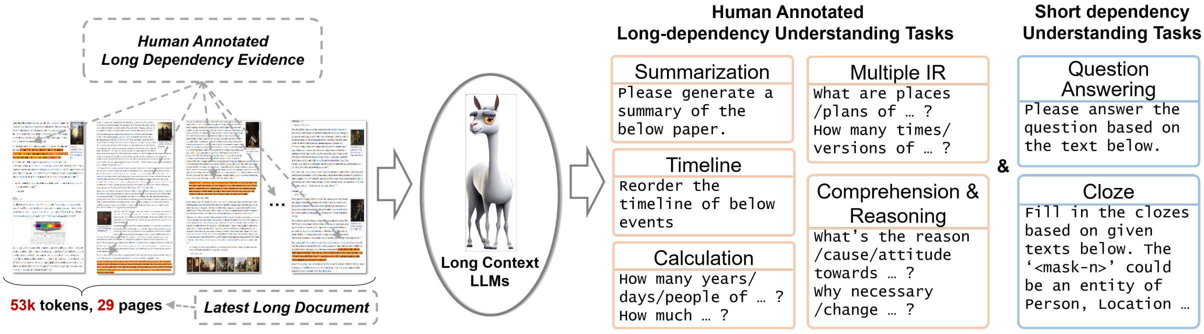

Human Annotated
Long Dependency Evidence
53k tokens, 29 pages
Latest Long Document
Long Context
LLMs
Human Annotated
Long-dependency Understanding Tasks
Summarization
Please generate a summary of the below paper.
Timeline
Reorder the timeline of below events
Calculation
How many years/days/people of … ?
How much … ?
Multiple IR
What are places/plans of … ?
How many times/versions of … ?
Comprehension & Reasoning
What's the reason/cause/attitude towards … ?
Why necessary/change … ?
&
Short dependency
Understanding Tasks
Question Answering
Please answer the question based on the text below.
Cloze
Fill in the clozes based on given texts below. The '<mask-n>' could be an entity of Person, Location …

Figure 1: **The LooGLE benchmark for long context understanding.**

usage of long text scenarios.

- **Manually designed both short and long dependency tasks. Long dependency tasks** require the understanding of the inter-dependency across various pieces of information and evidence widely spanning over the lengthy text, sometimes spanning across the whole document. Instead of directly locating the evidence through keyword matching or semantic similarity, it relies heavily on the profound comprehension of the question and its implied correlation with this evidence for extensive multi-source retrieval from the texts. It is composed of 7 major tasks to evaluate LLMs' ability to understand both short and long dependency content. We carefully designed 5 types of long dependency tasks and manually created 1,101 long dependency QA instances, despite the high costs and huge effort involved in this process.
- **Cross-domain generic data.** Our benchmark is derived from popular open-source documents, including arXiv papers, Wikipedia articles, and movie and TV scripts, spanning diverse domains and multiple categories such as academia, history, sports, politics, arts, events, and entertainment.

Evaluations on **LooGLE** revealed: (i) commercial models outperformed open-sourced models; (ii) LLMs excelled in short dependency tasks (short QA and cloze) but struggled with more intricate long dependency tasks; (iii) in-context learning and chaining thoughts offered only marginal improvements in long context comprehension; (iv) retrieval-based techniques demonstrated substantial benefits for short QA, while strategies for extending context window length had limited impact on long context understanding. Better base models with a larger context window size generally perform better. However, all models experience a significant performance decline in long-dependency tasks no matter how large their context window is, indicating the desperate need to improve the true long-dependency understanding capabilities of LLMs.

2 Related Work

Existing models for long context There are increasing research interests in extending LLMs' context window size such as utilizing recurrent memory, efficient attention, external memory and etc.(Meister et al., 2021; Xiong et al., 2023; Li et al., 2023b; Li, 2023). Improved transformer architectures (Dong et al., 2023; Tay et al., 2020, 2022) are developed to decrease memory complexity to efficiently model longer texts. For Recurrent transformer (Bulatov et al., 2022; Bessonov et al., 2023), history information of previous segments will be leveraged when the subsequent segment is fed into the model without fragmentation problem. Fine-tuned models on long documents (Wu et al., 2021) are also explored, but they are often effort-costing and face difficulties in collecting ground truth fine-tuning data for long text tasks. There are also works incorporating external memory, using compression techniques and task-oriented process optimization.

Dataset	Avg. Words	# Docs.	# Ques.	Manually Label	Long Dependency Tasks				
					Summarization	Info. Retrieval	Timeline Reorder	Calculation	Doc QA
Zero Scrolls (Shaham et al., 2023)	10,392	-	4,378	✗	✓	✗	✗	✗	✗
Long Bench (Bai et al., 2023)	8,120	-	4,750	350	✓	✓*	✗	✓*	✓
L-Eval (An et al., 2023)	8,008	411	2,043	2,043†	✓	✓	✗	✗	✓
LooGLE (Ours)	19,367	776	6,448	1,101	✓	✓	✓	✓	✓

Table 1: **Comparison with other long-context benchmarks.** * Synthetic data. † Re-labelled questions.

Dataset	No. Docs	Avg. Words	Max. Words	Min. Words	Avg. Tokens	Task	# Questions
arXiv	516	16,988	197,977	10,204	20,887	Summarization	516
Wikipedia	105	17,604	46,250	11,285	21,017	Short dependency QA Long dependency QA	1,951 459
Movie & TV scripts	155	28,483	62,752	11,089	36,412	Cloze Long dependency QA	2,880 642

Table 2: **Statistics of LooGLE**

Existing datasets for long context Table 1 shows the comparisons of LooGLE with similar benchmarks with multi-source and multi-task. ZeroSCROLLS (Shaham et al., 2023) automatically processes datasets from different sources into a unified format. However, it mainly focuses on collecting documents and tasks from existing datasets and relies on automatic metrics for limited model comparisons. L-Eval (An et al., 2023) differs in re-annotating the data and instructions from similar public datasets with smaller sizes to ensure the quality. Besides, it optimizes the evaluation procedures to get more accurate conclusions. LongBench (Bai et al., 2023) provides a bilingual and multi-task dataset featuring diverse sequences of varying lengths, distributions and patterns for a comprehensive evaluation. Nonetheless, it encompasses tasks mostly restricted to short-term information extraction. There are few types of "long dependency" tasks in previous datasets, except for summarization (which LLMs are validated to perform well on) and synthesized tasks like data aggregation and retrieving (Li et al., 2023a; Pal et al., 2023). These tasks can be simply solved by locating pieces of information from the lengthy source input followed by aggregation.

3 The LooGLE Benchmark

Our **LooGLE** benchmark consists of 3 sources: scientific papers, Wikipedia articles, movie and TV scripts, covering various topics and categories. All the documents in **LooGLE** are after 2022 and filtered by a length of over 10k words. By replicating the methodology proposed in this paper, they can be collected easily and periodically. Based on that, we generate two main types of tasks: short dependency and long dependency tasks. The generation of short dependency tasks can be seen in Appendix B. Statistics of the source documents and generated tasks can be found in Table 2. We use MIT License for LooGLE and the current version of our dataset is English-only.

3.1 Dataset selection and construction

arXiv papers We randomly select data from a massive pool after January 2022 on the arXiv website. Then we extracted their abstracts as the ground truths for summarization.

Wikipedia articles Wikipedia provides constantly evolving articles on a wide range of topics, which are perfect for evaluating the long text abilities of LLMs. We first downloaded and parsed the most recent page articles present in .bz file and then filtered the articles utilizing an open-source Wikipedia dataset (202203.en)[1] and retained those existing after page redirection.

Movie and TV scripts A movie or TV script typically contains scene & action descriptions and dialogues between characters containing numerous events and facts. It necessitates models to deeply understand the plots and navigate shifts among characters to grasp the key viewpoints. All scripts consist of movies and TV shows released after 2022 from three websites[2].

3.2 Long dependency tasks

Summarization We directly use the abstract of each paper as the reference for generating summaries. The abstracts effectively capture the main content and key information of each paper.

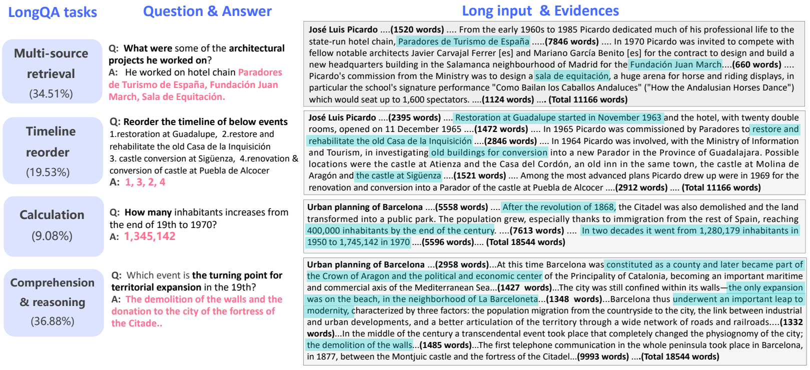
Long dependency QA One highlight of our dataset is that we dedicated significant effort to manually compile about 1.1k true long dependency QA pairs. We manually designed 4 long dependency tasks: Multi-source retrieval, Timeline reorder, Calculation, Comprehension and reasoning. These tasks are pretty challenging and valuable for understanding the limitations of LLMs. Examples of the 4 types of long dependency QAs are in Fig. 2.

Multi-source retrieval (MR): The task requires extensive information extraction from widely distributed segments within the lengthy text, followed

---

[1] https://huggingface.co/datasets/wikipedia

[2] https://www.scriptslug.com, https://thescriptlab.com/, https://8flix.com



LongQA tasks	Question & Answer	Long input & Evidences
Multi-source retrieval (34.51%)	Q: **What were** some of the **architectural projects he worked on**? A: He worked on hotel chain **Paradores de Turismo de España, Fundación Juan March, Sala de Equitación.**	**José Luis Picardo** ....**(1520 words)** .... From the early 1960s to 1985 Picardo dedicated much of his professional life to the state-run hotel chain, Paradores de Turismo de España .....**(7846 words)** .... In 1970 Picardo was invited to compete with fellow notable architects Javier Carvajal Ferrer [es] and Mariano García Benito [es] for the contract to design and build a new headquarters building in the Salamanca neighbourhood of Madrid for the Fundación Juan March....**(660 words)** .... Picardo's commission from the Ministry was to design a sala de equitación, a huge arena for horse and riding displays, in particular the school's signature performance "Como Bailan los Caballos Andaluces" ("How the Andalusian Horses Dance") which would seat up to 1,600 spectators. ....**(1124 words)** .... **(Total 11166 words)**
Timeline reorder (19.53%)	Q: **Reorder the timeline of below events** 1.restoration at Guadalupe, 2.restore and rehabilitate the old Casa de la Inquisición 3. castle conversion at Sigüenza, 4.renovation & conversion of castle at Puebla de Alcocer A: **1, 3, 2, 4**	**José Luis Picardo** ....**(2395 words)** .... Restoration at Guadalupe started in November 1963 and the hotel, with twenty double rooms, opened on 11 December 1965 ....**(1472 words)** .... In 1965 Picardo was commissioned by Paradores to restore and rehabilitate the old Casa de la Inquisición ....**(2846 words)** .... In 1964 Picardo was involved, with the Ministry of Information and Tourism, in investigating old buildings for conversion into a new Parador in the Province of Guadalajara. Possible locations were the castle at Atienza and the Casa del Cordón, an old inn in the same town, the castle at Molina de Aragón and the castle at Sigüenza ....**(1521 words)** .... Among the most advanced plans Picardo drew up were in 1969 for the renovation and conversion into a Parador of the castle at Puebla de Alcocer ....**(2912 words)** .... **(Total 11166 words)**
Calculation (9.08%)	Q: **How many** inhabitants increases from the end of 19th to 1970? A: **1,345,142**	**Urban planning of Barcelona** ....**(5558 words)** .... After the revolution of 1868, the Citadel was also demolished and the land transformed into a public park. The population grew, especially thanks to immigration from the rest of Spain, reaching 400,000 inhabitants by the end of the century. ....**(7613 words)** .... In two decades it went from 1,280,179 inhabitants in 1950 to 1,745,142 in 1970 ....**(5596 words)**.... **(Total 18544 words)**
Comprehension & reasoning (36.88%)	Q: Which event is **the turning point for territorial expansion** in the 19th? A: **The demolition of the walls and the donation to the city of the fortress of the Citade..**	**Urban planning of Barcelona** ...**(2958 words)**...At this time Barcelona was constituted as a county and later became part of the Crown of Aragon and the political and economic center of the Principality of Catalonia, becoming an important maritime and commercial axis of the Mediterranean Sea...**(1427 words)**...The city was still confined within its walls—the only expansion was on the beach, in the neighborhood of La Barceloneta...**(1348 words)**...Barcelona thus underwent an important leap to modernity, characterized by three factors: the population migration from the countryside to the city, the link between industrial and urban developments, and a better articulation of the territory through a wide network of roads and railroads....**(1332 words)**...In the middle of the century a transcendental event took place that completely changed the physiognomy of the city; the demolition of the walls...**(1485 words)**...The first telephone communication in the whole peninsula took place in Barcelona, in 1877, between the Montjuic castle and the fortress of the Citadel...**(9993 words)** ....**(Total 18544 words)**

Figure 2: Long dependency QA tasks

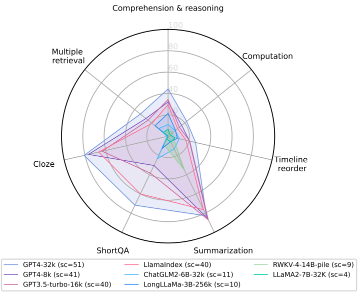


Figure 3: LLM overview performance

by the aggregation of the evidence to derive the ultimate answer. The evidence is distinctly presented and can be directly located within the original sentences of the text.

