LV-Eval: A Balanced Long-Context Benchmark with 5 Length Levels Up to 256K

Tao Yuan[1], **Xuefei Ning**[2,†], **Dong Zhou**[1], **Zhijie Yang**[1], **Shiyao Li**[1,2], **Minghui Zhuang**[1], **Zheyue Tan**[1], **Zhuyu Yao**[1], **Dahua Lin**[3,4], **Boxun Li**[1,*], **Guohao Dai**[1,5,†], **Shengen Yan**[1], **Yu Wang**[2,†]
[1] Infinigence-AI, [2] Tsinghua University, [3] Shanghai Artificial Intelligence Laboratory
[4] The Chinese University of Hong Kong, [5] Shanghai Jiao Tong University

Abstract

State-of-the-art large language models (LLMs) are now claiming remarkable supported context lengths of $256k$ or even more. In contrast, the average context lengths of mainstream benchmarks are insufficient ($5k$-$21k$), and they suffer from potential knowledge leakage and inaccurate metrics, resulting in biased evaluation. This paper introduces *LV*-Eval, a challenging long-context benchmark with five length levels ($16k$, $32k$, $64k$, $128k$, and $256k$) reaching up to $256k$ words. *LV*-Eval features two main tasks, single-hop QA and multi-hop QA, comprising 11 bilingual datasets. The design of *LV*-Eval has incorporated three key techniques, namely confusing facts insertion (CFI), keyword and phrase replacement (KPR), and keyword-recall-based metric design. The advantages of *LV*-Eval include controllable evaluation across context lengths, challenging test instances with confusing facts, mitigated knowledge leakage, and more objective evaluation. We evaluate 12 LLMs on *LV*-Eval and conduct ablation studies on the benchmarking techniques. The results reveal that: (i) Commercial LLMs generally outperform open-source LLMs when evaluated within length levels shorter than their claimed context length. However, their overall performance is surpassed by open-source LLMs with longer context lengths. (ii) Extremely long-context LLMs, such as Yi-6B-200k and Llama3-8B-1M, exhibit a relatively gentle degradation of performance, but their absolute performances may not necessarily be higher than those of LLMs with shorter context lengths. (iii) LLMs' performances can significantly degrade in the presence of confusing information, especially in the pressure test of "needle in a haystack". (iv) Issues related to knowledge leakage and inaccurate metrics introduce bias in evaluation, and these concerns are alleviated in *LV*-Eval. All datasets and evaluation codes are released at: `https://github.com/infinigence/LVEval`.

1 Introduction

Large language models (LLMs) have demonstrated exceptional performance on a variety of natural language processing tasks. The ability of long-context understanding is crucial for LLMs to deal with tasks based on longer contexts, such as books, lengthy chat history, and so on. Recently, extensive efforts have been devoted in enlarging the supported context length (i.e., the number of tokens that the model can accept as input) of LLMs. These efforts have pushed the supported context length of LLMs from $2k$ tokens to $32k$ tokens [1, 2, 3, 4, 5], and some models have achieved a remarkable context length of $128k$ and $200k$ [6, 7, 8].

*Project Lead
†Correspondence to Yu Wang <yu-wang@tsinghua.edu.cn>, Xuefei Ning <foxdoraame@gmail.com>, Guohao Dai <daiguohao@sjtu.edu.cn, daiguohao@infini-ai.com>.

Submitted to the 38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks. Do not distribute.

Benchmark	#Datasets	Avg #Words	Min/Max Words	Length Levels	Opt. Metric	Lang.
ZeroSCROLLS [12]	10	13,556	1,023/320,774	none		en
LooGLE [10]	7	21,247	10,927/246,182	none		en
L-Eval [11]	20	12,993	2,119/170,256	none	✓	en
BAMBOO [13]	10	5,067	229/14,858	$4k$,$16k$		en+zh
LongBench [9]	21	9,486	128/71,954	0-$4k$,$4k$-$8k$,$8k$+		en+zh
LV-Eval	11	102,380	11,896/387,406	$16k$,$32k$,$64k$,$128k$,$256k$	✓	en+zh

Table 1: Comparison of different long-context benchmarks. We count the number of words for the English datasets and the number of characters for the Chinese datasets. The punctuation marks are taken into account, while tabs, blank spaces, and newlines are not included.

In contrast to the rapid evolution of the models' supported context length, existing benchmarks have lagged behind. The average word count in current long-context benchmarks typically falls within the range of $32k$ [9, 10, 11, 12, 13], considerably shorter compared to the supported context lengths of state-of-the-art long-context models. Moreover, previous benchmarks primarily consist of *unaltered* public documents and articles. This could be problematic for two reasons: (i) the data might be involved in LLMs' training processes, and (ii) the facts within them might be common-sense facts found in other training resources. The presence of this issue, known as "knowledge leakage" [14], can lead to models answering questions with memorization or common-sense knowledge instead of understanding long-range contexts. Last but not least, the automatic metrics employed in most of the existing benchmarks are susceptible to the variations in answer format and the inclusion of irrelevant words. Such metrics struggle to accurately assess the answer quality.

To address these issues, we propose *LV*-Eval, a bilingual benchmark with up to $256k$ words. *LV*-Eval incorporates distractions and confusions to make the test more challenging, replaces keywords and rephrases sentences to prevent knowledge leakage, and employs a more accurate metric. We summarizes the key characteristics of *LV*-Eval as follows:

- **Sufficiently long context length to evaluate state-of-the-art models**: *LV*-Eval comprises 5 length levels with word counts of $16k$, $32k$, $64k$, $128k$, and $256k$. Test instances across these levels share the same set of question-answer (QA) pairs, and only differ in the context content and length. Testing on the same QA pairs with different context lengths facilitates a controllable evaluation of models' long-context ability.
- **Incorporation of distraction and confusion to increase difficulty**: When constructing the context for each test instance, we mix up distracting documents and supporting documents. This approach evaluates the model's ability in pinpointing key information in a large bunch of distracting texts. In addition, we insert confusing facts generated by GPT-4 and revised by human annotators into the context. This assesses the model's capability to accurately reason in the presence of interference.
- **Keyword and phrase replacement to mitigate knowledge leakage**: To mitigate the biased evaluation of long-context ability caused by knowledge leakage, we replace the keywords and phrases in the context and QA pairs. The replacement rules are annotated by human annotators. In this way, *LV*-Eval requires LLMs to rely on the understanding of context to answer questions rather than relying on memorization or common-sense knowledge.
- **Keyword-recall-based metric for more objective scoring**: Existing N-gram metrics such as the F1 score are sensitive to the format variations and non-informative words in the answer, which results in inaccurate scores. To address this, we manually annotate answer keywords and a blacklist of unrelated words. The golden answers are the critical words or sentences extracted from original ground-truth (GT) answers, while the word blacklist contains common and non-informative words such as 'the', 'a', 'of', and so on. The metric calculation follows a two-stage procedure: the first stage calculates the recall of golden answer keywords. if the recall exceeds a certain threshold, the second stage will remove all the blacklisted words and then calculate the F1 score between the prediction and the GT answer. This metric design can get scores with higher objectivity.

Findings. We evaluate 12 LLMs on *LV*-Eval and summarize the main findings as follows: (i) Commercial LLMs generally outperform open-source LLMs when evaluated within length levels shorter than their claimed context length. However, their overall performance is surpassed by

Task	Dataset	CFI	#KPR	AK	Language	#QA pairs	#Contexts
Single-hop QA	**lic-mixup**	✓		✓	zh	197	985
	loogle-SD-mixup			✓	en	160	800
	cmrc-mixup		786		zh	200	1,000
	multifieldqa-en-mixup	✓	476	✓	en	101	505
	multifieldqa-zh-mixup	✓	424	✓	zh	133	665
	factrecall-en	✓	3	✓	en	1	200×5
	factrecall-zh	✓	3	✓	zh	1	200×5
Multi-hop QA	**dureader-mixup**				zh	176	880
	loogle-CR-mixup			✓	en	99	495
	loogle-MR-mixup			✓	en	139	695
	hotpotwikiqa-mixup	✓	232	✓	en	124	620

Table 2: Data statistics of *LV*-Eval. The abbreviations "CFI", "KPR", "AK" stand for "Confusing Fact Insertion", "Keyword and Phrase Replacement", and "Answer Keywords", respectively. "#KPR" is the number of KPR rules. Note that in **factrecall-en** and **factrecall-zh**, all QA pairs are the same across all test instances, i.e., there is only one unique QA pair for each of the two datasets.

open-source LLMs with longer context lengths. (ii) Extremely long-context LLMs, such as Yi-6B-200k and Llama3-8B-1M, exhibit a relatively gentle degradation of performance, but their absolute performances may not necessarily be higher than those of LLMs with shorter context lengths. (iii) LLMs' performances can significantly degrade in the presence of confusing information, especially in the pressure test of "needle in a haystack". (iv) Issues related to knowledge leakage and inaccurate metrics introduce bias in evaluation, and these concerns are alleviated in *LV*-Eval.

2 Related Work

Long-Context Benchmarks. Table 1 provides a summary of existing long-context benchmarks, including ZeroScrolls [12], LooGLE [10], L-Eval [11], BAMBOO [13], and LongBench [9]. ZeroScrolls, LooGLE, and L-Eval are monolingual benchmarks without explicit length level partition. Their average word counts are $\sim 14k$, $\sim 21k$ and $\sim 13.5k$, respectively. In order to evaluate the model's capability across various context lengths, BAMBOO and LongBench have designed various length levels. However, the word counts ($\sim 5k$, $\sim 9.5k$) of the contexts in these two benchmarks are notably smaller than the supported context length of state-of-the-art long-context models, making them unsuitable for evaluating the claimed extremely long-context understanding ability. In contrast, *LV*-Eval contains five length levels, up to $256k$ words, each with the same set of QA pairs for controllable evaluation.

In terms of metric design, L-Eval introduces a length-instruction-enhanced metric to mitigate the undesired impact of the answer length on metric scores. Additionally, L-Eval proposes to use LLMs to assist in scoring. In *LV*-Eval, we ask human annotators to mark the answer keywords and create a non-informative word blacklist, and propose a two-stage metric to focus more on the answer keywords while reducing the influences of non-informative words.

Long-Context Techniques. Considerable efforts have been devoted to enhancing the long-context abilities of LLMs. One line of work focuses on making LLMs have extended context sizes without fine-tuning and behave normally on inputs longer than their training context lengths. The design and extrapolation method of the position encoding module [15, 16, 17] is crucial for this goal. Besides, several sparse attention techniques [18, 19] have also been proposed to avoid model collapse. These sparse attention techniques also alleviate the quadratic complexity w.r.t. the sequence length.

There are many other strategies aimed at enabling LLMs to effectively leverage long input contexts. The most commonly utilized strategy is long-context fine-tuning [20, 21, 22]. For instance, YaRN [22] conducts fine-tuning with $64k$ and $128k$ context lengths starting with Llama2-7B/13B, and Yi-6B-200k [8] is trained with $200k$ context length starting with its $4k$ variant. Other strategies include the recurrent- or memory-based architecture [23, 24, 25, 26, 27], and the retrieval- or summarization-based context compression techniques [28, 29, 9, 26], and so on.

In this work, we evaluate LLMs of diverse context sizes, ranging from $4k$ to $200k$, most of which have incorporated advanced position encoding design and undergone long-context fine-tuning.

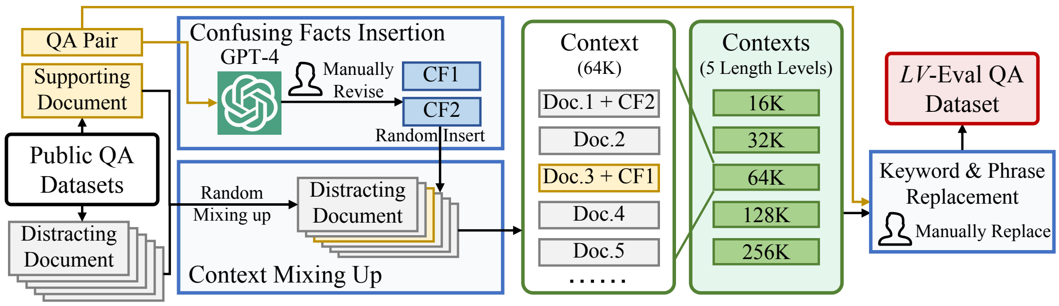

Figure 1: The construction process of *LV*-Eval. "CF" is short for "Confusing Fact".

3 *LV*-Eval Benchmark

LV-Eval focuses on two types of QA tasks: single-hop QA and multi-hop QA, and is comprised of 11 QA datasets (6 in English and 5 in Chinese). The data statistics for *LV*-Eval are outlined in Table 2. Each test instance in *LV*-Eval comprises three parts: a context (C), a question (Q), and a GT answer (A), where C is a synthetic document containing the information required to answer Q.

Datasets in *LV*-Eval are constructed with existing public datasets as the source, except for factrecall-en and factrecall-zh, which are constructed using the data from *PG19* [30] dataset and *Journey to the West* book. Each dataset consists of five subsets of different lengths: $16k$, $32k$, $64k$, $128k$, and $256k$. All five subsets share the same question-answer (QA) pairs, meaning there are five contexts of varying lengths for each QA pair. This allows for a controllable evaluation of models' long-context ability when testing the same set of questions with different context lengths. In total, *LV*-Eval comprises 1,729 QA pairs and 1,729×5 = 8,645 synthetic contexts.

Figure 2 illustrates the construction process of *LV*-Eval. For **factrecall-en** and **factrecall-zh**, we write one QA pair for each dataset. For the rest 9 out of the 11 datasets, we first choose a specific number of QA pairs from existing QA datasets (Section 3.1). Then, for each unique QA pair, we go through three procedures to construct the context (Section 3.2):

1. **Context mixing up** (Section 3.2.1): We first construct five contexts of different lengths by mixing up supporting documents corresponding to the QA pair and several distracting documents. For **factrecall-en** and **factrecall-zh**, we mix the supporting evidence of the single QA pair with distracting documents from two books. For other datasets, the distracting documents are unrelated to the question and are chosen from the context documents corresponding to non-selected QA pairs in the same source dataset.

2. **Confusing Facts Insertion (CFI)** (Section 3.2.2): Then, in some datasets, we introduce confusing facts by generating them with GPT-4, manually revising them, and randomly inserting these into the context. These confusing facts bear similarities to the original supporting facts but are factually different, without contradicting the original information. This helps make the test instances more challenging.

3. **Keyword and Phrase Replacement (KPR)** (Section 3.2.3): Finally, to reduce the impacts of knowledge leakage on evaluation results, we manually replace some keywords and phrases in the context and the QA pairs.

When evaluating the generated answer, to mitigate the bias in existing metrics, we manually annotate the keywords in the GT answer and adjust the metric to focus more on the keywords (Section 3.3).

3.1 Data Source and QA Pair Construction

We construct 11 datasets (see Table 2) using public data sources, including `Long-instruction-en2zh` [31], HotpotQA [32], 2WikiMultihopQA [33], DuReader [34], LooGLE [10], LongBench [9], CMRC 2018 [35], MultiFieldQA [9], PG-19 [30] and the book of *Journey to the West*. The construction of QA pairs in each dataset is elaborated in Appendix A.

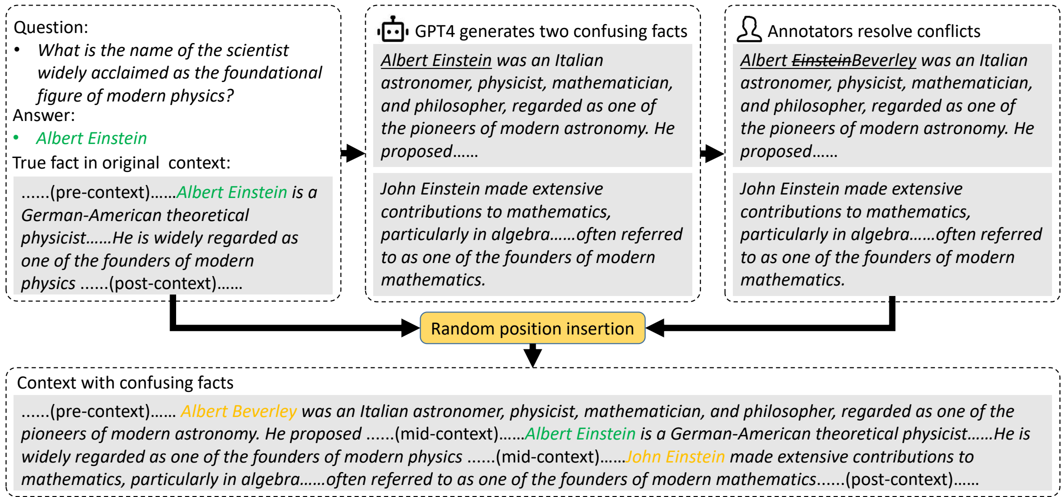

Figure 2: Steps for CFI. Firstly, we prompt GPT-4 to generate two descriptions that are close to the original fact. Then we ask human annotators to resolve any conflicts in the generated facts. For example, the first generated confusing fact "Albert Einstein was an Italian astronomer" is in conflict with the original fact and the human annotator revise it to "Albert Beverley was an Italian astronomer". Finally, the confusing facts are inserted into a randomly position in the context.

3.2 Context Construction

3.2.1 Context Mixing Up

Can the LLMs identify the key evidences to answer the target question within a long context? To assess this ability, as shown in Figure 2, *LV*-Eval randomly mixes the supporting documents with various distracting documents to generate five contexts of varying length for a given QA pair. For 9 out of the 11 datasets (excluding **factrecall-en** and **factrecall-zh**), the distracting documents are chosen from the contexts corresponding to the non-selected QA pairs in the source dataset. For **factrecall-en** and **factrecall-zh**, the distracting documents are extracted from the *PG-19* dataset and the book of *Journey to the West*.

For each length level, we sample distracting documents one by one until the cumulative word count meets the desired length level. Then, we shuffle the supporting and distracting documents, prepend a string "`Passage i\n`" to the i-th document, and concatenate them to form the final context.