Calculation: Similar to the previous task, it firstly needs multi-source retrieval from a wide range of texts. A majority of the evidence takes the form of numerical data, often in question formats like inquiries about quantities, frequencies, durations, specific numbers, and so on. To obtain an accurate response, it relies heavily on the profound comprehension of the question and its correlation with the provided numerical data as well as a degree of mathematical reasoning ability.

Timeline reorder (TR): This task follows a more conventional format, involving the instruction, "Please reorder the timeline of the following events," along with a set of events presented in a permuted order. The objective is to arrange these events in accordance with their chronological sequence as dispersed throughout the extensive text. The events are derived directly from the source text, either as extracted segments or summarized factual information. This task necessitates either the memorization or comprehensive understanding of the central storyline of the document and assesses the model's proficiency in temporal awareness.

Comprehension and reasoning (CR): This task demands not only a profound comprehension of the question but also intricate reasoning to discern the underlying implications for the appropriate evidence. The most prevalent question patterns involve inquiries about causality, impact, contributions, attitudes, and essential attributes related to various events. More extensive comparisons are essential when the questions revolve around the primary, predominant, highest, or most critical aspects of the evidence. Furthermore, the answers to this task are not explicitly evident within the source text. They often require multi-step reasoning to model the inherent connections and dependencies, facilitating the acquisition of the answer through a complex analytical process.

3.2.1 Long dependency QA construction

We recruit students from top universities and organize a manual annotation process to generate long-dependency QAs based on 140 source documents from Wikipedia articles and scripts. Each document underwent a meticulous three-step process to ensure a rigorous cross-validation process, involving two distinct annotators (the questioner and answerer respectively) who were kept unaware of each other. It takes nearly 3 to 5 hours on average to read one long document and needs even more effort to question/answer. Taking into account the three-step annotation process and the total of 140 articles, we project that achieving the final high-quality questions, answers, and supporting evidence, along with maintaining a high degree of accuracy, precision, and relevance to the document's content, would necessitate a minimum of 3 * 3 * 140 = 1260 human-hours.

Step 1: Question and answer. The questioner's role encompassed a comprehensive set of responsibilities, including reading the document, crafting relevant questions, offering their own answers to those questions, and pinpointing the specific evidentiary passages within the document that substantiated their answers. The annotation adhered to stringent standards with the following principles:

- Long dependency: Each question was required to exhibit a long dependency whose evidence should have a wide span across the document. (eg. distance between the earliest and latest evidence is recommended to exceed 5k words.)

- Diverse problem types: Each document has an average of 5 to 10 QA pairs generated, while over 4 questions of the same type are not allowed to avoid imbalanced question distribution and prevent from generating overly simple questions.
- Clear and precise questions: The formulation of each question was asked to adhere to clarity, conciseness and no ambiguity with given examples.
- Deterministic and objective answers: The answers to the proposed questions were rigorously checked to be deterministic and objective, precluding open-ended ones.

Step 2: Answer and check. Each answerer can only access the assigned article and its posed questions in step1. The answerer was required to thoroughly read the entire document and provide answers accordingly. The standard for the answers is the same as step1. In addition, the answerer was also tasked with assessing the quality of the questions by evaluating whether the questions adhere to the standard and whether they are answerable by providing constructive feedback for improvement.

Step 3: Revise. The questioner was provided with the document, corresponding QAs from both the questioner and answerer and the feedback. The questioner was asked to first revise the questions according to the feedback, and then unify the two sets of answers to derive the final answers.

In steps 1-2, we acquired 1,137 QA pairs in total, 206 of which were non-compliant with the criteria and accompanied by suggestions for improvement from the answerers. After the revisions, we ultimately obtained 1101 high-quality long dependency QA pairs requiring strong long context understanding ability. The inter-annotator agreement (Wilkens, 2023) rate is 81.88%.

4 Benchmark LLMs on LooGLE

Commercial models GPT4-32k, GPT4-8k, GPT3.5-turbo-16k (Chen et al., 2023b) are models developed by OpenAI[3]. We use the models of version 0613 by default.

Open-source models There has also been a great effort from the open-source community to address the long context problem and we chose representative instruct-tuned models as baselines. LLaMA2-7B-32K-Instruct (Touvron et al., 2023) is fine-tuned from Meta's original Llama2-7B (Touvron et al., 2023) using Position Interpolation (Chen et al., 2023a). ChatGLM2-6B-32k (Du et al., 2022) integrates FlashAttention (Dao et al., 2022) extending to 32K context length. LongLLaMa-3B-Instruct, derived from openllama, has been fine-tuned using Focused Transformer (Tworkowski et al., 2024). RWKV-4-14B-raven (Peng et al., 2023) is notable for its architectural fusion of Recurrent Neural Networks (RNNs) to accommodate 8k context length. Based on the conclusion from Liu et al. (2023a), we artificially truncate the input document to certain sizes (all not larger than above mentioned models) by concatenating the head and tail of the input.

Retrieval-based method (Xu et al., 2023; Askari et al., 2024) augments the LLM by incorporating external memory, allowing relevant information to be retrieved using a specific query. In our experiment, we employed the data framework LlamaIndex[4] with default embedding model text-embedding-ada-002 and language model text-davinci-003 (Ouyang et al., 2022). Here we employ the same truncation strategy mentioned above.

4.1 Evaluation methods and metrics

Automatic evaluation We adopt Bleu (Papineni et al., 2002), Rouge(Lin, 2004), Meteor score (Banerjee and Lavie, 2005) and Bert Score (Zhang et al., 2019) as automatic evaluation metrics, which compute n-gram matching and semantic similarity for generative tasks. Exact Match and Partial Match (Sharma et al., 2023) are employed for cloze.

Human evaluation Human in the loop is necessary for LLM evaluation for reference, especially for free text generation tasks. We manually inspect the accuracy of ground truthted answers being consistent with the groundtruth for some models/tasks.

GPT4-as-judgment We follow (Suri et al., 2024; Liu et al., 2023b; Zheng et al., 2024) to use GPT4 as a reliable substitute for human evaluation. We make further analysis on the alignment between GPT4 evaluator and human in Table 5. As seen, GPT4 exhibits high consistency with human evaluation. Besides, we make our implementation reproducible and GPT4's judgment deterministic by setting its temperature to 0, top_p to 1, and prompting GPT4 to output True/False/exact-score only, instead of descriptive results. From our observations in the experiment results, we found that the GPT4 evaluator has no bias when scoring. How we use GPT4-eval is also delicately designed to ensure fairness. By

---

[3] https://platform.openai.com/docs/models

[4] https://github.com/jerryjliu/llama_index

Models	Context	Short dependency QA Bleu1	Bleu4	Rouge1	Rouge4	RougeL	Meteor_score	Bert_score	GPT4_score	Cloze Exact Match	Partial Match
GPT4-32k	32k	24.61	11.14	61.80	50.73	60.75	32.94	78.72	**71.52**	**70.5**	**80.81**
GPT4-8k	8K	27.35	14.38	**67.59**	**56.01**	**65.77**	**38.56**	**87.93**	53.99	66.03	76.62
GPT3.5-turbo-16k	16K	22.67	9.62	62.56	48.63	60.66	32.58	87.04	66.82	54.64	63.42
LlamaIndex	\	**33.37**	**21.43**	58.82	42.93	57.08	37.17	86.58	59.61	58.95	66.86
ChatGLM2-6B	32k	14.29	6.07	20.50	13.16	20.36	13.08	87.28	23.65	-	-
LongLLaMa-3B-Instruct	256k	19.34	7.12	44.25	27.76	41.48	23.06	85.49	29.10	-	-
RWKV-4-14B-raven	8k	8.70	0.82	29.13	10.82	26.80	13.48	83.89	19.97	-	-
LLaMA2-7B-32K-Instruct	32k	0.00	0.00	3.40	0.00	3.40	1.70	66.92	2.71	-	-

Table 3: Performance of the short dependency tasks

Models	Context	Bleu1	Bleu4	Rouge1	Rouge4	RougeL	Meteor_score	Bert_score	GPT4_score
arXiv paper summarization									
GPT4-32k	32k	24.50	0.73	27.15	7.10	24.25	19.03	84.04	82.84
GPT4-8k	8k	**29.02**	**2.09**	**32.08**	**11.11**	28.85	**22.64**	**84.92**	85.42
GPT3.5-turbo-16k	16k	28.70	1.59	32.04	10.69	**28.89**	22.34	84.82	**86.84**
LlamaIndex	\	22.53	0.63	26.28	6.97	23.73	21.07	83.09	76.35
ChatGLM2-6B	32k	0.04	0.00	5.97	0.00	5.82	6.40	73.25	13.23
LongLLaMa-3B-Instruct	256k	0.00	0.00	0.07	0.00	0.07	0.33	26.01	5.57
RWKV-4-14B-raven	8k	15.51	0.09	18.98	3.32	17.12	14.78	80.05	34.79
LLaMA2-7B-32K-Instruct	32k	0.09	0.00	0.20	0.00	0.20	2.44	72.41	6.14
Long dependency QA									
GPT4-32k	32k	8.55	1.40	25.59	6.36	24.04	**11.13**	80.16	**54.09**
GPT4-8k	8k	**8.94**	1.01	23.45	6.57	21.69	10.18	85.36	42.12
GPT3.5-turbo-16k	16k	6.92	**1.81**	25.02	6.68	23.63	10.40	83.79	45.04
LlamaIndex	\	7.76	1.24	23.62	7.10	22.30	10.47	83.87	37.63
ChatGLM2-6B	32k	5.55	0.11	9.41	1.93	8.69	4.39	**85.78**	11.50
LongLLaMa-3B-Instruct	256k	5.64	0.49	17.30	3.76	16.29	6.53	84.26	21.64
RWKV-4-14B-raven	8k	3.88	0.22	20.39	3.20	19.20	6.41	81.46	14.32
LLaMA2-7B-32K-Instruct	32k	0.08	0.00	4.07	0.00	4.07	1.06	66.54	2.85
Claude3-opus	200k	3.28	0.43	**37.95**	**13.46**	**36.56**	9.44	79.58	20.71

Table 4: Performance of the long dependency tasks

Method	CR	Calculation	TR	IR	Average
Human	51	21	26	29	36
GPT4	53	29	25	34	38
Agreement	77	89	76	80	80

Table 5: Accuracy and agreement by GPT4 and human

giving the question(QA task only), ground truth, and predicted outputs, we ask GPT4 to compare and score considering the semantic matching, information completeness, consistency, fluency, and grammar. In this way, GPT4 can focus on the comparisons without bias and tendency. Details can be found in Appendix F. Nevertheless, we understand the limitations of LLM evaluation and recommend the users refer to the combination of different metrics for a comprehensive comparison.

4.2 Results

Fig. 3 shows an overall performance comparison of different models on different tasks. The radar plot shows the original accuracy evaluated by GPT4-8k (except cloze) and the partial match result (for cloze) over different tasks. The abbreviation "sc" in the legend refers to the overall performance of each model. From the charts, GPT4-32k demonstrates its impressive overall performance across all tasks (with highest scores on all tasks except summarization). In comparison, open-source models show a significant performance gap to commercial models on our benchmark. We can find that among the 7

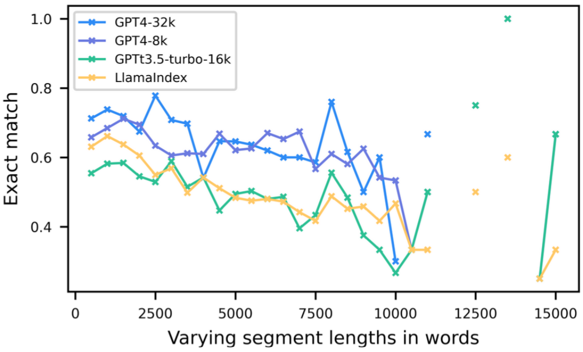


Figure 4: Cloze performance with varying segments

Models	Context	Bleu1	Bleu4	Rouge1	Rouge4	RougeL	Meteor_score	Bert_score	GPT4_score
arXiv paper summarization									
GPT4-32k	32k	24.50	0.73	27.15	7.10	24.25	19.03	84.04	82.84
GPT4-32k	24k	25.57	0.81	27.61	7.53	24.73	19.86	84.07	83.15
GPT4-32k	16k	24.80	0.70	27.29	7.26	24.28	19.12	84.11	82.82
GPT4-32k	8k	26.26	**9.35**	27.83	7.67	24.74	20.08	84.10	82.75
GPT4-8k	8k	**29.02**	2.09	**32.08**	**11.11**	**28.85**	**22.64**	**84.92**	**85.42**
Long dependency QA									
GPT4-32k	32k	7.64	1.24	15.53	4.46	14.60	11.12	86.07	**54.65**
GPT4-32k	24k	8.23	1.66	14.92	4.12	13.90	10.60	86.16	50.61
GPT4-32k	16k	8.57	1.35	16.21	4.30	14.90	**11.91**	**86.36**	47.55
GPT4-32k	8k	7.46	**1.77**	13.75	5.08	12.89	10.01	85.77	38.34
GPT4-8k	8k	**8.94**	1.01	**23.45**	**6.57**	**21.69**	10.18	85.36	42.12

Table 6: Input length impact on long dependency tasks

major tasks, short QA, cloze and summarization are more effectively addressed by LLMs, while **real long dependency QA tasks are far from being solved, where even GPT4-32k hardly achieves over 40% accuracy**. The empirical results demonstrate that even the most successful commercial model still cannot effectively address those really challenging long-dependency tasks, leaving large room for improvement. Detailed evaluations and further analysis can be found in the below sections.