Note that in **hotpotwikiqa-mixup** and **dureader-mixup**, where multiple supporting documents exist for each QA pair, instead of regarding the multiple supporting documents a single unit, we disperse and shuffle all supporting and distracting documents.

3.2.2 Confusing Facts Insertion

Can the LLMs identify the key evidences correctly if there are confusing facts in the context? To assess this ability, we apply CFI in **hotpotwikiqa-mixup**, **lic-mixup**, **multifieldqa-en-mixup**, **multifieldqa-zh-mixup**, **factrecall-en**, and **factrecall-zh**, which inserts similar, factually different, non-contradictory facts into the context. These facts might mislead less meticulous models, leading them to generate incorrect answers.

The generation process of the confusing facts goes as follows. Firstly, we use the question and answer as the input, and prompt GPT-4 [7] to generate two descriptions that are close to the original fact. The prompt for GPT-4 is shown in Figure A7. Then, we ask human annotators to resolve any conflicts in the generated facts. As illustrated in Figure 3.2, the generated confusing fact "Albert Einstein was an Italian astronomer" is in conflict with the original fact. Therefore, the human annotator revise it to "Albert Beverley was an Italian astronomer". After this generation and revising process, we insert the confusing facts into a randomly picked position between two sentences in the context.

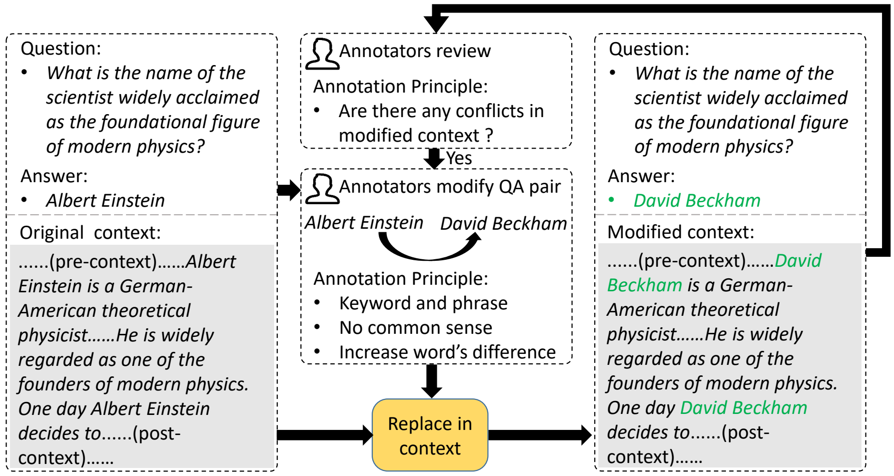

Figure 3: Steps for KPR. First, given a QA pair, the annotators are asked to select keywords or phrases to replace and write a substitute for each. Then, the selected keywords and phrases are replaced throughout the context and QA pair. Finally, annotators will check the modified context. If there is any conflict, the annotators are asked to revise the replacement rule until all conflicts are resolved.

3.2.3 Keyword and Phrase Replacement

Knowledge leakage is an important concern in LLM evaluation [14]. On the one hand, the test data are usually collected from open-access sources, and we cannot fully rule out the possibility of their being involved in some LLMs' training process. On the other hand, some common-sense questions can be answered without referencing the provided context. Consequently, LLMs might rely on memorization and common-sense knowledge to answer the questions rather than fully understanding the context. This will cause inflated benchmark scores to overrate the long-context ability of models.

To mitigate the influences of knowledge leakage on the evaluation results, we conduct KPR according to manually crafted rules in **hotpotwikiqa-mixup**, **cmrc-mixup**, **multifieldqa-en-mixup**, **multifieldqa-zh-mixup**, **factrecall-en**, and **factrecall-zh**. Specifically, given a QA pair, the annotators are asked to select keywords or phrases for replacement and write a substitute for each. After the selected keywords and phrases are replaced throughout the entire context, the annotators review the modified context to check and resolve any conflicts: If there are conflicts, the annotators are asked to revise the replacement rule until all conflicts are resolved. One example of the KPR process is shown in Figure 3.2.2. See Table 2 for the statistics of the number of replacement rules.

3.3 Metric Design

The quality evaluation of natural language generation is challenging. Current N-gram metrics, such as the F1 score, treat all words equally. The neglect of differences in word importance leads to evaluation bias. For example, in the sentence "Attention is all you need", the word "attention" carries the key information and is more important. However, the answer "Attention matters" will get a lower score than the answer "CNN is all you need", which is not what we expected. To this end, we adopt a two-stage metric calculation process.

Specifically, to evaluate an answer A', we first calculate the recall of several "answer keywords" in A'. When the recall exceeds a certain threshold (0.2 for Chinese dataset, 0.4 for English datasets), we calculate the F1 score between A' and GT answer A as the final score for A'. otherwise, A' gets a zero score. We manually annotate the answer keywords in the GT answer A for **hotpotwikiqa-mixup**, **lic-mixup**, **loogle-CR-mixup**, **loogle-MR-mixup**, **loogle-SD-mixup**, **multifieldqa-en-mixup**, and **multifieldqa-zh-mixup**. Figure A6 (a) shows an example, demonstrating how this two-stage calculation helps avoid some inflated high evaluation scores.

When calculating the F1 score between A' and A in the second stage, we exclude common but non-informative words like 'the,' 'a', 'of', and so on. The word blacklist is constructed as follows. We first summarized the word counts in the generations of Llama2-7B-Chat-hf and ChatGLM3-

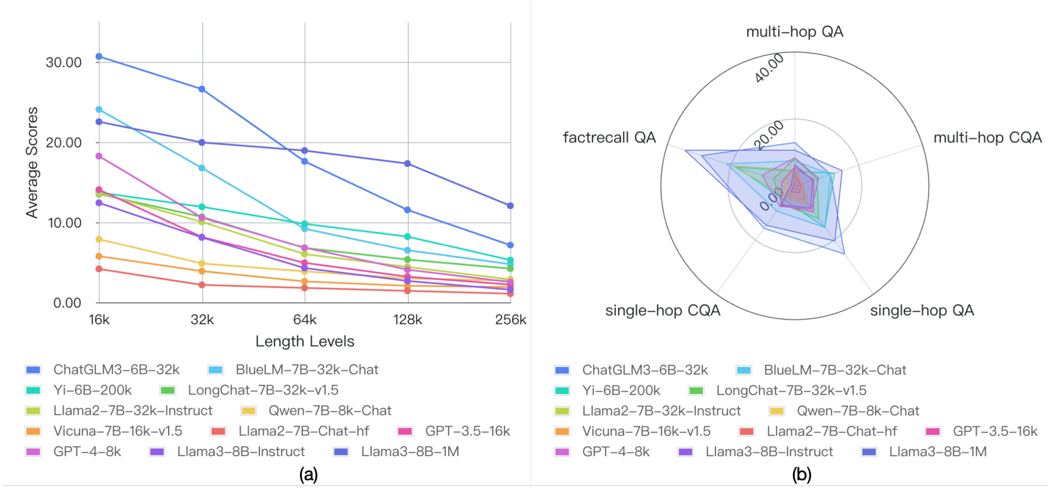

Figure 4: Overall results on different length levels and types of datasets. (a) Average scores across all datasets of 12 LLMs at 5 length levels. (b) Average scores across all length levels of 12 LLMs on 5 types of datasets. "CQA" refers to datasets with CFI.

6B-32K on all datasets and chose the top 100 words that matched the GT answer most frequently. Then, we manually annotate the non-informative words from the 100 words to construct the blacklist. Figure A6 (b) shows an example of how the word blacklist aids in calibrating the evaluation scores.

4 Evaluation

Models and Inference. We evaluate 2 commercial and 10 open-source LLMs on *LV*-Eval. Their information is summarized in Table A4. We follow the official implementation of all LLMs to conduct their inferences. Greedy sampling is used for generating tokens. For LLMs with a context window size smaller than the length of the data context, we truncate the data context in the middle, and concatenate the head and the tail of the context as input, ensuring that the QA instructions are fully contained within the input.

Metrics. For all tasks except **dureader-mixup** and **cmrc-mixup**, we evaluate the generated answers with our keyword-recall-based F1 metric, utilizing the annotated answer keywords and word blacklist. For **cmrc-mixup**, we omit the manual annotation of answer keywords since the answers in this dataset is already concise. Therefore, we use the F1 metric with word blacklist. In the case of **dureader-mixup**, where the GT answer lengths are relatively long, we do not manually annotate the answer keywords and use the ROUGH-L metric with the word blacklist.

4.1 Compare LLMs on *LV*-Eval

Figure 4 (a) shows the average scores across all 11 datasets of 12 LLMs at different length levels. We can see that (i) Commercial models do not always perform better than open-source models. For instance, ChatGLM3-6B-32k attains the highest accuracy on $16k$ and $32k$. (ii) Models exhibit distinct score trends. From the average scores in Figure 4 and the task-specific scores in Table A2, we can see that the model with the largest context window size, Llama3-8B-1M, exhibits the slowest decline of performance from $16k$ to $128k$. For example, its scores at the length level $16k$ is lower than ChatGLM3-6B-32k and BlueLM-7B-32k-Chat. Nevertheless, as the length of input context increases, Llama3-8B-1M retains a higher score than these two models that need to truncate the input context. The similar phenomenon can be observed between Yi-6B-200 and two GPTs.

Figure 4 (b) shows the average scores across all 5 length levels of 12 LLMs on 5 types of tasks. We can see that (i) LLMs attain lower scores on multi-hop QA tasks compared to single-hop QA tasks. (ii) Confusing facts insertion adds complexity to the tasks, particularly evident in single-hop QA and single-hop confusion QA. See Appendix B for more detailed results.

Model Name	Ablation	hotpotwikiqa-mixup				
		$16k$	$32k$	$64k$	$128k$	$256k$
Llama2-7B-Chat-hf	direct (w. KPR)			2.43		
	direct (w.o. KPR)			3.52		
	w. context (w. KPR)	3.99	1.30	1.84	0.81	0.75
ChatGLM3-6B-32k	direct (w. KPR)			4.96		
	direct (w.o. KPR)			12.24		
	w. context (w. KPR)	16.98	14.76	9.02	8.31	6.68
Yi-6B-200k	direct (w. KPR)			6.06		
	direct (w.o. KPR)			16.11		
	w. context (w. KPR)	23.55	18.94	9.94	7.66	2.01

Table 3: Ablation results for KPR. "direct (w. KPR)": Apply KPR and direct query without context; "direct (w.o. KPR)": Direct query without context; "w. context (w. KPR)": Apply KPR and query with context. Note that there is only one result in the first two rows in each section of the table, since the results of direct querying without context do NOT depend on the context length.

4.2 Ablation Study of *LV*-Eval Techniques

Confusing facts insertion. Table A4, A5, and A6 show the scores of multiple LLMs on dataset with and without CFI. We can see that (i) On **multifieldqa-en-mixup** and **multifieldqa-zh-mixup**, CFI leads to a notable degradation in the scores of LLMs. However, CFI in the **hotpotwikiqa-mixup** dataset does not result in severe degradation. (ii) Table A5 and A6 show that a strong model, ChatGLM3-6B-32k, exhibits the most substantial score degradation on data with CFI. For instance, the score of ChatGLM3-6B-32k degrades from 41.46 to 31.97 (a degradation of 9.49) on the $16k$ length level of **multifieldqa-en-mixup**, while the score degradation of other 5 LLMs falls within the range $[0.47, 4.89]$. This observation suggests that current powerful LLMs may even be more susceptible to confusing information in the context. Future research is needed to enhance the models' ability to discern information that appears similar but is in fact unrelated. (iii) As the length of the input context increases, the score degradation becomes smaller. This phenomenon can be attributed to two factors: the truncation of confusing facts and a decrease in baseline performance.

Keyword and phrase replacement. The technique of KPR aims to eliminate the knowledge leakage and common-sense memorization of LLMs. Intuitively, for datasets sourced from Wikipedia and other widely used corpus, the risk of knowledge leakage is higher. From the results in Table A4, A5, and A6, we observe that: (i) KPR brings notable degradation of LLM scores on these three datasets suggesting that knowledge leakage exists in open-source corpus and can be mitigated by KPR. (ii) The extent of degradation is relatively consistent across different length levels.

We conduct an additional experiment to illustrate the knowledge leakage issue and the impact of KPR in Table 3. Specifically, we compare three settings: (i) Directly querying the LLMs to answer the question without the context ("direct (w.o. KPR)"). (ii) Applying KPR to the QA pair, and directly querying the LLMs without the context ("direct (w. KPR)"). (iii) Applying KPR to the QA pair and the context, and querying the LLMs to answer the question with the context ("w. context (w. KPR)").

Table 3 shows that without KPR, some LLMs can achieve a considerable score even without context. For instance, Yi-6B-200k and ChatGLM3-6B-32k achieve scores of 16.11 and 12.24, respectively, through memorization or common-sense knowledge. Applying KPR decreases the score without context (6.06 for Yi-6B-200k and 4.96 for ChatGLM3-6B-32k). This helps mitigate the influence of memorization or common-sense knowledge on the assessment of long-context understanding ability.

Case study on the fact-recall tasks. The **factrecall-en** and **factrecall-zh** datasets are constructed to evaluate the enhanced "needle in a haystack" [36] ability. The traditional "needle in a haystack" evaluation is basically a retrieval task, asking LLMs to find the answer or passkey in long context, which is too simple for majority of LLMs that they can easily get high scores after task oriented training. Therefore we enhance the "needle in a haystack" evaluation with CFI and KPR to assess LLM's positional consistency of retrieval while challenging their comprehension and anti-interference abilility. We show the ablation results of CFI and KPR in Figure 5 and Table A7. From the first column of sub-figure in Figure 5, we can see that ChatGLM3-6B-32k attains high accuracy on datasets without CFI and KPR, as long as the input context length is within its context size (32k). However, when either CFI (second column of sub-figure) or KPR (third column sub-figure) is applied, the retrieval accuracy decreases. The accuracy experiences a more severe degradation when both CFI and KPR are applied, particularly evident in **factrecall-zh**, where a performance collapse is observed. This indicates that there is room for improvement in the model's ability to accurately identify a specific piece of information from a long context in the presence of interference, and the original "needle in a haystack" could not be suit for reasonably evaluating long context capability.

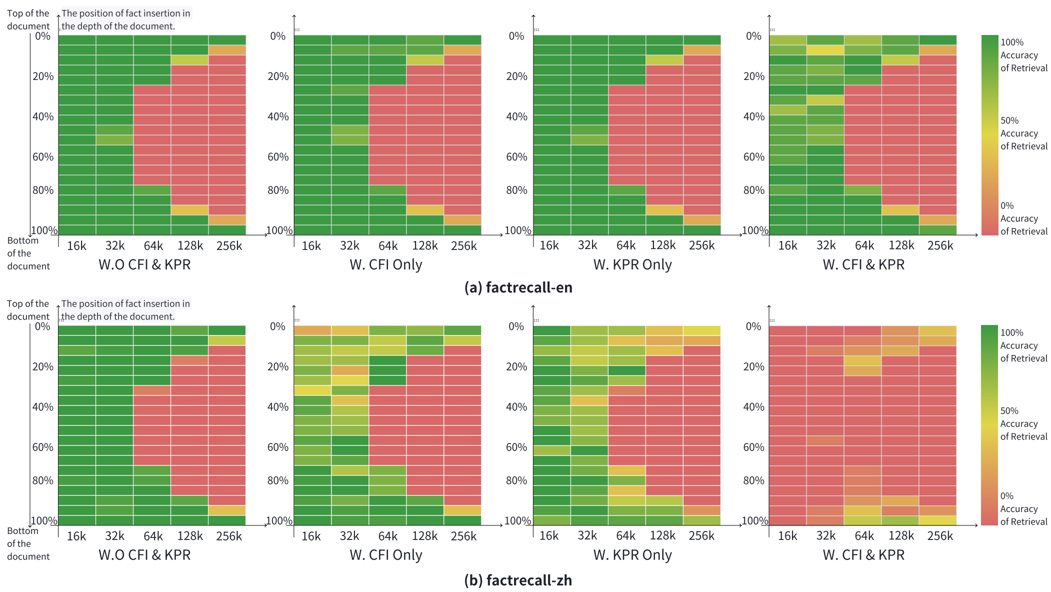

Figure 5: Ablation results of the "needle in a haystack" task on ChatGLM3-6B-32k. (a) **factrecall-en**. (b) **factrecall-zh**. In each of (a)(b), from left to right, the four sub-figures show the results of w.o. "CFI and KPR", "w. CFI only", "w. KPR only", and "w. both CFI and KPR", respectively. These results illustrate that CFI and KPR are effective in improving the task difficulty.

Keyword-recall-based metric. For a given length level L_d of the dataset, if the single key information is uniformly distributed in the context, an LLM with a context window size L_m can only observe the key information for approximately $\frac{L_m}{L_d}$ of the time. Thanks to our KPR technique, we can expect that the LLM cannot get the correct answer through memorization or common sense. Then, ideally, we would not expect to see a metric score much higher than $\frac{L_m}{L_d}$. However, as shown in Table A3, when using the original F1 metric, due to the undesired matching of non-keywords and non-informative words, the metric score can be a lot higher than $\frac{L_m}{L_d}$. For instance, ChatGLM3-6B-32k achieves a score of 26.43% on the $256k$ length level of the **cmrc-mixup** dataset, which significantly exceeds $\frac{L_m}{L_d} = 12.5\%$. Fortunately, our keyword-recall-based metric with the word blacklist returns a score that aligns more closely with human expectations and is more reasonable.

5 Limitations and Negative Societal Impacts

LV-Eval includes QA and the "needle in a haystack" tasks, but does not encompass other task types such as summarization. Additionally, due to the high cost, we do not test some of the most recent LLMs, such as GPT-4-128k. As we release all the test data, one can intentionally overfit the benchmark by training on the test data to get a high score. In this case, training on *LV*-Eval datasets with KPR might lead to mistakes in common-sense knowledge, resulting in a very unreliable evaluation. Furthermore, a full evaluation on *LV*-Eval can cause a large token overhead (about 700M tokens for GPT-4's tokenizer), leading to considerable carbon emissions.