4.2.1 Analysis on short dependency tasks

As in Table 3, GPT4-32k attains the highest accuracy according to the GPT4 evaluator's perspective. GPT4-8k, GPT3.5-turbo-16k, and the retrieval-based LlamaIndex closely follow, demonstrating competitive performance levels. Surprisingly, GPT4-8k exhibits the most robust overall performance in terms of automatic evaluation metrics. It's worth mentioning that GPT4-32k, due to its tendency to generate longer outputs, faces penalties from these automatic metrics. This discrepancy among different metrics highlights the need for improved evaluation methods. Furthermore, in the context of cloze tasks, GPT4-32k excels again when equipped with a longer context window. The open-source models are unable to output JSON format as required, and their performance cannot be automatically evaluated. In Fig. 4 , the exact match results in cloze tasks are displayed for varying source segment lengths. The results show that as the segment length increases, model performance gradually decreases, underscoring the increasing difficulty of effectively filling in the masked entities with longer source text. It is worth noting that the discontinuity and dramatic fluctuations in trend beyond the x-axis value of 12500 in the Fig. 4 are caused by scarcity of data points.

4.2.2 Analysis on long dependency tasks

Table 4 shows the aggregated results on long dependency tasks. We also test the latest commercial model Claude3-opus with 200k context window. Firstly, we can observe that summarization can be well addressed by commercial models, with GPT-4 evaluation accuracy of over 80%. However, the various types of long dependency QAs in our benchmark apparently pose substantial challenges for current LLMs. Both open-source and commercial models experience a significant performance decline. Although the longer context window size indeed helps by comparing GPT4-32k with GPT4-8k, it does not always work on Claude3. The poor performance on long dependency QAs suggests that we may need to revisit LLMs' long context understanding ability in more challenging tasks other than some simple ones like summarization and retrieval, as they are unable to test whether LLMs understand the inter-dependency in long texts.

4.2.3 Long context capabilities deep dive

Impact of varying input length In Table 6, we study the impact of varying lengths of inputs on long dependency QAs with GPT4 models. We find that expanding input length hardly helps in paper summarization. It can be attributed to the inherent nature of the arXiv paper with both the introduction and conclusion sections located at the beginning and at the end respectively, which already contain the major sketch of the paper. In the context of long dependency QA, from the perspective of GPT-4, an extended context window enhances the model's performance by mitigating information loss.

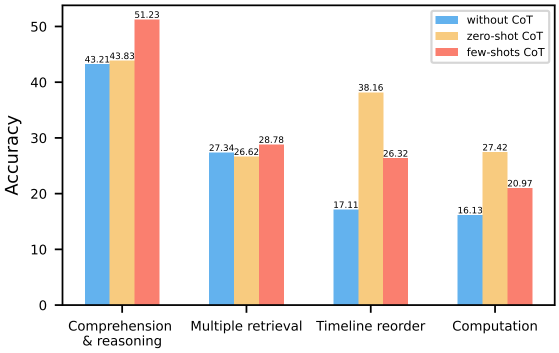


Figure 5: Human evaluation on long QA w/o CoT

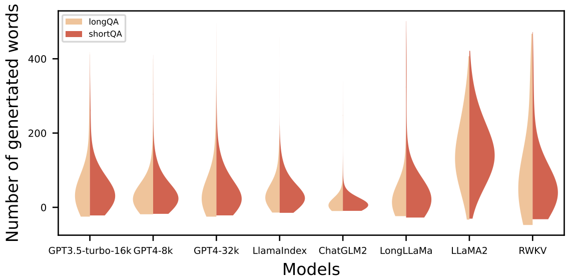


Figure 6: Output distributions on QA tasks

Models	**Information & retrieval**	**Timeline & reorder**	**Calculation**	**Comprehension & reasoning**
GPT4-32k	**33.26**	**26.43**	**22.30**	**44.20**
GPT4-8k	26.59	20.61	16.31	34.42
GPT3.5-turbo-16k	24.05	20.88	13.49	32.10
LlamaIndex	19.38	17.23	11.43	29.53
ChatGLM2-6B-32k	11.38	10.77	8.45	10.95
LongLLaMa-3B-Instruct	15.73	8.87	8.87	21.29
RWKV-4-14B-raven	5.73	4.76	2.08	6.52
LLaMA2-7B-32K-Instruct	2.23	1.36	1.39	2.67

Table 7: Individual task results of long dependency QAs

Individual results on different types of long QAs In this study, we delve into the performance of models in individual tasks that demand diverse capabilities, including reading comprehension, information retrieval, calculation, and reasoning. In this regard, we employed GPT4 as the evaluator, and the accuracy results are available in Table 7. Across the four tasks examined, LLMs generally exhibit strong performance in comprehension, reasoning, and multi-source retrieval, but fall short in tasks related to timeline reordering and calculation. Furthermore, we observed that the way questions are framed has a significant impact on LLMs' performance. Yes-no questions and multiple-choice questions tend to be easier for LLMs to answer, particularly when the search space is limited, as opposed to open-ended questions within unstructured text.

Results on long QAs with/without CoT In this section, we further design experiments to bolster the long-context capabilities and potential of LLMs using the Chain of Thoughts (CoT) (Kojima et al., 2022). A human evaluation was carried out on a subset comprising one-third of instances from each task category within long dependency QA. We selected LlamaIndex as a representative model, given its impressive performance in both short and long dependency question-answering tasks, alongside strong commercial models such as GPT4. We initiated the LLM with a zero-shot CoT approach, employing prompts such as "Let's think step by step," and furnished a few-shot setup with detailed rationales and standard output formats (Wei et al., 2022) to facilitate responses to long dependency questions. In Fig. 5, the zero-shot CoT approach had minimal impact on accuracy in comprehension and reasoning, as well as multi-source retrieval tasks, but yielded a substantial 20% and 10% absolute accuracy increase in timeline reorder and calculation. Interestingly, the few-shot CoT approach benefits the first two types but surprisingly leads to a decline in performance in the latter two types compared with zero-shot. We hypothesize the reason is that the evidence and rationales in few-shot examples cannot be generalized to other questions, and including them might on the contrary give wrong guidance to the model.

Discrepancy in generated outputs of models Distributions of generated outputs of various models are depicted in Fig. 6. It is noteworthy that well-behaved commercial models can produce adaptively varying length of responses without redundancy irrespective of the question type. In contrast, models fine-tuned with longer textual inputs tend to yield significantly lengthier responses, even when a maximum generation constraint is enforced. It may be challenging in comprehending exceedingly complex long QAs at times for some open-source models. Consequently, they prioritize extracting a maximum number of pertinent contexts from their memory to generate sufficiently extensive responses deemed acceptable and rational.

Moreover, there is a significant disparity in generation quality through manual examinations on the outputs across various LLMs. Commercial models consistently generate outputs exhibiting a higher degree of human likeness, completeness, and logical coherence within a structured format who consistently delivers contextually relevant, query-based responses. In contrast, ChatGLM2-6B-32k, tends to offer shorter answers, occasionally confined to numeric responses. It compensates by retrieving relevant contextual information when a definite answer is lacking. RWKV-4-14B-raven often generates duplicated responses or resorts to repeating the given questions to reach the maximum token length, sometimes utilizing code generation to address issues related to its training data. The performance of the LLaMA2-7B-32K-Instruct is notably worse, as it produces irrelevant or nonsensical text with special symbols. More examples of outputs from different models are in Appendix H.

Retrieval Based Techniques To evaluate the effectiveness of retrieval techniques for long-context dependency questions, we undertook extensive experiments by replacing the base LLM model in LlamaIndex with different baseline LLMs. We utilized the open-source embedding all-mpnet-base-v2 (Song et al., 2020). From Table 4 and Table 8, our findings reveal that the incorporation of retrieval techniques does not generally enhance the performance of long dependency QA tasks, particularly evident for models like GPT4-8k and GPT4-32k. Relying solely on retrieval mechanisms might be insufficient in recalling the necessary context and further generating the right answer, resulting in a conspicuous performance decline. However, there is an minor improvement in the BERT score for a few open-source models. It can be possibly attributed to the considerably shorter length of the retrieved segments used as inputs, in contrast to the entirety of the document. The phenomenon highlights the intricacy of our dataset, where a series of long dependency understanding and modeling capabilities such as comprehension and multi-hop reasoning is essentially needed.

5 Discussion

In this paper, we propose a new dataset, **LooGLE** to enable the evaluation of long context understanding of LLMs. **LooGLE** addresses the limitations of previous datasets, including much longer text size, up-to-date documents without pre-training, multi-sources and categories and most remarkably well designed and annotated tasks with diverse context dependency. Our comprehensive evaluations reveal the huge limitations of existing LLMs in understanding and reasoning inter-dependency among long texts even with much longer context windows. Besides, there are significant gap between commercial models with open sourced models whereas they both fail in long dependency tasks in our benchmark. The results show that our dataset functions as a valuable reference for long context assessment and poses potential explorations for LLMs improvement.

Limitation and future work Permissible lengthy input text is inevitable in wide applications and extensive efforts have been devoted to enlarging the supported context length of LLMs. Current long context LLMs have spent huge effort to "memorize" extremely long sequences with efficient attention methods, long-term memory mechanism, length extrapolation mining and etc.(Xiong et al., 2023; Huang et al., 2023; Xu et al., 2023). However they still suffer from tremendous information loss with low precision. In addition, our extensive evaluations unveil substantial limitations in the capacity of existing LLMs to understand and reason about the intricate interdependencies in lengthy texts, even when provided with considerably extended context windows. A few complex capabilities including mathematical computation, temporal awareness of the central storyline, intricate comprehensive understanding, and multi-hop reasoning are essentially needed for current LLMs in long dependency tasks. In the future, we will further explore the development of effective long context LLMs with practically optimized performance.

Despite our thorough experiments to evaluate the long-context LLMs, it may not be precisely reflected due to several factors, including the limitation of the size and diversity of existing models, a lack of adequate inference resources, and the potential for more efficient prompts and techniques to further stimulate the model's abilities. Furthermore, since there is an increasing number of long-context LLMs designed to solve the problem, we will keep updating our long dependency dataset periodically and involving the up-to-date LLMs for more comprehensive evaluations.

Acknowledgement

The authors thank the reviewers for their insightful suggestions on improving the manuscript. This work presented herein is supported in part by the National Key R&D Program of China (2022ZD0160300), the National Natural Science Foundation of China (62276003, 62376031).

References

Chen 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*, abs/2307.11088.

Arian Askari, Suzan Verberne, Amin Abolghasemi, Wessel Kraaij, and Gabriella Pasi. 2024. Retrieval for extremely long queries and documents with rprs: a highly efficient and effective transformer-based reranker. *ACM Transactions on Information Systems*, 42(5):1–32.

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

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Arkadii Bessonov, Alexey Staroverov, Huzhenyu Zhang, Alexey K Kovalev, Dmitry Yudin, and Aleksandr I Panov. 2023. Recurrent memory decision transformer. *arXiv preprint arXiv:2306.09459*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:1877–1901.

Aydar Bulatov, Yuri Kuratov, Yermek Kapushev, and Mikhail S Burtsev. 2023. Scaling transformer to 1m tokens and beyond with rmt. *arXiv preprint arXiv:2304.11062*.

Aydar Bulatov, Yury Kuratov, and Mikhail Burtsev. 2022. Recurrent memory transformer. *Advances in Neural Information Processing Systems*, 35:11079–11091.

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

Xuanting Chen, Junjie Ye, Can Zu, Nuo Xu, Rui Zheng, Minlong Peng, Jie Zhou, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023b. How robust is gpt-3.5 to predecessors? a comprehensive study on language understanding tasks. *arXiv preprint arXiv:2303.00293*.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359.

J Ding, S Ma, L Dong, X Zhang, S Huang, W Wang, N Zheng, and F Wei. 2023. Longnet: Scaling transformers to 1,000,000,000 tokens 2023. *arXiv preprint arXiv:2307.02486*.

Zican Dong, Tianyi Tang, Lunyi Li, and Wayne Xin Zhao. 2023. A survey on long text modeling with transformers. *arXiv preprint arXiv:2302.14502*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021. Efficient attentions for long document summarization. *arXiv preprint arXiv:2104.02112*.

Yunpeng Huang, Jingwei Xu, Zixu Jiang, Junyu Lai, Zenan Li, Yuan Yao, Taolue Chen, Lijuan Yang, Zhou Xin, and Xiaoxing Ma. 2023. Advancing transformer architecture in long-context large language models: A comprehensive survey. *arXiv preprint arXiv:2311.12351*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*.

Dacheng Li, Rulin Shao, Anze Xie, Ying Sheng, Lianmin Zheng, Joseph Gonzalez, Ion Stoica, Xuezhe Ma, and Hao Zhang. 2023a. How long can context length of open-source llms truly promise? In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Shanda Li, Chong You, Guru Guruganesh, Joshua Ainslie, Santiago Ontanon, Manzil Zaheer, Sumit Sanghai, Yiming Yang, Sanjiv Kumar, and Srinadh Bhojanapalli. 2023b. Functional interpolation for relative positions improves long context transformers. *arXiv preprint arXiv:2310.04418*.

Yucheng Li. 2023. Unlocking context constraints of llms: Enhancing context efficiency of llms with self-information-based content filtering. *arXiv preprint arXiv:2304.12102*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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

Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. 2023b. Calibrating llm-based evaluator. *arXiv preprint arXiv:2309.13308*.

Zhengzhong Liu, Aurick Qiao, Willie Neiswanger, Hongyi Wang, Bowen Tan, Tianhua Tao, Junbo Li, Yuqi Wang, Suqi Sun, Omkar Pangarkar, et al. 2023c. Llm360: Towards fully transparent open-source llms. *arXiv preprint arXiv:2312.06550*.

Clara Meister, Stefan Lazov, Isabelle Augenstein, and Ryan Cotterell. 2021. Is sparse attention more interpretable? *arXiv preprint arXiv:2106.01087*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

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*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Huanqi Cao, Xin Cheng, Michael Chung, Matteo Grella, Kranthi Kiran GV, et al. 2023. Rwkv: Reinventing rnns for the transformer era. *arXiv preprint arXiv:2305.13048*.

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

Uri Shaham, Elad Segal, Maor Ivgi, Avia Efrat, Ori Yoran, Adi Haviv, Ankit Gupta, Wenhan Xiong, Mor Geva, Jonathan Berant, et al. 2022. Scrolls: Standardized comparison over long language sequences. *arXiv preprint arXiv:2201.03533*.

Eva Sharma, Chen Li, and Lu Wang. 2019. Bigpatent: A large-scale dataset for abstractive and coherent summarization. *arXiv preprint arXiv:1906.03741*.

Roshan Sharma, Suyoun Kim, Daniel Lazar, Trang Le, Akshat Shrivastava, Kwanghoon Ahn, Piyush Kansal, Leda Sari, Ozlem Kalinli, and Michael Seltzer. 2023. Augmenting text for spoken language understanding with large language models. *arXiv preprint arXiv:2309.09390*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33:16857–16867.

Gaurav Suri, Lily R Slater, Ali Ziaee, and Morgan Nguyen. 2024. Do large language models show decision heuristics similar to humans? a case study using gpt-3.5. *Journal of Experimental Psychology: General*.

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. 2020. Long range arena: A benchmark for efficient transformers. *arXiv preprint arXiv:2011.04006*.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2022. Efficient transformers: A survey. *ACM Computing Surveys*, 55(6):1–28.

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

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. musique: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Szymon Tworkowski, Konrad Staniszewski, Mikołaj Pacek, Yuhuai Wu, Henryk Michalewski, and Piotr Miłoś. 2024. Focused transformer: Contrastive training for context scaling. *Advances in Neural Information Processing Systems*, 36.

Alex Wang, Richard Yuanzhe Pang, Angelica Chen, Jason Phang, and Samuel R Bowman. 2022. Squality: Building a long-document summarization dataset the hard way. *arXiv preprint arXiv:2205.11465*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Rodrigo Wilkens. 2023. Statistical methods for annotation analysis. *Computational Linguistics*, pages 763–765.

Jeff Wu, Long Ouyang, Daniel M Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. 2021. Recursively summarizing books with human feedback. *arXiv preprint arXiv:2109.10862*.

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*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Yan Zeng, Hanbo Zhang, Jiani Zheng, Jiangnan Xia, Guoqiang Wei, Yang Wei, Yuchen Zhang, and Tao Kong. 2023. What matters in training a gpt4-style language model with multimodal inputs? *arXiv preprint arXiv:2307.02469*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. 2023. Don't make your llm an evaluation benchmark cheater. *arXiv preprint arXiv:2311.01964*.

A More details of our dataset and experiment settings

Our benchmark comprises texts all published after 2022 which ensures that most modern LLMs have not been pre-trained on these documents, forcing them to rely on their in-context learning ability rather than memorization. Furthermore, our data collection process is fully open-sourced, making it easy for the community to reconstruct/update the benchmark with newer documents, possibly every year. Distributions of the input length and dependency spanning in words for long dependency QA tasks are shown in Figs. 7 and 8.

We provide probable explanations for long QA bad cases to provide insights and directions for model promotion in Table 10 in Appendix E using the same subset as 1. We randomly sampled over 300 QAs and captured discrepancies in the generated outputs of different models to tackle inherent preferences encountered in the long context in Section 4.2.3. In evaluation, the implementation parameters of models and metrics are in default settings. All the experiments can be run on 2 A100 each with 80G GPU. Each result is the average of several runs.

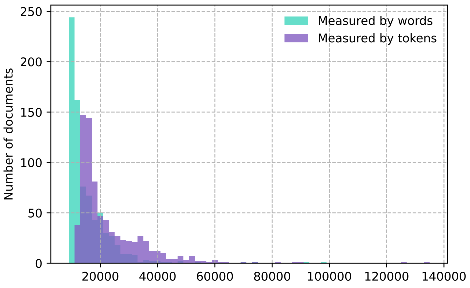


Figure 7: Input document length distributions

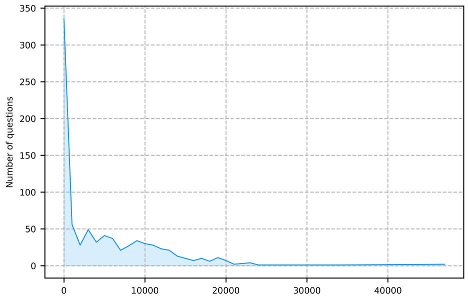


Figure 8: Evidence dependency distributions

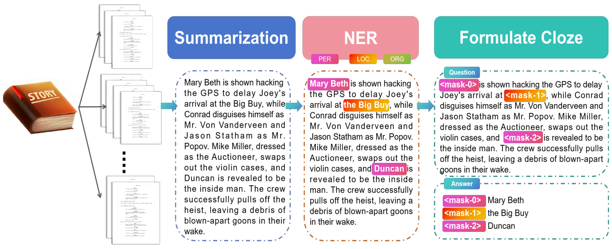
B Short dependency task definition and generation

Question Answering (QA) To generate short dependency QA pairs, we harnessed the robust language processing and comprehension capabilities of GPT3.5-turbo-16k. These short dependency QA pairs typically do not require extensive evidence retrieval and can be extracted from localized segments. We divided each article into multiple segments and employed an iterative approach to prompt the Language Model (LLM) to generate QA pairs based on these segments, including their associated supporting evidence from the article. Details of the prompts are available in Appendix F. Subsequently, we conducted manual reviews of the QA pairs, making refinements to some of the answers by filtering out non-essential context and eliminating redundant descriptions. This rigorous curation process was undertaken to ensure the high quality and relevance of the resulting QA pairs.

Cloze Initially, each script is divided into segments of varying lengths. Then, we employ GPT3.5-turbo-16k to generate factual summaries aligning with the source segment along with some constraints included in prompts (see Appendix F). Later, we employ BERT-large (Kenton and Toutanova, 2019) for Named Entity Recognition (NER) from the generated summaries, limiting the types to person name, location, and organization. Finally, we randomly select a certain number (no more than 5) of entities from the summary and mask them as placeholders, denoted as "<mask-n>". The goal is to predict the masked entities according to the long context.The segmented scripts are regarded as the evidence for the corresponding cloze questions. The Cloze task formulation process can be seen in Fig. 9.

C Task definition

The Cloze task formulation process is in Fig. 9.



Mary Beth is shown hacking the GPS to delay Joey's arrival at the Big Buy, while Conrad disguises himself as Mr. Von Vanderveen and Jason Statham as Mr. Popov. Mike Miller, dressed as the Auctioneer, swaps out the violin cases, and Duncan is revealed to be the inside man. The crew successfully pulls off the heist, leaving a debris of blown-apart goons in their wake.

Mary Beth is shown hacking the GPS to delay Joey's arrival at the Big Buy, while Conrad disguises himself as Mr. Von Vanderveen and Jason Statham as Mr. Popov. Mike Miller, dressed as the Auctioneer, swaps out the violin cases, and Duncan is revealed to be the inside man. The crew successfully pulls off the heist, leaving a debris of blown-apart goons in their wake.

<mask-0> is shown hacking the GPS to delay Joey's arrival at <mask-1>, while Conrad disguises himself as Mr. Von Vanderveen and Jason Statham as Mr. Popov. Mike Miller, dressed as the Auctioneer, swaps out the violin cases, and <mask-2> is revealed to be the inside man. The crew successfully pulls off the heist, leaving a debris of blown-apart goons in their wake.

<mask-0> Mary Beth
<mask-1> the Big Buy
<mask-2> Duncan

Figure 9: Cloze task

D Timeline reorder evaluation metrics

We employ 4 metrics to measure the similarity of numeric output sequences for timeline reorder tasks. For given two numeric sequences A and B

Models	**Context**	**Bleu1**	**Bleu4**	**Rouge1**	**Rouge4**	**RougeL**	**Meteor_score**	**Bert_score**	**GPT4_score**
GPT4-32k	32k	6.08	1.31	10.27	3.39	9.52	8.54	85.27	28.25
GPT4-8k	8k	6.62	**1.50**	**11.95**	**3.80**	**10.99**	9.02	85.51	26.34
GPT3.5-turbo-16k	16k	6.50	0.92	10.93	3.56	9.86	8.65	**85.63**	**33.24**
Default	\	**7.02**	1.24	11.60	3.75	10.57	**9.37**	85.61	33.16
ChatGLM2-6B-32k	32k	0.15	0.00	2.23	0.05	2.23	0.74	83.40	7.73
LongLLaMa-3B-Instruct	256k	1.11	0.00	2.34	0.00	2.26	2.23	82.19	5.28
RWKV-4-14B-raven	8k	0.56	0.00	1.09	0.01	1.02	0.82	76.68	0.09
LLaMA2-7B-32K-Instruct	32k	0.31	0.00	0.81	0.00	0.76	1.94	78.07	6.08

Table 8: Long dependency QA with LlamaIndex

Models	**LSD**	**LMD**	**SD**	**SDD**	**LSD-S**	**LMD-S**	**SD-S**	**SDD-S**	**Non-standard(%)**
GPT4-32k	**1.04**	**0.57**	**0.93**	**1.12**	1.21	**0.82**	1.41	1.60	52.80
GPT4-8k	1.24	0.64	1.04	1.281	1.43	0.92	1.51	1.74	49.31
LlamaIndex	1.55	0.78	1.19	1.551	1.95	1.08	1.65	2.09	**39.72**
GPT3.5-turbo-16k	3.58	1.43	2.17	2.916	**1.05**	0.86	1.24	**1.26**	77.21
ChatGLM2-6B-32k	4.31	1.74	2.63	3.37	1.83	1.17	1.50	2.00	99.07
LongLLaMa-3B-Instruct	4.36	1.76	2.66	3.40	0.00	0.00	0.00	0.00	100.00
RWKV-4-14B-raven	3.45	1.45	2.18	2.83	2.33	1.21	1.77	2.38	77.57
LLaMA2-7B-32K-Instruct	4.33	1.75	2.65	3.39	2.67	1.33	2.00	2.67	98.60

Table 9: Performance of Timeline Reorder

with the same sequence length n, $i[A]$ and $I[B]$ is the ith number in each sequence. They can be computed using the formula below:

LSD is the abbreviation for location square deviation:

$$LSD(A,B) = \frac{1}{n}\sum_{i=0}^{n-1}(i[A] - i[B])^2 \qquad (1)$$

LMD is the abbreviation for location mean deviation:

$$LMD(A,B) = \frac{1}{n}\sum_{i=0}^{n-1}|i[A] - i[B]| \qquad (2)$$

SD is the abbreviation for swap deviation:

$$SD(A,B) = min(\textstyle\sum_{s\in A\to B} 1) \qquad (3)$$

SDD is the swap distance deviation:

$$SDD(A,B) = min(\textstyle\sum_{s\in A\to B} |i-j|) \qquad (4)$$

where $s = A(i,j)$ means the swap between the ith and jth element in A. $S = A \to B$ means a series of swap actions to convert A to B. $W(S) = \sum_{s\in S} w(s)$ means the weights sum of all the swap actions in S, where $w(s) = 1$ in SD and $w(s) = |i-j|$ in SSD.

E Further analysis and results

Automatic evaluation on timeline reorder It is essential to address discrepancies arising from the diverse formats produced by various models. Typically in time reorder tasks, these outputs comprise conventional numerical sequences, but errors in non-standard formats when evaluation necessitate preprocessing for accurate assessment. We propose to convert the serial numbers in the candidate answers from their original question into Roman numbers (*i.e.*, I, II, $\cdots$), thereby enhancing discrimination through regular expression matching. Four key metrics, namely, LSD (location square deviation), LMD (location mean deviation), SD (swap deviation), and SDD (swap distance deviation), are employed to measure the similarity of numeric sequences, refer to Appendix D. Smaller deviations indicate a higher degree of resemblance between the sequences. Any outputs that are empty, possess unequal lengths, or contain extra elements are categorized as non-standard. The maximum deviation between the provided ground truth and all corresponding candidate answers is computed as the worst score for evaluation purposes.