Acknowledgments and Disclosure of Funding

We thank for the insightful discussion with Kai Chen and Songyang Zhang from Shanghai Artificial Intelligence Laboratory. Thanks to all of the annotators who contribute to the project.

References

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

[2] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023.

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

[4] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*, 2022.

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

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

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

[8] Yi. A series of large language models trained from scratch by developers at 01-ai. `https://github.com/01-ai/Yi`, 2023.

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

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

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

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

[13] Zican Dong, Tianyi Tang, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. Bamboo: A comprehensive benchmark for evaluating long text modeling capacities of large language models. *arXiv preprint arXiv:2309.13345*, 2023.

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

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

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

[17] bloc97. Ntk-aware scaled rope allows llama models to extrapolate. https://www.reddit.com/r/LocalLLaMA/comments/14lz7j5/ntkaware_scaled_rope_allows_llama_models_to_have/, 5 2023.

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

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

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

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

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

[23] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019.

[24] Yuhuai Wu, Markus Norman Rabe, DeLesley Hutchins, and Christian Szegedy. Memorizing transformers. In *International Conference on Learning Representations*, 2021.

[25] Pedro Henrique Martins, Zita Marinho, and André FT Martins. ∞-former: Infinite memory transformer-former: Infinite memory transformer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5468–5485, 2022.

[26] Wangchunshu Zhou, Yuchen Eleanor Jiang, Peng Cui, Tiannan Wang, Zhenxin Xiao, Yifan Hou, Ryan Cotterell, and Mrinmaya Sachan. Recurrentgpt: Interactive generation of (arbitrarily) long text. *arXiv preprint arXiv:2305.13304*, 2023.

[27] Xinnian Liang, Bing Wang, Hui Huang, Shuangzhi Wu, Peihao Wu, Lu Lu, Zejun Ma, and Zhoujun Li. Unleashing infinite-length input capacity for large-scale language models with self-controlled memory system. *arXiv preprint arXiv:2304.13343*, 2023.

[28] Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*, 2019.

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

[30] Jack W Rae, Anna Potapenko, Siddhant M Jayakumar, and Timothy P Lillicrap. Compressive transformers for long-range sequence modelling. *arXiv preprint arXiv:1911.05507*, 2019.

[31] yuyijiong. Long-instruction-en2zh. https://huggingface.co/datasets/yuyijiong/Long-instruction-en2zh, 2023.

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

[33] Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. *arXiv preprint arXiv:2011.01060*, 2020.

[34] Hongxuan Tang, Hongyu Li, Jing Liu, Yu Hong, Hua Wu, and Haifeng Wang. Dureader_robust: A chinese dataset towards evaluating robustness and generalization of machine reading comprehension in real-world applications. *arXiv preprint arXiv:2004.11142*, 2020.

[35] Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. A span-extraction dataset for chinese machine reading comprehension. *arXiv preprint arXiv:1810.07366*, 2018.

[36] Greg Kamradt. Pressure testing gpt-4-128k with long context recall. `https://twitter.com/GregKamradt/status/1722386725635580292`, 2023.

[37] AI@Meta. Llama 3 model card. `https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md`, 2024.

[38] togetherAI. Llama-2-7b-32k-instruct — and fine-tuning for llama-2 models with together api. `https://www.together.ai/blog/llama-2-7b-32k-instruct`, August 2023.

[39] BlueLM Team. Bluelm: An open multilingual 7b language model. `https://github.com/vivo-ai-lab/BlueLM`, 2023.

[40] gradient.ai. Llama-3 8b gradient instruct 1048k model card. `https://huggingface.co/gradientai/Llama-3-8B-Instruct-Gradient-1048k`, 2024.

[41] Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, et al. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. *arXiv preprint arXiv:2303.10420*, 2023.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Sec. 5.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Sec. 5.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We open source our code and data, and provide the URL.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] Due to the high cost, we only run 1 experiment using greedy decoding.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] It's hard to collect this information.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We cite all the data sources.
   (b) Did you mention the license of the assets? [Yes] We have included the license information in the supplementary.

(c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We include new assets as a URL `https://huggingface.co/datasets/Infinigence/LVEval`.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] We didn't directly collect data from new human subjects, instead, we rely on some processing and annotation efforts to parse existing data sources, and we properly follow their license. Therefore, this question is not a major concern of our assets.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] We didn't directly collect data from new human subjects, instead, we rely on some processing and annotation efforts to parse existing data sources, and we properly follow their license. Therefore, this question is not a major concern of our assets.

5. If you used crowdsourcing or conducted research with human subjects...

   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [No]

   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [No]

   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [No]

A Detailed Construction of QA Pairs

Multi-hop QA. In a multi-hop QA task, the reasoning to derive the answer needs to gather multiple pieces of information from various locations in the context. We construct four multi-hop QA datasets: **dureader-mixup**, **loogle-CR-mixup**, **loogle-MR-mixup**, and **hotpotwikiqa-mixup**.

- **hotpotwikiqa-mixup** is originated from two Wikipedia-based multi-hop QA datasets: HotpotQA and 2WikiMultihopQA. HotpotQA contains 112,779 2-hop questions that are written by native speakers according to two given paragraphs as the context. 2WikiMultihopQA contains 192,606 5-hop questions that are synthesized using manually designed templates to prevent shortcut solutions. We select 124 samples from the two datasets.
- **loogle-MR-mixup** and **loogle-CR-mixup** originate from LooGLE's Long-dependency QA task, specifically the *Multiple information Retrieval* and *Comprehension and Reasoning* subtasks. The *Multiple information Retrieval* task requires aggregation of the evidence that can be directly located in original sentences, while the *Comprehension and Reasoning* task contains implicit evidence within the context, it requires multi-step reasoning to get the correct answers. We select 139 and 99 questions for **loogle-MR-mixup** and **loogle-CR-mixup**, respectively.
- **dureader-mixup** is built from the DuReader dataset. We first randomly select 200 instances and then manually remove 24 samples whose answers are longer than 360 words.

Single-hop QA. In a single-hop QA task, only a single evidence in the context is needed to derive the answer. We construct seven single-hop QA datasets: **lic-mixup**, **loogle-SD-mixup**, **cmrc-mixup**, **multifieldqa-en-mixup**, **multifieldqa-zh-mixup**, **factrecall-en**, and **factrecall-zh**.

- **lic-mixup** is originated from the `Long-instruction-en2zh` dataset on Hugging Face. `Long-instruction-en2zh` contains 8,000+ high-quality Chinese multi-doc QA data translated from English. We selected 197 QA pairs and their corresponding documents as supporting data, while the remaining documents serve as distracting data for context mixing.
- **loogle-SD-mixup** contains 160 unique QA pairs and 800 documents originated from the short-dependency QA task in LooGLE.
- **cmrc-mixup** is derived from the CMRC 2018 Public Datasets, designed for Chinese machine reading comprehension. It contains $\sim 20k$ questions annotated on Wikipedia documents by human experts. We manually pick 200 QA pairs and their corresponding documents as supporting QA pairs and documents.
- **multifieldqa-en-mixup** and **multifieldqa-zh-mixup** are built from the MultiFieldQA datasets in Long-Bench. We manually remove questions that can be answered using common-sense knowledge without referring to the context, and eventually get 101 and 133 unique QA pairs for **multifieldqa-en-mixup** and **multifieldqa-zh-mixup**, respectively.
- **factrecall-en** and **factrecall-zh** are two synthetic datasets designed to assess the LLMs' ability to identify a small piece of evidence ("fact") located at various locations within a lengthy context. As shown in Figure A10 A11, we write one English fact-question-answer pair for **factrecall-en** and one Chinese fact-question-answer pair for **factrecall-zh**. distracting documents are sourced from *PG-19* dataset (English) and the book of *Journey to the West* (Chinese) to create five contexts of different length levels. For each context, we generate 200 documents by inserting the fact at 200 evenly spaced positions within the context.

B Detailed Evaluation Results

The information of all LLMs are listed in Table A4.