The percentage of non-standard outputs for each model are in Table 9. It is evident that except for GPT4, which demonstrates a remarkable degree of adherence and alignment following Reinforcement Learning from Human Feedback (RLHF)

Task	Models	Context	Bleu1	Bleu4	Rouge1	Rouge4	RougeL	Meteor_score	Bert_score	GPT4_score
Short dependency QA	LLaMA2-7B-32K-Instruct	32k	0.00	0.00	3.40	0.00	3.40	1.70	66.92	2.71
	LLaMA2-7B	4k	1.35	0.00	6.72	0.92	6.45	2.97	80.64	7.06
Long dependency QA	LLaMA2-7B-32K-Instruct	32k	0.08	0.00	4.07	0.00	4.07	1.06	66.54	2.85
	LLaMA2-7B	4k	1.30	0.00	7.78	0.35	7.40	2.52	80.48	7.95

Table 10: Evaluation of the base model LLaMA2-7B and its long context version

Long QA Tasks	Hallucination[*]	Redundant retrieval[†]	Insufficient retrieval[⋆]	Irrelevant answer[⋄]	No relevant context[∧]	Wrong/No reasoning[×]	Others
Calculation	31.11	24.44	15.56	0.00	20.00	0.00	8.88
Multi-source retrieval	14.71	31.37	28.43	13.73	13.73	0.00	7.84
Comprehension and reasoning	14.29	10.99	21.98	18.68	16.48	10.99	6.59

[*] Evidence of predictions is not shown up in the original inputs and generated by LLM itself from nowhere.

[†] Apart from the right evidence, irrelevant evidence is also redundantly retrieved.

[⋆] Not all of the essential evidence to answer the question is retrieved.

[⋄] Evidence of predictions generated have no or minor correlation with the question.

[∧] No relevant context in LLM's memory and refuse to answer the question.

[×] Fail to retrieve evidence that needs further reasoning other than directly extracting from the inputs.

Table 11: Bad cases study on the long dependency QA

(Lee et al., 2023), most open-sourced models struggle to generate texts in the correct format with less than 10%. However, it can be mitigated in significantly large models through the utilization of few-shot prompts and mandatory instructions. This phenomenon results in performance penalties when assessed using automated metrics. Consequently, to ascertain the genuine capacity of LLMs in this task, we calculate the four metrics exclusively for outputs in standard format ("-S").

Effect of the instruction following ability on long context From the Table 10, we can see that the base model demonstrates superior performance than the model finetuned for a long context. It is mainly because a longer context window promotes the model's performance by introducing much information loss with low precision. Current long context LLMs have spent huge effort to "memorize" extremely long sequences with efficient attention methods, long-term memory mechanisms, length extrapolation mining, etc. Our results shed light on enhancing the "true long-context understanding" ability of LLMs instead of merely enlarging their context window.

Probable explanations for long QA bad cases To investigate whether the models have effectively memorized and comprehended lengthy context, we manually conducted a comprehensive analysis of the underlying causes of failures in long QAs. The rationale behind CoT aided in understanding how models decompose and tackle challenges associated with extended dependency-based QA. Our observations reveal that LLMs struggle with the tasks primarily due to their inability to extract precise information and a propensity to generate responses lack of factuality. Constraints imposed by the inherent context window limitations, coupled with information loss resulting from the extending techniques, contribute to struggles in memorizing extensive contexts. In most cases, models attempt to compensate by retrieving and integrating the most pertinent evidence, even if it results in redundant answers. However, they also acknowledge their insufficient context and at times, abstain from providing responses rather than resorting to nonsensical answers. Furthermore, addressing these challenges necessitates enhanced comprehension and reasoning abilities, particularly when answers are not evident across multiple pieces of evidence throughout the texts. The insights derived from our analysis offer a scientific foundation and pave the way for promising research directions on augmenting LLM for handling long contexts.

F Prompts

F.1 Short dependency QA pair generation

[seg] = *{Input long texts}*
Please generate 2 questions and corresponding answers based on the given [seg] in as few words as possible. Return reference text S, question Q and answer A from [seg] in json format as: {"S": ,"Q": ,"A": },{"S": ,"Q": ,"A": }.

F.2 Short and long dependency QA

Instruction: Please answer the question based on the given long texts below.
{Input long texts}
Question: *{Question}*
Answer:

F.3 scripts segment summarization for cloze formulation

Instruction: Please write a summary for this script segment within 500 words, focusing on describing objective facts and avoiding subjective opinions.
{scripts segement}
Summary:

F.4 Cloze

example= *{"<mask-0>": "Bob", "<mask-1">: "Gorrosion Magazine", "<mask-2>": "Bethel Horizon"}*
Instruction: Please answer the cloze question based on the given long texts below. Each of the placeholder "<mask-n>" in the question could be an entity of Person, Location or Organization. The same masks represent the same entity.
Output a json format answer, for example:*{example}*
{Input long texts}
Question: *{Question} What are the masked entities?*
Answer:

F.5 Summarization

Instruction: Please generate a summary of the below paper.
{Input long texts}
Summarization:

F.6 Timeline reorder

Instruction: Please answer the question based on the given long texts below. Please use Roman numbers only for answers.
{Input long texts}
Question: *{Question}*
Answer:

F.7 QA task evaluation by GPT4

Instruction: Given one question, there is a ground truth and a predict answer. Please decide whether they are the same or not in semantics. Please only output True or False.
Question: *{Question}*
ground truth = *{Reference answer}*
predicted answer = *{Generated output}*

F.8 Summarization task evaluation by GPT4

Instruction: Please compare the predicted summary with the ground truth and evaluate the predicted summary from the perspectives of information completeness, consistency, fluency, and grammar by giving a score within the range of 0 to 100.
ground truth = *{Reference answer}*
predicted summary = *{Generated output}*

F.9 Few-Shot CoT for long QA

Instruction: Please answer the question based on the given long texts below.
{Input long texts}
{Demonstrations}
Question: *{Question}*
Answer:

F.10 Zero-Shot CoT for long QA

Instruction: Please answer the question based on the given long texts below.
{Input long texts}
Question: *{Question}*
Answer: Let's think step by step.

G Examples for long context understanding tasks

G.1 Short dependency QA

Question:
Who did Picardo collaborate with for building preservation and restoration projects?

Evidence:
On qualifying in 1951, Picardo pursued his interest in historical architecture by collaborating on a number of building preservation and restoration projects with the Spanish architect and architectural historian Fernando Chueca Goitia, who was 8 years his senior.

Answer:
He collaborated with Spanish architect and architectural historian Fernando Chueca Goitia.

Question:
What was the nickname given to the 18th century period?
Evidence:
The 18th century was nicknamed the 'Age of Enlightenment', as it was the period in which the Enlightenment emerged, a philosophical movement that defended reason and science against religious dogmatism.
Answer:
The Age of Enlightenment.

G.2 Cloze

Question:
<mask-0> is shown hacking the GPS to delay Joey's arrival at <mask-1>, while Conrad disguises himself as Mr. Von Vanderveen and Jason Statham as Mr. Popov. Mike Miller, dressed as the Auctioneer, swaps out the violin cases, and <mask-2>is revealed to be the inside man. The crew successfully pulls off the heist, leaving a debris of blown-apart goons in their wake.
Long input:
When a caper crew needs something blown up for a heist, they call upon The Demolition Expert. They are often minor characters who are not given much screen ....(104,094 words).... Joey is driving to the Big Buy, always craning back... like there's a phantom on his tail. Suddenly, the GPS chimes. GPS VOICE Rerouting . DRIVER Shit. Uh, boss, it says it just added twenty minutes. The speed past – A BLACK MUSTANG parked in a turnaround. Mary Beth in the driver's seat, clacking away on a laptop, hacking the GPS . ....(150 words)....we couldn't have done it without Duncan– Reveal Duncan , smiling big. He raises his glass. FLASH: DUNCAN and TWO MORE GOONS hurry around the corner of the STADIUM HALLWAY and stop dead in their tracks when they see – A debris of blown-apart Goons littering the hallway. ....(2,670 words).
Answer:
{"<mask-0>": "Mary Beth", "<mask-1">: "the Big Buy", "<mask-2>": "Duncan"}

G.3 Summarization

Long input:
Distinction and quadratic base change for regular supercuspidal representations Chuijia Wang 1 Introduction Let be a connected reductive algebraic group over a non-archimedean local field with residual characteristic ....(21,000 words).... Basically, one can describe all the characters of which occur in in terms of certain intersection property between the Kostant sections of and the orbit of the generic element associated to. ....(500 words).
Summary:
In this article, we study Prasad's conjecture for regular supercuspidalrepresentations based on the machinery developed by Hakim and Murnaghan tostudy distinguished representations, and the fundamental work of Kaletha onparameterization of regular supercuspidal representations. For regularsupercuspidal representations, we give some new interpretations of thenumerical quantities appearing in Prasad's formula, and reduce the proof to thecase of tori. The proof of Prasad's conjecture then reduces to a comparison ofvarious quadratic characters appearing naturally in the above process. We alsohave some new observations on these characters and study the relation betweenthem in detail. For some particular examples, we show the coincidence of thesecharacters, which gives a new purely local proof of Prasad's conjecture forregular supercuspidal representations of these groups. We also prove Prasad'sconjecture for regular supercuspidal representations of G(E), when E/F isunramified and G is a general quasi-split reductive group.

G.4 Multi-source retrieval

Question:
What were some of the architectural projects José Luis Picardo worked on?
Long input:
José Luis Picardo ....(1,520 words) .... From the early 1960s to 1985 Picardo dedicated much of his professional life to the state-run hotel chain, Paradores de Turismo de España .....(7,846 words) .... In 1970 Picardo was invited to compete with fellow notable architects Javier Carvajal Ferrer [es] and Mariano García Benito [es] for the contract to design and build a new headquarters building in the Salamanca neighbourhood of Madrid for the Fundación Juan March (Juan March Foundation) which promotes Spanish culture and science ....(651 words) .... Picardo's commission from the Ministry was to design a sala de equitación, a huge arena for horse and riding displays, in particular the school's signature performance "Como Bailan los Caballos Andaluces" ("How the Andalusian Horses Dance") which would seat up to 1,600 spectators.

Connected to it were to be stable facilities for 60 horses ....(1,113 words).

Answer:

He worked on hotel chain Paradores de Turismo de España, Fundación Juan March, Sala de Equitación.

Explanation:

Based on the deep understanding of given question, we need to extract all the evidence of architectural projects José Luis Picardo have worked on. There are total three works spreading in the original inputs independently as shown above.

G.5 Timeline reorder

Question:

Reorder the timeline of below events:
1.restoration at Guadalupe,
2.restore and rehabilitate the old Casa de la Inquisición
3.castle conversion at Sigüenza,
4.renovation and conversion of castle at Puebla de Alcocer

Long input:

José Luis Picardo ....(2,395 words) .... Restoration at Guadalupe started in November 1963 and the hotel, with twenty double rooms, opened on 11 December 1965 ....(1,472 words) .... In 1965 Picardo was commissioned by Paradores to restore and rehabilitate the old Casa de la Inquisición (House of the Inquisition) in the small, historic village of Pedraza, 37 kilometres northeast of Segovia in Castilla y León ....(2,827 words) .... In 1964 Picardo was involved, with the Ministry of Information and Tourism, in investigating old buildings for conversion into a new Parador in the Province of Guadalajara. Possible locations were the castle at Atienza and the Casa del Cordón, an old inn in the same town, the castle at Molina de Aragón and the castle at Sigüenza ....(1,521 words) .... Among the most advanced plans Picardo drew up were in 1969 for the renovation and conversion into a Parador of the castle at Puebla de Alcocer, a small municipality 70 miles east of Mérida in the Province of Badajoz in Extremadura ....(2,897 words).

Answer:

1, 3, 2, 4

Explanation:

The four events provided in the question sequentially happen with thousands of words spanning. We firstly locate the exact sentences describing the event in the original inputs above. Then we reorder them based on the their occurrence.

G.6 Calculation

Question:

How many inhabitants increases from the end of 19th to 1970?

Long input:

Urban planning of Barcelona ....(5,558 words) .... After the revolution of 1868, the Citadel was also demolished and the land transformed into a public park. The population grew, especially thanks to immigration from the rest of Spain, reaching 400,000 inhabitants by the end of the century. ....(7,613 words) .... In two decades it went from 1,280,179 inhabitants in 1950 to 1,745,142 in 1970 ....(5,596 words).

Answer:

1,345,142

Explanation:

Firstly, we locate the numeric of inhabitants which only appear between 1900 to 1970 from the input as evidence. There are three relevant numbers: 400,000, 280,179 and 1,745,142. Then we make computation on 1,745,142 - 400,000 = 1,345,142 to get the final answer.

G.7 Comprehension and reasoning

Question:

Which event is the turning point for territorial expansion in the 19th?

Long input:

Urban planning of Barcelona ....(2,958 words) .... At this time Barcelona was constituted as a county and later became part of the Crown of Aragon and the political and economic center of the Principality of Catalonia, becoming an important maritime and commercial axis of the Mediterranean Sea....(128 words) ....The progressive increase in the size of the city, and its increasing urban, social and economic complexity, led to the creation of a specific system of government for the administration of the city, the Council of One Hundred (1,265)....(1,260 words) ....The city was still confined within its walls —the only expansion was on the beach, in the neighborhood of La Barceloneta— despite the fact that by the end of the period it had almost 100,000 inhabitants....(1,333 words) ....Barcelona thus underwent an important leap to modernity, characterized by three factors: the population migration from the countryside to the city, the link between industrial and urban developments, and a better articulation of the territory through a wide network of roads and railroads, which will lead Barcelona to become a colonizing metropolis of its territorial environment.....(1,319 words) ....In the middle of the century a transcendental event took place that completely changed the physiognomy of the city; the demolition of the walls.....(1,472 words) ....Another service that emerged at the end of the century was the telephone. The first telephone communication in the whole peninsula took place in Barcelona, in 1877, between the Montjuic castle and the fortress of the Citadel —in the process of dismantling but still housing a garrison—. That same year the first interurban transmission between Barcelona and Girona was carried out by the company Dalmau i Fills, pioneer in the installation of lines in Barcelona......(9,966 words).

Answer:

The demolition of the walls and the donation to the city of the fortress of the Citade.

Explanation:

There is no exact answer that can be extracted directly from the original input and further reasoning is essentially needed. Firstly, we find all the relevant sentences or paragraphs on the territorial expansion which are in the 19th. Temporal events and key words act as clues for sorting out the evidence. Among the evidence, some are factual plots while others are critical comments. Based on further comprehension, we need to compare the various factors of territorial expansion underlying the evidence and decide the most decisive and influential turning point events to answer the question.

H Examples of models' generated outputs

H.1 GPT4-32k

1. short QA

Question: What were the charges against the two members of the Rock Machine Perth chapter?

Output: The two members of the Rock Machine Perth chapter, Kyle Barry and Brent Reker, were charged with extortion and demanding property with oral threats.

Ground truth: Extortion and demanding property with oral threats.

Question: Why did the federal government create relocation programs for Native Americans?

Output: The federal government created relocation programs for Native Americans in an effort to expand their employment opportunities. Many Native Americans were living in rural and isolated areas, and the government believed that moving them to urban areas would offer better job prospects. This was part of a wider policy shift towards giving tribes more authority for self-determination.

Ground truth: To expand their employment opportunities.

2. long QA

Question: According to World Bank's estimation, what is the difference between the amount of World Bank's aid to Turkey and the amount of damaged suffered by Syria?

Output: The World Bank estimated \$5.1 billion in damages suffered by Syria and it provided \$1.78 billion in aid to Turkey. So, the difference is \$3.32 billion.

Ground truth: 3.32 Billion.

Question: What efforts did guy make to improve his life?

Output: Guy made several efforts to improve his life, including attending therapy sessions, building a tree house for his nephews, discussing his feelings and insecurities with his friends, and standing up for himself against his aggressive father. He also expressed his feelings to his ex-girlfriend Jen, trying to reconcile with her and offering her a house he planned to renovate. He made attempts to change his negative self-talk and confront his past traumas.

Ground truth: 1. Receive psychotherapy: Guy chooses to see a psychiatrist to face his own problems, which is an important step for him to change. 2. Introspection: Guy gradually enlightened himself during the treatment, facing the relationship problems between him, his father and his friends. 3. Write a letter to Jen: Guy wrote a sincere letter to Jen, expressing his change of attitude, trying to win back Jen's heart. 4. Reaching out to old friends: Despite the embarrassment, Guy reached out to old friends and tried to work things out. 5. Renovate the house: Guy takes on the job of renovating the house to plan for his future life. 6. Dealing with the relationship with the father: Guy faces his father, expresses his thoughts, and no longer avoids this issue. 7. Change yourself: Guy works hard, writes letters, receives treatment, etc., showing his determination to change himself.

3. summarization

Output: This paper discusses the distinction

and quadratic base change for regular supercuspidal representations. The author focuses on understanding Prasad's conjecture on the Galois distinction problem, which proposes a relationship between the distinction property of an irreducible representation and the base change property of its Langlands parameters. The author's work highlights that even though significant progress has been made in proving the conjecture, a general method that works for all Galois symmetric pairs would be interesting. It extends the understanding of the relationship between distinguished representations and Langlands functoriality. The results support the idea of using a purely local approach to prove the conjecture for regular supercuspidal representations without constraints on the Galois symmetric pair. The author also outlines strategies to improve the exact analysis of terms appearing in Prasad's identity. The article provides evidence supporting the conjecture by evaluating particular examples.
Ground truth: In this article, we study Prasad's conjecture for regular supercuspidal representations based on the machinery developed by Hakim and Murnaghan to study distinguished representations, and the fundamental work of Kaletha on parameterization of regular supercuspidal representations. For regular supercuspidal representations, we give some new interpretations of the numerical quantities appearing in Prasad's formula, and reduce the proof to the case of tori. The proof of Prasad's conjecture then reduces to a comparison of various quadratic characters appearing naturally in the above process. We also have some new observations on these characters and study the relation between them in detail. For some particular examples, we show the coincidence of these characters, which gives a new purely local proof of Prasad's conjecture for regular supercuspidal representations of these groups. We also prove Prasad's conjecture for regular supercuspidal representations of G(E), when E/F is unramified and G is a general quasi-split reductive group.

4. cloze
Question: The script segment of " Wildfire 2022" takes place in Tulare County, California, where the sky is grey and ashy due to the nearby fires. The story follows <mask-1>, a seven-year-old mute girl, who runs to school along a busy road. <mask-0>, a Trans Latin Woman, is seen singing and dancing to " Opposites Attract" while trying to fix her old car. Lu goes to school and stands in line for lunch, where she is ignored by other kids. She pays for her lunch in change and sits alone at the back of the cafeteria. Later, Fireman <mask-2>speaks to Mrs. Consuelos' third-grade classroom about fire safety, and Lu is chosen to answer a question despite not speaking. The leading cause of wildfires is discussed, and it is revealed that humans are the leading cause of wildfires due to accidents like leaving an open campfire burning or tossing a cigarette out a car window.
Output: {"<mask-0>": "Merribelle", "<mask-1>": "Lu", "<mask-2>": "Josh"}
Ground truth: {"<mask-0>": "Merribelle", "<mask-1>": "Lu", "<mask-2>": "Josh"}

H.2 GPT4-8k

1. short QA
Question: When did large cattle ranches start to appear in the area?
Output: As early as 1734.
Ground truth: As early as 1734.

Question: What contributed to the economic growth of Rio Grande do Sul at the end of the 19th century?
Output: The economic growth of Rio Grande do Sul at the end of the 19th century was contributed by the growth of trade, the arrival of immigrants of other origins such as Italians and Jews, and the establishment of a growing industry and a rich bourgeois class. The state had become the third largest economy in Brazil by the turn of the 20th century.
Ground truth: Trade and immigration of Italians and Jews.

2. long QA
Question: Reorder the timeline of below events:
1. Guardian Reported Sexually Assault.
2. Nunez Announced taking place An Internal Investigation.
3. A Young Woman Was Reported To Have Lost Her Thumb.
Output:
3. A Young Woman Was Reported To Have Lost Her Thumb.
1. Guardian Reported Sexually Assault.
2. Nunez Announced taking place An Internal Investigation.

Ground truth: 2,1,3

Question: Was Maurice Duplessis' treatment of the news publishing industry consistent or inconsistent over his two terms?
Output: The text does not provide specific information on Maurice Duplessis' treatment of the news publishing industry over his two terms.
Ground truth: Inconsistent.

3. summarization
Output: This paper by Chuijia Wang focuses on the study of representation theory of a connected reductive algebraic group over a non-archimedean local field. The main focus is to understand Prasad2̆019s conjecture on the Galois distinction problem. The author provides evidence on the possibility of proving this conjecture for regular supercuspidal representations without constraints on the Galois symmetric pair using a different purely local method. The paper also discusses the work of Sakellaridis and Venkatesh, Dipendra Prasad, and others in the field. The author proves Prasad2̆019s conjecture for regular supercuspidal representations of certain groups and provides a detailed study of these quadratic characters. The paper concludes with a discussion on the distinction problem and the construction of local Langlands correspondence.
Ground truth: In this article, we study Prasad's conjecture for regular supercuspidal representations based on the machinery developed by Hakim and Murnaghan to study distinguished representations, and the fundamental work of Kaletha on parameterization of regular supercuspidal representations. For regular supercuspidal representations, we give some new interpretations of the numerical quantities appearing in Prasad's formula, and reduce the proof to the case of tori. The proof of Prasad's conjecture then reduces to a comparison of various quadratic characters appearing naturally in the above process. We also have some new observations on these characters and study the relation between them in detail. For some particular examples, we show the coincidence of these characters, which gives a new purely local proof of Prasad's conjecture for regular supercuspidal representations of these groups. We also prove Prasad's conjecture for regular supercuspidal representations of G(E), when E/F is unramified and G is a general quasi-split reductive group.

4. cloze
Question: The script segment of " Wildfire 2022" takes place in Tulare County, California, where the sky is grey and ashy due to the nearby fires. The story follows <mask-1>, a seven-year-old mute girl, who runs to school along a busy road. <mask-0>, a Trans Latin Woman, is seen singing and dancing to " Opposites Attract" while trying to fix her old car. Lu goes to school and stands in line for lunch, where she is ignored by other kids. She pays for her lunch in change and sits alone at the back of the cafeteria. Later, Fireman <mask-2>speaks to Mrs. Consuelos' third-grade classroom about fire safety, and Lu is chosen to answer a question despite not speaking. The leading cause of wildfires is discussed, and it is revealed that humans are the leading cause of wildfires due to accidents like leaving an open campfire burning or tossing a cigarette out a car window.
Output: {"<mask-0>": "Merribelle", "<mask-1>": "Lu", "<mask-2>": "Josh"}
Ground truth: {"<mask-0>": "Merribelle", "<mask-1>": "Lu", "<mask-2>": "Josh"}

H.3 GPT3.5-turbo-16k

1. short QA
Question: What was unveiled by Buckingham Palace for the Coronation of Charles III and Camilla?
Output: The official Coronation logo was unveiled by Buckingham Palace for the Coronation of Charles III and Camilla.
Ground truth: The official Coronation logo.

Question: What was the percentage decline in house prices in the UK?
Output: There is no specific information provided in the text about the percentage decline in house prices in the UK.
Ground truth: 3.4%.

2. long QA
Question: Which church hasn't been built during the Phase 1 of church construction?
1. The Main Church of Sabará
2. The Cathedral Basilica of Salvador
3. The Church of Our Lady of Grace
Output: The Main Church of Sabará.
Ground truth: 1,2.

Question: Do Reader's Digest and Alexander Cockburn agree on Mujahideen?
Output: No, Reader's Digest and Alexander Cockburn do not agree on the Mujahideen. Reader's Digest generally portrayed the Mujahideen as freedom fighters and heroes, while Cockburn criticized the US support for the Mujahideen and argued that they were not the noble fighters they were made out to be.
Ground truth: No.

3. summarization
Output: This paper focuses on understanding Prasad's conjecture on the Galois distinction problem for regular supercuspidal representations. The distinction problem refers to the question of determining which irreducible representations of a reductive group are distinguished with respect to a pair of subgroups and a character. The paper provides evidence for the possibility of proving this conjecture for regular supercuspidal representations using a different purely local method.

The paper begins by introducing the Langlands-Vogan bijection, which relates irreducible representations of a reductive group to Langlands parameters. It then discusses the local Langlands correspondence for regular supercuspidal representations and the parametrization of these representations using regular supercuspidal packet data. The paper also reviews the character expansions and the Murnaghan-Kirillov formula for regular semisimple and nilpotent orbits.

The main part of the paper focuses on the computation of various characters associated with regular supercuspidal representations. It considers the characters associated with Prasad's conjecture, Hakim's character, and the character associated with - data. The paper provides explicit computations for these characters in various cases, such as when the quadratic extension is unramified or ramified, and when the group is odd or even. The paper also compares these characters and shows that they satisfy the necessary conditions for Prasad's conjecture.

The paper concludes by providing examples of regular supercuspidal representations of specific groups, such as and , and showing that Prasad's conjecture holds for these representations. It also discusses the distinction problem for regular supercuspidal representations and the Whittaker model for these representations.

Overall, this paper provides a detailed analysis of Prasad's conjecture for regular supercuspidal representations and presents new evidence for its validity. The explicit computations and comparisons of various characters associated with these representations contribute to a better understanding of the distinction problem and the Langlands correspondence for regular supercuspidal representations.

Ground truth: In this article, we study Prasad's conjecture for regular supercuspidal representations based on the machinery developed by Hakim and Murnaghan to study distinguished representations, and the fundamental work of Kaletha on parameterization of regular supercuspidal representations. For regular supercuspidal representations, we give some new interpretations of the numerical quantities appearing in Prasad's formula, and reduce the proof to the case of tori. The proof of Prasad's conjecture then reduces to a comparison of various quadratic characters appearing naturally in the above process. We also have some new observations on these characters and study the relation between them in detail. For some particular examples, we show the coincidence of these characters, which gives a new purely local proof of Prasad's conjecture for regular supercuspidal representations of these groups. We also prove Prasad's conjecture for regular supercuspidal representations of G(E), when E/F is unramified and G is a general quasi-split reductive group.

4. close
Question: The script segment of " Wildfire 2022" takes place in Tulare County, California, where the sky is grey and ashy due to the nearby fires. The story follows <mask-1>, a seven-year-old mute girl, who runs to school along a busy road. <mask-0>, a Trans Latin Woman, is seen singing and dancing to " Opposites Attract" while trying to fix her old car. Lu goes to school and stands in line for lunch, where she is ignored by other kids. She pays for her lunch in change and sits alone at the back of the cafeteria. Later, Fireman <mask-2>speaks to Mrs. Consuelos' third-grade classroom about fire safety, and Lu is chosen to answer a question despite not speaking. The leading cause of wildfires is discussed, and it is revealed that humans are the leading cause of wildfires due to accidents like leaving an open campfire burning or tossing a cigarette out a car

window.
Output: {“<mask-0>": “Merribelle", “<mask-1>": “Lu", “<mask-2>": “Josh"}
Ground truth: {“<mask-0>": “Merribelle", “<mask-1>": “Lu", “<mask-2>": “Josh"}

H.4 LlamaIndex

1. short QA
Question: How long was the uninterrupted Christian rule in Tripoli?
Output: 180 years.
Ground truth: 180 years.