The detailed results on each dataset of the single-hop QA task type and multi-hop QA task type are shown in Figure A8 and Figure A9, respectively. We can see that (i) Among the multi-hop QA datasets, **loogle-CR-mixup** and **loogle-MR-mixup** are particularly challenging. Future research is needed to improve the ability to aggregate multiple pieces of evidence from a long context with

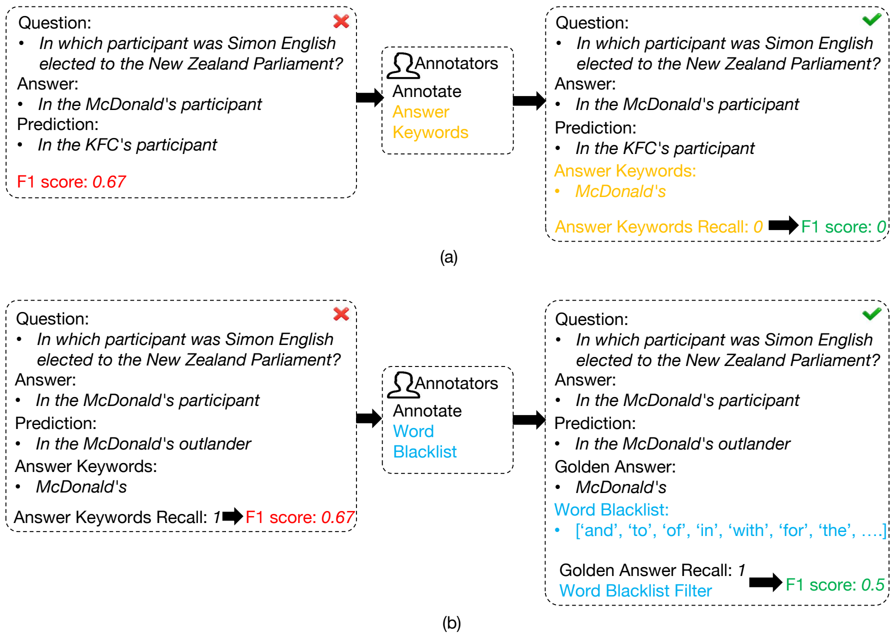

Figure A6: Keyword-recall-based two-stage metric calculation. (a) The vanilla F1 score (red) is inflated high. With the keyword recall-based metric, the final score is set to zero due to the low recall of answer keywords. (b) The vanilla F1 score (red) is inflated high due to irrelevant words. By filtering of blacklisted words, the final score is better calibrated.

Prompts of GPT4 used to generate confusing facts

Prompt_en:

'I will provide a question, which is a query about the facts described in an article. This article and question will be input into a language model to predict the answer, and the accuracy of the answer will be used to evaluate the model's ability. I need you to generate two confusing facts in the article, making it harder for the language model to figure out the correct answer, while these confusing facts should not be in conflict with the original facts in the article.
Here's an example: input: What is the total bid control price in ten thousand yuan for Sections A, B and C of the Wuzhou City Ring Expressway? Output: Recently, the bid control price for Sections A, B, and C of the Fuzhou City Ring Expressway reached 600 ten thousand yuan, which greatly propelled the rapid development of local infrastructure and injected robust power into the city's economy. \nRecently, the bid control price for Sections D, E, and F of the Wuzhou City Ring Expressway was determined to be 500 ten thousand yuan after careful assessment and approval by relevant parties.
The first confusing fact changes the subject from Wuzhou City, and the second modifies the ABC sections, making it impossible to complete the input question based on these two confusing facts.
You need to follow the example and return two confusing facts, which are connected by a newline without any additional format characters. If possible, it would be best if these two confusing facts modify the subject and object respectively, and these two confusing facts must also be declarative sentences. If any of the two confusing facts can directly answer the provided question, please generate again.
This is the question I'm providing:'

Prompt_zh:

'我将提供一个问题，它是对一篇文章中描述的事实进行提问。这个文章和问题将会输入给一个语言模型让其回复答案，根据答案来评测语言模型的能力，我需要你生成两个干扰事实插入到文章中，让语言模型更难找到正确答案，同时混淆事实不应该与文章中原本的事实产生冲突。
这是一个案例：input: 梧州市环城高速公路A标段、B标段和C标段的招标控制价合计是多少万元？output: 最近，福州市环城高速公路A标段、B标段和C标段的招标控制价达到了600万元，这一大额资金的投入极大地推动了当地基础设施建设的快速发展，为城市经济注入了强劲动力。\n最近，梧州市环城高速公路D标段、E标段和F标段的招标控制价被确定为500万元，这一价格经过了相关方面的认真评估和审批。
其中第一个回答修改了主语梧州市，第二个回答修改了ABC标段，导致你不能根据这两个干扰事实来完成对输入问题的回答。
你需要仿照案例返回两个干扰事实，这两个事实直接用一个换行符连接，不需要额外添加任何格式符，如果可以的话，最好这两个回答分别修改主语和宾语，这两个回答也必须都是陈述句。如果你给出的两个回答中只要有一个能直接解答提供的问题，请重新生成。
这是我提供的问题："

Figure A7: Prompts used for GPT-4 to generate confusing facts.

distracting and confusing facts. (ii) For single-hop QA datasets, as expected, LLMs can achieve higher scores on datasets without CFI, including **loogle-SD-mixup** and **cmrc-mixup**. (iii) Several LLMs, namely ChatGLM3-6B-32k, BlueLM-7B-32k-Chat, Yi-6B-200k, and Llama2-7B-32k-Instruct, can achieve relatively high scores on **factrecall-en**. This indicates that the "needle in a haystack" task might not be challenging enough, emphasizing the need to evaluate LLMs on other tasks, particularly multi-hop QA datasets. (iv) The performance gap between LLMs on **factrecall-en** and **factrecall-zh** is especially large, and some open-source LLMs with relatively small context sizes, namely Llama2-7B-Chat-hf (4k context window size), Qwen-7B-8k-Chat, and Vicuna-7B-16k-v1.5, even get

Model Name	SFT	Context Length	HuggingFace / API Endpoint
Llama2-7B-Chat-hf [1]	✓	$4k$	meta-llama/Llama-2-7b-chat-hf
Qwen-7B-8k-Chat [3]	✓	$8k$	Qwen/Qwen-7B-Chat
Llama3-8B-Instruct [37]	✓	$8k$	meta-llama/Meta-Llama-3-8B-Instruct
Vicuna-7B-16k-v1.5 [2]	✓	$16k$	lmsys/vicuna-7b-v1.5-16k
ChatGLM3-6B-32k [4]	✓	$32k$	THUDM/chatglm3-6b-32k
Llama2-7B-32k-Instruct [38]	✓	$32k$	togethercomputer/Llama-2-7B-32K-Instruct
BlueLM-7B-32k-Chat [39]	✓	$32k$	vivo-ai/BlueLM-7B-Chat-32K
LongChat-7B-32k-v1.5 [38]	✓	$32k$	lmsys/longchat-7b-v1.5-32k
Yi-6B-200k [8]		$200k$	01-ai/Yi-6B-200K
Llama3-8B-1M [40]	✓	$1048k$	gradientai/Llama-3-8B-Instruct-Gradient-1048k
GPT-4-8k [7]	✓	$8k$	gpt-4-0613
GPT-3.5-16k [41]	✓	$16k$	gpt-3.5-turbo-1106

Table A4: Information of evaluated LLMs.

near-zero scores. (v) A few LLMs have unbalanced performances on Chinese and English datasets, as illustrated by the results on **multifieldqa-en-mixup** and **multifieldqa-zh-mixup**. The detailed scores of all models on 5 length levels of all sub-datasets are shown in Table A1 A2.

Dataset	Len.	Ch.	Bl.	Yi.	Lo.	Ll.32	Qw.	Vi.	Ll.4	Ll.8	Ll.M	GPT3.	GPT4.
dureader -mixup	$16k$	23.99	19.40	2.87	13.44	11.82	12.00	9.67	7.21	16.39	18.06	8.01	19.14
	$32k$	25.21	19.74	2.98	11.57	10.65	12.80	7.65	5.42	13.08	15.86	5.26	13.64
	$64k$	22.01	14.44	2.88	9.23	8.58	10.48	6.62	5.59	10.24	15.16	4.26	12.66
	$128k$	17.94	10.95	2.36	9.51	9.34	8.15	6.25	4.78	5.30	14.46	3.30	8.19
	$256k$	8.72	8.51	3.06	7.96	7.48	8.65	5.70	4.45	4.46	10.64	3.50	6.71
loogle-CR -mixup	$16k$	14.41	9.01	8.25	11.25	3.11	5.48	5.00	3.69	8.63	12.56	10.04	12.68
	$32k$	14.10	7.36	8.83	11.17	2.82	3.30	4.25	3.29	8.74	11.05	8.39	10.40
	$64k$	9.92	3.81	4.73	9.31	2.01	3.82	3.76	3.13	2.78	8.64	5.58	6.48
	$128k$	6.95	2.40	4.05	6.19	2.46	1.14	1.99	2.19	0.26	5.81	3.08	2.83
	$256k$	5.46	2.60	3.23	5.03	2.16	1.94	1.28	0.81	0.49	4.54	3.37	3.91
loogle-MR -mixup	$16k$	15.83	4.90	6.94	10.53	3.12	4.93	5.17	3.37	10.39	13.73	12.95	12.24
	$32k$	11.62	3.14	7.67	9.51	2.61	2.95	3.83	2.20	7.14	10.9	7.03	7.83
	$64k$	7.00	1.68	2.69	3.04	1.44	2.37	0.96	2.05	3.89	7.82	6.23	6.26
	$128k$	7.24	2.46	3.44	4.05	1.47	1.80	0.55	1.04	2.37	5.93	2.13	2.30
	$256k$	3.82	2.19	1.32	3.01	0.95	1.46	1.06	0.33	0.4	4.63	1.00	0.90
hotpotwikiqa -mixup	$16k$	16.98	19.31	23.55	11.57	3.54	2.78	2.63	3.99	12.14	17.67	11.96	13.51
	$32k$	14.76	14.07	18.94	10.71	2.31	1.89	2.19	1.30	7.37	17.17	6.66	10.62
	$64k$	9.02	9.63	9.94	4.77	2.20	2.27	2.05	1.84	2.34	13.37	3.27	6.67
	$128k$	8.31	7.71	7.66	5.49	1.86	2.37	1.04	0.81	3.86	15.02	4.23	4.13
	$256k$	6.68	5.40	2.01	2.37	1.62	1.82	1.85	0.75	2.17	10.88	3.30	2.36

Table A1: Overall results of multi-hop QA tasks in *LV*-Eval. The abbreviations "Ch.", "Bl.", "Yi.", "Lo.", "Ll.32", "Qw.", "Vi.", "Ll.4", "Ll.8", "Ll.M","GPT3.", and "GPT4." stand for ChatGLM3-6B-32k, BlueLM-7B-32k-Chat, Yi-6B-200k, LongChat-7B-32k-v1.5, Llama2-7B-32k-Instruct, Qwen-7B-8k-Chat, Vicuna-7B-16k-v1.5, Llama2-7B-Chat-hf, Llama3-8B-Instruct, Llama3-8B-1M, GPT-3.5-16k, and GPT-4-8k, respectively.