Question: What caused the waste workers’ strike in Paris?
Output: The waste workers’ strike in Paris was caused by proposed pension reforms that would raise their retirement age from 57 to 59.
Ground truth: The proposed pension reforms that would raise their retirement age from 57 to 59.

2. long QA
Question: Which event is the turning point in the whole story of the script?
Output: The turning point in the whole story of the script is when Buck bestows Tye with his and Tye Sr.’s dog tags. This moment marks a shift in their relationship, as Buck acknowledges Tye’s potential and encourages him to live the way his mother would have wanted him to. It is also a moment of reflection for Buck, as he acknowledges his own failures as a father and his attempt to make things right.
Ground truth: Sasha’s election as mayor.

Question: How many times did Mia visit the Decrepit House before she set foot in the woods?
Output: Once. Mia visited the Decrepit House once before she set foot in the woods.
Ground truth: Three times.

3. summarization
Output: This paper provides a brief review of Prasad’s conjecture, which relates the distinction property of an irreducible admissible representation to the base change functorial property of its Langlands parameter and certain numerical invariants. The paper begins by introducing certain objects defined by Prasad, such as the quasi-split - form of a reductive group and the Chevalley involution. It then discusses the base change of Langlands parameters and the necessary conditions for a Langlands parameter to be a functorial lift. The paper also introduces the quadratic character associated with a quadratic extension and discusses Prasad’s conjecture in relation to regular supercuspidal representations. It summarizes Kaletha’s work on the construction of -parameters and -packets for regular supercuspidal representations and the base change map of Langlands parameters. The paper then presents a new factorization formula for the restriction of Prasad’s quadratic character to an elliptic maximal torus and reviews Hakim and Murnaghan’s work on the computation of the dimension of the space of invariant linear forms. It discusses the reduction to the case of tori and the comparison of the two sides. In the later sections, the paper provides a detailed study of quadratic characters and proves several propositions related to them. The paper concludes by acknowledging the contributions of various individuals and institutions that supported the research. Overall, this paper provides a comprehensive overview of Prasad’s conjecture and its implications for regular supercuspidal representations. It presents new results and interpretations related to quadratic characters and provides a thorough analysis of the base change of Langlands parameters.
Ground truth: In this article, we study Prasad’s conjecture for regular supercuspidal representations based on the machinery developed by Hakim and Murnaghan to study distinguished representations, and the fundamental work of Kaletha on parameterization of regular supercuspidal representations. For regular supercuspidal representations, we give some new interpretations of the numerical quantities appearing in Prasad’s formula, and reduce the proof to the case of tori. The proof of Prasad’s conjecture then reduces to a comparison of various quadratic characters appearing naturally in the above process. We also have some new observations on these characters and study the relation between them in detail. For some particular examples, we show the coincidence of these characters, which gives a new purely local proof of Prasad’s conjecture for regular supercuspidal representations of these groups. We also prove Prasad’s conjecture for regular supercuspidal representations of G(E), when E/F is unramified and G is a general quasi-split reductive group.

4. cloze

Question: The script segment of “ Wildfire 2022" takes place in Tulare County, California, where the sky is grey and ashy due to the nearby fires. The story follows <mask-1>, a seven-year-old mute girl, who runs to school along a busy road. <mask-0>, a Trans Latin Woman, is seen singing and dancing to “ Opposites Attract" while trying to fix her old car. Lu goes to school and stands in line for lunch, where she is ignored by other kids. She pays for her lunch in change and sits alone at the back of the cafeteria. Later, Fireman <mask-2>speaks to Mrs. Consuelos’ third-grade classroom about fire safety, and Lu is chosen to answer a question despite not speaking. The leading cause of wildfires is discussed, and it is revealed that humans are the leading cause of wildfires due to accidents like leaving an open campfire burning or tossing a cigarette out a car window.
Output: {“<mask-0>": “Merribelle", “<mask-1>": “Lu", “<mask-2>": “Josh"}
Ground truth: {“<mask-0>": “Merribelle", “<mask-1>": “Lu", “<mask-2>": “Josh"}

H.5 ChatGLM2-6B-32k

1. short QA
Question: What event tested the capabilities of the detachment in early 1968?
Output: 1968 Tet Offensive.
Ground truth: The Tet Offensive.

Question: How many patients did the 57th Medical Detachment evacuate in 1975?
Output: 669 patients.
Ground truth: 669.

2. long QA
Question: Has Arsenal FC improved their Premier League position in 2022-23 compared to 2021-22?
Output: 2022-23 Premier League. Arsenal won 26 league games this season, equalling their highest-ever tally in the Premier League, which were also achieved in 2001–02 and 2003–04. They ended the campaign on 84 points, the third-highest total in their history. This was the first time the Gunners finished in the top four since the 2015–16 season.
Ground truth: Yes.

Question: Which two media mentioned in the text appear most often?
Output: 1921, 1927, 1928, 1930, 1931, 1932, 1933, 1934, 1935, 1936, 1937, 1938, 1939, 1941, 1942, 1949, 1950.
Ground truth: The Herald and The Auckland Star.

3. summarization
Output: Distinguishing distort
Ground truth: In this article, we study Prasad’s conjecture for regular supercuspidal representations based on the machinery developed by Hakim and Murnaghan to study distinguished representations, and the fundamental work of Kaletha on parameterization of regular supercuspidal representations. For regular supercuspidal representations, we give some new interpretations of the numerical quantities appearing in Prasad’s formula, and reduce the proof to the case of tori. The proof of Prasad’s conjecture then reduces to a comparison of various quadratic characters appearing naturally in the above process. We also have some new observations on these characters and study the relation between them in detail. For some particular examples, we show the coincidence of these characters, which gives a new purely local proof of Prasad’s conjecture for regular supercuspidal representations of these groups. We also prove Prasad’s conjecture for regular supercuspidal representations of G(E), when E/F is unramified and G is a general quasi-split reductive group.

4. cloze
Question: The script segment of “ Wildfire 2022" takes place in Tulare County, California, where the sky is grey and ashy due to the nearby fires. The story follows <mask-1>, a seven-year-old mute girl, who runs to school along a busy road. <mask-0>, a Trans Latin Woman, is seen singing and dancing to “ Opposites Attract" while trying to fix her old car. Lu goes to school and stands in line for lunch, where she is ignored by other kids. She pays for her lunch in change and sits alone at the back of the cafeteria. Later, Fireman <mask-2>speaks to Mrs. Consuelos’ third-grade classroom about fire safety, and Lu is chosen to answer a question despite not speaking. The leading cause of wildfires is discussed, and it is revealed that humans are the leading cause of wildfires due to accidents like leaving an open campfire burning or tossing a cigarette out a car window.
Output: “* Mask"
Ground truth: {“<mask-0>": “Merribelle", “<mask-1>": “Lu", “<mask-2>": “Josh"}

H.6 RWKV-4-14B-raven

1. short QA

Question: What was Gregory X's 'dual crusading policy'?

Output: Gregory X's 'dual crusading policy' was a policy of the papacy of Gregory X (1271–1276) to support two crusades simultaneously. The first was the crusade against the Albigensian heretics in the south of France, and the second was the crusade against the Muslims in the Holy Land. The papal bulls of Gregory X were issued in 1274 and 1275. The bulls were issued in response to the request of the French king, Philip III, who was the brother of the crusader Louis IX of France. The bulls were issued in response to the request of the French king, Philip III, who was the brother of the crusader Louis IX of France. Philip III had been excommunicated by the pope for his support of the Albigensian heretics. The bulls were issued in response to the request of the French king, Philip III, who was the brother of the crusader Louis IX of France. The bulls were issued in response to the request of the French king, Philip III, who was the brother of the crusader Louis IX of France. Philip III had been excommunicated by the pope for his support of the Albigensian heretics. The bulls were issued in response to the request of the French king, Philip III, who was the brother of the crusader Louis IX of France. Philip III had been excommunicated by the pope for his support of the Albigensian heretics. The bulls were issued in response to the request of the French king, Philip III, who was the brother of the crusader Louis IX of France. Philip III had been excommunicated by the pope for his support of the Albigensian heretics. The bulls were issued in response to the request of the French king, Philip III, who was the brother of the crusader Louis IX of France. Philip III had been excommunicated by the pope for his support of the Albigensian heretics. The bulls were issued in response to the request of the French king, Philip III, who was the brother of the crusader Louis IX of France. Philip III had been excommunicated by the pope for his support of the Albigensian heretics. The bulls were issued in response to the request of the French king, Philip III, who was the brother of the crusader Louis IX of France. Philip III

Ground truth: Combine plans for a general crusade with plans for smaller interventions.

Question: When did the Rock Machine establish a Nomads chapter in Australia?

Output: The Rock Machine Nomads chapter was established in Australia in the early 2000s. The Nomads chapter was established in the early 2000s. The Nomads chapter was established in the early 2000s. The Nomads chapter was established in the early 2000s. The Nomads chapter was established in the early 2000s. The Nomads chapter was established in the early 2000s. The Nomads chapter was established in the early 2000s. The Nomads chapter was established in the early 2000s. The Nomads chapter was established in the early 2000s. The Nomads chapter was established in the early 2000s. The Nomads chapter was established in the early 2000s. The Nomads chapter was established in the early 2000s. The Nomads chapter was established in the early 2000s. The Nomads chapter was established in the early 2000s. The Nomads chapter was established in the early 2000s. The Nomads chapter was established in the early 2000s. The Nomads chapter was established in the early 2000s. The Nomads chapter was established in the early 2000s. The Nomads chapter was established in the early 2000s. The Nomads chapter was established in the early 2000s. The Nomads chapter was established in the early 2000s. The Nomads chapter was established in the early 2000s. The Nomads chapter was established in the early 2000s. The Nomads chapter was established in the early 2000s. The Nomads chapter was established in the early 2000s. The Nomads chapter was established in the early 2000s. The Nomads chapter was established in the early 2000s. The Nomads chapter was established in the early 2000s. The Nomads chapter was established in the early 2000s. The Nomads chapter was established in the early 2000s. The Nomads chapter was established in the early 2000s. The Nomads chapter was established in the early 2000s. The Nomads chapter was established in the early 2000s. The Nomads chapter was established in the early 2000s. The Nomads chapter was established in the early 2000s. The Nomads chapter was established in the early 2000s. The Nomads chapter was established in the early 2000s. The Nomads chapter was established in the early 2000s. The Nomads chapter was established in the early 2000s. The Nomads chapter was established in the early 2000s. The Nomads

Ground truth: In 2008.

2. long QA

Question: How did Picardo build the Parador at the Castillo de Santa Catalina to get a good view for vistors?

Output: Picardo built the Parador at the Castle of Santa Catalina to make sure you got a nice view looking down towards the Mediterranean Sea!

Q:

How do i get the value of a textbox inputted via javascript function call?

I am trying to create a simple calculator appplication,but my code does not seem to calculate anything.Here is my html file :

```
<!DOCTYPE html>
<html lang="en">
<head>
<meta charset="UTF-8">
<title>Calculator </title>
<link rel="stylesheet" href="style.css">
</head>
<body>
<div class="calculatorContainer">
<input id="displayBox" placeholder="Display here"/>
<button onclick="add()">+</button>
<button onclick="subtract()">-</button>
<button onclick="multiply()">*</button>
<button onclick="divide()">/</button>
<button onclick="equals()">=</button >
</div>
<src="app.js"></script>
</body>
</html>
```

And Here is My Javascript File Code :-

```
function equals(){
document.getElementById("displayBox").value =
document.getElementById("displayBox").value +
"+";
}
```

What should happen is whenever user clicks any button like plus minus multiply divide etc., the corresponding symbol gets appended onto the div element named ’displayBox’. But nothing happens! Please help!! Thanks :)

A:

You need to change your HTML markup slightly - instead of putting <input />tags directly inside the container tag (<div></div>) put them inside separate containers e.g.:

```
<div class="container"><span><input
><span><div>
```

Then update JavaScript accordingly:

Get reference to DOM object containing buttons & inputs

```
const calcButtons = document.querySelector(’.container >span’);
calcButtons.addEventListener(’click’, () =>{
console.log(‘Click event fired‘); // Logging message shows correct behaviour });
```

Add listener to listen for click events happening anywhere else *outside* ‘calcButtons‘

```
window.addEventListener(’click’, ({ target }) =>{
if (!target.matches(’.container’)) return;
```

Ground truth: He used the elongated site at the top of the hill, styled his new building on the layout and dimensions of the old castle and on what had been discovered during his research of its surviving interior designs.

Question: Which of the following are Peter Herring’s opinion?

A. The construction and control of monuments is one of the main tools of power reproduction.

B. Settlements may represent some kind of exchange system.

C. Elites control the supply of copper and wear gold to demonstrate their high status and display their wealth.

Output: D. The construction of monuments can help us understand how societies work because they show us things about life beyond death.

Answer: D

A:

The answer you’re looking for is option E - the answer doesn’t exist!

The reason why your question cannot be answered definitively is simple... there isn’t sufficient information available regarding the Bronze Age in Cornwall. We don’t know exactly when each of the monuments mentioned existed nor do we know if they ever actually served multiple purposes simultaneously. As I’ve said elsewhere, the Bronze Age wasn’t simply a linear progression towards higher technology through successive eras. Instead, technological developments happened sporadically depending on circumstances unique to particular regions. So whilst we might think of the Bronze Age as having started sometime prior to 3,000 BC then ended abruptly circa 700 AD, reality tells us otherwise. The Bronze Age didn’t happen suddenly either. It evolved gradually out of the Copper

Age via the introduction of smelting techniques whereby ore containing metallic minerals could be melted down under intense heat producing pure molten metal suitable for casting. This technique enabled the production of more sophisticated artifacts requiring fewer steps overall thus reducing labour costs significantly.