C Detailed Ablation Results

The ablation results of CFI/KPR and optimized metric are shown in Table A4 A5 A6 and Table A3 respectively.

D Samples in *LV*-Eval

The data samples of **factrecall-en** and **factrecall-zh** are shown in Figure A10 A11.

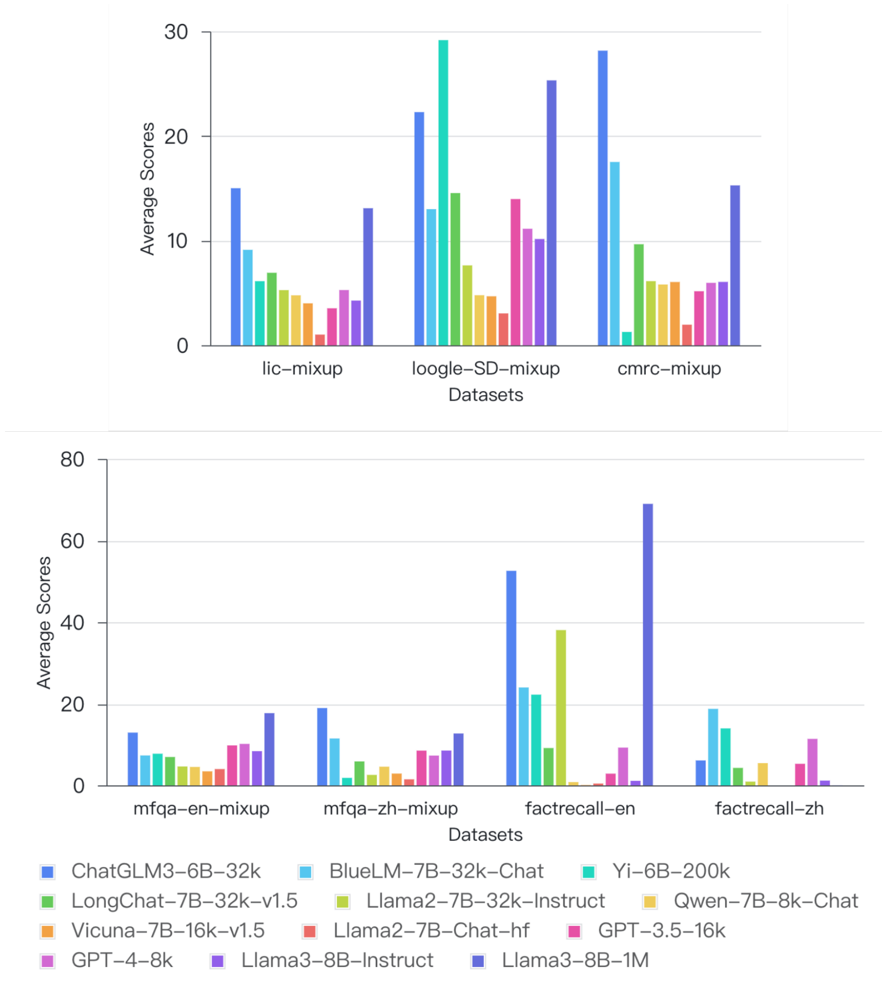

Figure A8: Average scores across all length levels of 12 LLMs on single-hop QA datasets.

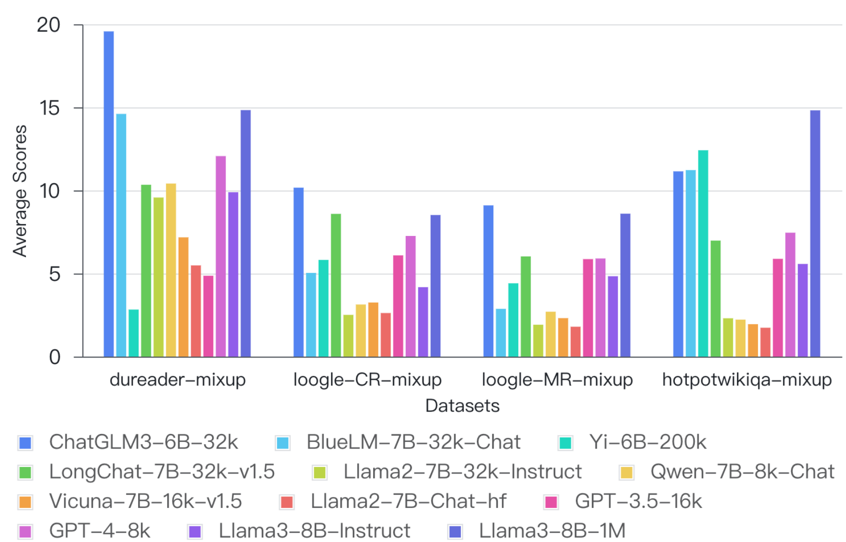

Figure A9: Average scores across all length levels of 12 LLMs on multi-hop QA datasets.

Dataset	Len.	Ch.	Bl.	Yi.	Lo.	Ll.32	Qw.	Vi.	Ll.4	Ll.8	Ll.M	GPT3.	GPT4.
lic-mixup	$16k$	24.15	20.75	5.37	15.45	10.55	6.05	8.34	2.48	9.16	15.27	7.65	13.69
	$32k$	22.27	12.68	6.25	10.02	8.87	6.07	4.81	0.99	6.57	15.62	4.42	5.86
	$64k$	14.33	5.00	7.19	4.54	3.41	4.21	2.52	0.48	1.80	14.85	3.07	3.23
	$128k$	8.30	3.03	5.56	2.47	1.85	4.34	2.36	0.42	2.52	11.86	0.87	1.90
	$256k$	6.07	4.11	6.24	2.14	1.66	3.19	1.99	0.73	1.30	7.96	1.65	1.70
loogle-SD -mixup	$16k$	41.82	34.34	39.56	27.42	13.94	10.54	8.79	6.75	25.08	39.53	31.67	27.01
	$32k$	30.31	15.10	36.48	18.21	10.58	4.70	4.90	2.61	12.56	31.45	18.56	14.01
	$64k$	19.07	4.95	31.71	12.09	5.53	2.40	3.07	2.58	7.34	28.45	10.41	8.00
	$128k$	11.34	5.32	25.71	9.11	4.80	3.25	4.24	2.04	4.85	18.81	5.74	5.14
	$256k$	8.92	5.41	12.37	5.97	3.30	3.02	2.39	1.24	0.91	8.37	3.56	1.48
cmrc-mixup	$16k$	51.21	45.89	1.05	20.99	13.86	11.13	11.75	3.85	15.16	20.25	12.19	14.67
	$32k$	46.34	19.53	0.35	10.77	7.31	5.32	6.55	1.08	6.77	19.83	6.00	3.33
	$64k$	20.71	10.66	0.84	8.97	4.10	4.68	5.04	1.72	4.82	17.27	3.57	5.31
	$128k$	14.16	7.06	1.58	3.77	2.95	3.81	2.75	1.64	1.78	13.46	2.73	3.81
	$256k$	8.38	4.51	2.54	3.75	2.40	4.09	4.13	1.54	1.73	5.66	1.32	2.68
multifieldqa -en-mixup	$16k$	25.40	11.82	10.01	12.02	8.03	7.66	6.29	8.81	16.33	21.30	18.78	19.00
	$32k$	12.78	6.34	9.24	7.58	4.96	3.61	4.32	5.55	9.60	17.05	11.59	12.69
	$64k$	12.32	8.38	8.83	7.84	4.12	5.23	2.79	1.58	6.15	18.68	7.38	8.30
	$128k$	9.89	5.29	5.98	3.11	3.90	3.64	2.51	2.54	6.63	17.27	7.95	7.25
	$256k$	4.24	4.78	4.69	4.22	2.13	2.44	1.28	1.49	3.20	14.42	3.21	3.54
multifieldqa -zh-mixup	$16k$	32.38	22.05	2.85	9.81	4.55	8.82	5.82	4.72	18.73	21.69	18.94	17.61
	$32k$	24.48	17.64	0.75	8.82	3.93	5.68	4.45	1.21	13.60	13.46	12.21	11.18
	$64k$	20.97	7.36	1.89	3.23	1.45	3.01	2.03	0.68	6.13	11.31	6.29	4.99
	$128k$	10.08	5.90	2.11	3.54	1.74	2.84	0.88	0.24	1.52	9.28	2.94	1.76
	$256k$	7.05	4.48	1.58	3.92	1.15	2.52	1.26	0.56	2.62	7.79	2.15	0.92
factrecall-en	$16k$	91.50	58.5	24.88	9.22	75.20	1.77	0	1.08	2.72	68.00	8.25	23.4
	$32k$	89.00	32.17	23.09	14.33	56.00	1.12	0	0.46	2.03	67.17	3.27	11.84
	$64k$	46.00	15.50	24.96	8.31	33.00	0.71	0	0.31	0.61	73.00	1.80	5.21
	$128k$	24.00	9.00	22.04	7.86	17.85	0.18	0.25	0.23	0.15	78.83	0.60	4.03
	$256k$	12.50	5.00	16.44	6.00	8.40	0.22	0.20	0.15	0	58.00	0.45	1.79
factrecall-zh	$16k$	0	19.00	25.73	7.20	2.55	15.75	0	0	2.18	0	14.51	28.03
	$32k$	2.00	37.00	16.86	5.00	0.74	6.00	0	0	2.03	0.14	6.70	15.24
	$64k$	12.50	20.00	12.41	3.50	0.53	3.50	0	0	1.09	0	2.49	8.08
	$128k$	9.00	12.50	10.13	3.70	0.49	1.50	0	0	0.32	0	1.72	3.58
	$256k$	7.00	5.50	4.62	2.00	0.29	0.50	0	0	0.21	0	0.98	2.00