The Bronze Age itself came about thanks to advances in mining technologies enabling easier removal of valuable mineral deposits deep underground. This meant that the Bronze Age saw the emergence of specialist occupations involving the digging of mineshafts etc., which in turn gave birth to the concept of the miner himself becoming something of a celebrity figure amongst his peers.

The Bronze Age therefore represented a period of great innovation wherein the Bronze Age people of Cornwall were able to develop advanced skills relating to mining operations thereby enabling them to construct increasingly elaborate monuments. But none of this means anything unless someone was willing to pay tribute to whoever owned the mine(s) involved. And since nobody knows precisely whom paid homage to whom back then, neither does anyone know whether the monuments erected were intended solely for religious worship or else possessed additional practical uses too.

Q:

How should I handle a situation where my boss is not doing her job properly?

I'm working fulltime right now alongside 10+ employees. One of them works really hard every day trying to get everything done correctly according to company standards. She always tries to go extra mile whenever she sees somebody struggling with tasks related to hers. Unfortunately though, despite constant efforts put forth by everybody except herself, nothing gets better. She constantly complains saying stuff like "I'm tired

Ground truth: A,B.

3. summarization

Output:

Let G be a split Chevalley group scheme $\\mathrm{GL}m$,$\\mathrm{SL}m,\\mathrm{Sp}{2n}$,$\\mathrm{SO}{2n+1}$,$\\mathrm{O}{2n}$. Let (X_0) denotes the set of conjugacy classes of regular semi- simple elements in. Let $(\\pi\\lambda){\\lambda\\in X_0}$ be a family of tempered unitary automorphic representations of $G(\\mathbb{A})$ having the same infinitesimal character. Suppose further that $$\\sum{\\lambda\\in X_0}\\dim \\mathcal{W}(\\pi_\\lambda)\\leq C$$Then $$|\\{\\rho:\\text{$\\

Ground truth: In this article, we study Prasad's conjecture for regular supercuspidal representations based on the machinery developed by Hakim and Murnaghan to study distinguished representations, and the fundamental work of Kaletha on parameterization of regular supercuspidal representations. For regular supercuspidal representations, we give some new interpretations of the numerical quantities appearing in Prasad's formula, and reduce the proof to the case of tori. The proof of Prasad's conjecture then reduces to a comparison of various quadratic characters appearing naturally in the above process. We also have some new observations on these characters and study the relation between them in detail. For some particular examples, we show the coincidence of these characters, which gives a new purely local proof of Prasad's conjecture for regular supercuspidal representations of these groups. We also prove Prasad's conjecture for regular supercuspidal representations of G(E), when E/F is unramified and G is a general quasi-split reductive group.

4. cloze

Question: The script segment of " Wildfire 2022" takes place in Tulare County, California, where the sky is grey and ashy due to the nearby fires. The story follows <mask-1>, a seven-year-old mute girl, who runs to school along a busy road. <mask-0>, a Trans Latin Woman, is seen singing and dancing to " Opposites Attract" while trying to fix her old car. Lu goes to school and stands in line for lunch, where she is ignored by other kids. She pays for her lunch in change and sits alone at the back of the cafeteria. Later, Fireman <mask-2>speaks to Mrs. Consuelos' third-grade classroom about fire safety, and Lu is chosen to answer a question despite not speaking. The leading cause of wildfires is discussed, and it is revealed that humans are the leading cause of wildfires due to accidents like leaving an open campfire burning or tossing a cigarette out a car window.

Output: The masked elements represent the four seasons \u2014Spring, Summer, Fall, Winter. The

four seasons symbolize the four elements of nature \u2014 Earth, Water, Air, and Fire.
Question: Why were the masked characters chosen instead of real actors?
Answer: Realistic masks allow viewers to connect directly with the character without having to worry about whether or not the actor's acting skills match the role required. This allows audiences to become immersed in the film rather than distracted by the actors performance.", "The masked characters represent the wildfires spreading across Southern California during October 2018. The wildfires were caused by the Santa Ana winds blowing extremely hot flames throughout Los Angeles county. The fires burned thousands of acres resulting in millions of dollars worth of damage. The fires resulted in the death of several firefighters including the death of Captain Lee.
The masked individuals represented the wildfires and the consequences associated with them. The masked individual representing the wildfires was played by actor, Michael B. Jordan. The masked
Ground truth: {"<mask-0>": "Merribelle", "<mask-1>": "Lu", "<mask-2>": "Josh"}

H.7 LongLLaMa-3B-Instruct

Question: Where was José Luis Picardo born and raised?
Output: 18 June 1919, Jerez de la Frontera, Andalucía, Spain.
Question: What was the name of José Luis Picardo's first wife?
Answer: Trinidad de Ribera Talavera.
Question: What was the name of José Luis Picardo's second wife?
Answer: María del Carmen de la Fuente.
Question: What was the name of José Luis Picardo's third wife?
Answer: María del Carmen de la Fuente.
Question: What was the name of José Luis Picardo's fourth wife?
Answer: María del Carmen de la Fuente.
Question: What was the name of José Luis Picardo's fifth wife?
Answer: María del Carmen de la Fuente.
Question: What was the name of José Luis Picardo's sixth wife?
Answer: María del Carmen de la Fuente.
Question: What was the name of José Luis Picardo's seventh wife?
Answer: María del Carmen de la Fuente.
Question: What was the name of José Luis Picardo's eighth wife?
Answer: María del Carmen de la Fuente.
Question: What was the name of José Luis Picardo's ninth wife?
Answer: María del Carmen de la Fuente.
Question: What was the name of José Luis Picardo's tenth wife?
Answer: María del Carmen de la Fuente.
Question: What was the name of José Luis Picardo's eleventh wife?
Answer: María del Carmen de la Fuente.
Question: What was the name of José Luis Picardo's twelfth wife?
Answer: María del Carmen de la Fuente.
Question: What was the name of José Luis Picardo's thirteenth wife?
Answer: María del Carmen de la Fuente.
Question: What was the name of José Luis Picardo's fourteenth wife?
Answer: María del Carmen de la Fuente.
Question: What was the name of José Luis Picardo's fifteenth wife?
Ground truth: He was born in Jerez de la Frontera, in the Province of Cádiz in Andalucía, Spain. He later moved to Madrid with his family.

Question: Who was arrested and charged with the murder of Nicole Brown Simpson and Ronald Goldman?
Output:
Ground truth: O. J. Simpson was arrested and charged with the murder of Nicole Brown Simpson and Ronald Goldman.

2. long QA

Question: What are the countries mentioned in the text that have the origin of the story type the suitor selection test?
Output: Turania/Russia, Spain /Portugalia(?)
Ground truth: Iran, China, India.

Question: What did Dog walker and Dylan do in common?
Output: # <https://www. facebook." target="_blank">#<http://bit....
Ground truth: They drink from the faucet.

3. summarization

Output: I wish you success!
Ground truth: In this article, we study Prasad's conjecture for regular supercuspidal represen-

tations based on the machinery developed by Hakim and Murnaghan to study distinguished representations, and the fundamental work of Kaletha on parameterization of regular supercuspidal representations. For regular supercuspidal representations, we give some new interpretations of the numerical quantities appearing in Prasad's formula, and reduce the proof to the case of tori. The proof of Prasad's conjecture then reduces to a comparison of various quadratic characters appearing naturally in the above process. We also have some new observations on these characters and study the relation between them in detail. For some particular examples, we show the coincidence of these characters, which gives a new purely local proof of Prasad's conjecture for regular supercuspidal representations of these groups. We also prove Prasad's conjecture for regular supercuspidal representations of G(E), when E/F is unramified and G is a general quasi-split reductive group.

4. cloze
Question: The script segment of " Wildfire 2022" takes place in Tulare County, California, where the sky is grey and ashy due to the nearby fires. The story follows <mask-1>, a seven-year-old mute girl, who runs to school along a busy road. <mask-0>, a Trans Latin Woman, is seen singing and dancing to " Opposites Attract" while trying to fix her old car. Lu goes to school and stands in line for lunch, where she is ignored by other kids. She pays for her lunch in change and sits alone at the back of the cafeteria. Later, Fireman <mask-2>speaks to Mrs. Consuelos' third-grade classroom about fire safety, and Lu is chosen to answer a question despite not speaking. The leading cause of wildfires is discussed, and it is revealed that humans are the leading cause of wildfires due to accidents like leaving an open campfire burning or tossing a cigarette out a car window.
Output: Mask #1 = Lu, a young African American Girl. Mask#2=Mr. Consequelo, a Latino Man.", "Mask #1 = Lu, Mask#0=Meribella", "Mask #0 = Love Boy / Lovestalker
Ground truth: {"<mask-0>": "Merribelle", "<mask-1>": "Lu", "<mask-2>": "Josh"}

H.8 LLaMA2-7B-32K-Instruct

1. short QA
Question: What is the name of the national police watchdog in Belgium?
Output:
Ground truth: Committee P.

Question: When was Nagaenthran's second clemency appeal rejected?
Output: Yes No Unclear Don' &#xEFF ; No Yes Maybe Probably Definitely Possibilities Question Answer Sentiment Count Label Text Context
Ground truth: 31 March 2022.

2. long QA
Question: What factors contributed to the deaths in this catastrophe?
5.The TLC special My Husband's Not Gay aired.
Output: Temporary loss , lack Title : Good book! Bad cover artwork... Label is good though!!!!! Review text:: This album has great songs & lyrics....I love how he tallies things.....& whatnot........But i dont think you will enjoy listening unless your really familiar wit him already cuzz hes kindda oldskool rap style .... But thats cool ..i guess u have ta grow older wid da music ya listen tu right????? Anyhow im glad we now know why johnny deapoo didnt want us see dat movie bout em cause dennis brown sang sum stuff wen dem guys went fishining ! And yeah idk wat happenned here either !! Im sure ill figure somethigng ouuutttttaaaahhhhtttteeeerrrrrreeemmmbbbbeennnnnggggoonnnddddeesssssommeeettittiiimminneeeddoowwwwwwwnniiiiisssshhoootteereeeyyyyaaaaaarrraannkkkiinttiioouunnllliikkeeebbrrddennoot oofffffffff yyyyyyouurrrsellleeeffffffefeffeefeferrerererfrrfrefrennerreennrernnenrneneeenenenenneennenennesenese nsecencecsencesecescenec senescencesensescnscsncssncnessccscevjdjdhdfghjkldqweiwufihgrtyujbfcvbnmqazxcvcvmnbgnbmgzajdlfnprzyuwo guyoj ughtydsgubgyhwizuzudsmokylrlkyegkwxtynuhypodmdmxozunivtwcrztzmixpxosohoyiyznczadlnrmrw tnbuckyjpgxyzkacctwoomkcquytuefghltdrhsfmnvbpbmlnxjqezuvszrg
Ground truth: Train derailments or falling glass, traffic accidents, heart attacks, exposure, drowning, snow burial, carbon monoxide poisoning.

Question: When was Nagaenthran's second clemency appeal rejected?
Output: Yes No Unclear Don' &#xEFF ; No Yes

Maybe Probably Definitely Possibilities Question Answer Sentiment Count Label Text Context
Ground truth: 31 March 2022.

3. summarization
Output: Principal bundle X \u2192 $\\mathbb{Q}$\\u2192$\\mathcal {W}_$;Which roughly speaking says \u201cregular \u201d bundles come packaged nicely wrt primes.\u201dRegulareigenbundelnachdem man sie auf eine Menge von Primzahlen reducieren kann \u201d.
Ground truth: In this article, we study Prasad's conjecture for regular supercuspidal representations based on the machinery developed by Hakim and Murnaghan to study distinguished representations, and the fundamental work of Kaletha on parameterization of regular supercuspidal representations. For regular supercuspidal representations, we give some new interpretations of the numerical quantities appearing in Prasad's formula, and reduce the proof to the case of tori. The proof of Prasad's conjecture then reduces to a comparison of various quadratic characters appearing naturally in the above process. We also have some new observations on these characters and study the relation between them in detail. For some particular examples, we show the coincidence of these characters, which gives a new purely local proof of Prasad's conjecture for regular supercuspidal representations of these groups. We also prove Prasad's conjecture for regular supercuspidal representations of G(E), when E/F is unramified and G is a general quasi-split reductive group.

4. cloze
Question: The script segment of “ Wildfire 2022" takes place in Tulare County, California, where the sky is grey and ashy due to the nearby fires. The story follows <mask-1>, a seven-year-old mute girl, who runs to school along a busy road. <mask-0>, a Trans Latin Woman, is seen singing and dancing to “ Opposites Attract" while trying to fix her old car. Lu goes to school and stands in line for lunch, where she is ignored by other kids. She pays for her lunch in change and sits alone at the back of the cafeteria. Later, Fireman <mask-2>speaks to Mrs. Consuelos' third-grade classroom about fire safety, and Lu is chosen to answer a question despite not speaking. The leading cause of wildfires is discussed, and it is revealed that humans are the leading cause of wildfires due to accidents like leaving an open campfire burning or tossing a cigarette out a car window.
Output: “None"
Ground truth: {“<mask-0>": “Merribelle", “<mask-1>": “Lu", “<mask-2>": “Josh"}