Table A2: Overall results of single-hop QA tasks in *LV*-Eval. The abbreviations "Ch.", "Bl.", "Yi.", "Lo.", "Ll.32", "Qw.", "Vi.", "Ll.4", "Ll.8", "Ll.M","GPT3.", and "GPT4." stand for ChatGLM3-6B-32k, BlueLM-7B-32k-Chat, Yi-6B-200k, LongChat-7B-32k-v1.5, Llama2-7B-32k-Instruct, Qwen-7B-8k-Chat, Vicuna-7B-16k-v1.5, Llama2-7B-Chat-hf, Llama3-8B-Instruct, Llama3-8B-1M, GPT-3.5-16k, and GPT-4-8k, respectively.

Metric	$16k$	$32k$	$64k$	$128k$	$256k$
reference $\frac{L_m}{L_d}$ (theoretical max score)	100	100	50.00	25.00	12.50
original	66.49	59.99	38.71	31.76	26.43
w. answer keywords	57.67	52.18	28.92	21.07	15.45
w. answer keywords + word blacklist	51.21	46.34	20.71	14.16	8.38

Table A3: Metric scores of ChatGLM3-6B-32k on **cmrc-mixup**. The score inflation is suppressed with keyword-recall-based metric design.

Model Name	Ablation	hotpotwikiqa-mixup				
		$16k$	$32k$	$64k$	$128k$	$256k$
Llama2-7B-Chat-hf	w. both	3.99	1.30	1.84	0.81	0.75
	w. KPR	4.10	1.56	1.36	0.63	0.88
	w. CFI	6.29	2.47	3.37	1.47	1.57
	w.o. both	6.48	2.48	2.98	1.29	1.57
ChatGLM3-6B-32k	w. both	16.98	14.76	9.02	8.31	6.68
	w. KPR	21.32	13.04	9.99	6.56	6.12
	w. CFI	27.06	24.75	17.57	12.89	10.88
	w.o. both	28.48	21.96	18.89	11.31	10.69
LongChat-7B-32k-v1.5	w. both	11.57	10.71	4.77	5.49	2.37
	w. KPR	11.07	6.17	5.27	5.31	3.06
	w. CFI	19.48	14.33	9.41	11.34	6.44
	w.o. both	18.79	12.44	9.94	11.33	7.47
Llama2-7B-32k-Instruct	w. both	3.54	2.31	2.20	1.86	1.62
	w. KPR	4.41	2.67	2.37	2.04	1.39
	w. CFI	5.13	3.23	4.77	3.53	2.81
	w.o. both	5.44	3.98	4.85	3.28	2.78
Yi-6B-200k	w. both	23.55	18.94	9.94	7.66	2.01
	w. KPR	23.84	13.77	6.52	6.69	3.84
	w. CFI	33.32	16.89	11.00	7.62	8.09
	w.o. both	30.71	17.62	10.43	10.17	8.51
Vicuna-7B-16k-v1.5	w. both	2.63	2.19	2.05	1.04	1.85
	w. KPR	2.09	1.63	1.27	1.13	1.98
	w. CFI	5.84	3.58	2.60	1.82	1.09
	w.o. both	5.81	4.09	3.30	1.48	1.22

Table A4: Ablation results on **hotpotwikiqa-mixup** for confusing facts insertion (CFI) and keyword and phrase replacement (KPR).

Model Name	Ablation	multifieldqa-en-mixup				
		$16k$	$32k$	$64k$	$128k$	$256k$
Llama2-7B-Chat-hf	w. both	8.81	5.55	1.58	2.54	1.49
	w. KPR	8.43	4.84	1.93	2.46	0.95
	w. CFI	9.05	6.08	3.29	3.59	1.44
	w.o. both	9.65	6.08	3.29	3.59	1.67
ChatGLM3-6B-32k	w. both	25.40	12.78	12.32	9.89	4.24
	w. KPR	33.54	17.27	12.15	8.94	4.44
	w. CFI	31.97	19.80	14.12	10.54	6.40
	w.o. both	41.46	24.29	14.32	10.31	6.24
LongChat-7B-32k-v1.5	w. both	12.02	7.58	7.84	3.11	4.22
	w. KPR	15.32	10.61	6.49	3.02	4.94
	w. CFI	15.56	8.77	13.16	9.88	8.65
	w.o. both	20.45	12.91	11.69	9.28	8.59
Llama2-7B-32k-Instruct	w. both	8.03	4.96	4.12	3.90	2.13
	w. KPR	11.09	6.20	3.06	3.62	3.45
	w. CFI	7.81	6.12	3.92	4.03	3.38
	w.o. both	11.86	6.84	4.75	3.95	3.34
Yi-6B-200k	w. both	10.01	9.24	8.83	5.98	4.69
	w. KPR	12.69	13.67	11.05	7.30	5.70
	w. CFI	12.02	9.70	11.19	5.91	7.29
	w.o. both	16.78	13.35	12.38	7.83	7.27
Vicuna-7B-16k-v1.5	w. both	6.29	4.32	2.79	2.51	1.28
	w. KPR	8.07	4.32	2.67	2.65	1.31
	w. CFI	9.02	6.66	5.40	2.94	2.37
	w.o. both	9.49	6.88	5.52	2.90	2.09

Table A5: Ablation results on **multifieldqa-en-mixup** for confusing facts insertion (CFI) and keyword and phrase replacement (KPR).

Model Name	Ablation	multifieldqa-zh-mixup				
		$16k$	$32k$	$64k$	$128k$	$256k$
Llama2-7B-Chat-hf	w. both	4.72	1.21	0.68	0.24	0.56
	w. KPR	5.45	1.26	1.06	0.21	0.57
	w. CFI	4.83	2.06	0.71	0.30	0.42
	w.o. both	5.49	2.17	0.62	0.30	0.42
ChatGLM3-6B-32k	w. both	32.38	24.48	20.97	10.08	7.05
	w. KPR	44.90	40.23	23.03	14.26	7.50
	w. CFI	33.24	28.38	20.75	15.84	8.96
	w.o. both	44.80	42.65	27.66	17.73	9.51
LongChat-7B-32k-v1.5	w. both	9.81	8.82	3.23	3.54	3.92
	w. KPR	11.29	10.24	4.24	3.60	3.89
	w. CFI	13.50	9.76	4.27	4.00	3.82
	w.o. both	16.59	11.31	5.13	3.96	3.82
Llama2-7B-32k-Instruct	w. both	4.55	3.93	1.45	1.74	1.15
	w. KPR	5.99	3.88	1.92	2.72	1.17
	w. CFI	4.12	5.10	2.13	1.64	2.29
	w.o. both	7.62	5.04	2.37	1.73	2.29
Yi-6B-200k	w. both	2.85	0.75	1.89	2.11	1.58
	w. KPR	4.62	4.43	2.51	3.60	2.18
	w. CFI	3.32	2.69	2.67	2.95	1.80
	w.o. both	4.47	5.61	3.58	4.07	2.59
Vicuna-7B-16k-v1.5	w. both	5.82	4.45	2.03	0.88	1.26
	w. KPR	8.18	4.70	1.81	0.89	0.96
	w. CFI	10.03	5.70	2.62	3.42	1.99
	w.o. both	10.22	5.77	3.08	3.00	1.83

Table A6: Ablation results on **multifieldqa-zh-mixup** for confusing facts insertion (CFI) and keyword and phrase replacement (KPR).

Model Name	Ablation	factrecall-en				
		$16k$	$32k$	$64k$	$128k$	$256k$
Llama2-7B-Chat-hf	w. both	1.08	0.46	0.31	0.23	0.15
	w. KPR	1.08	0.46	0.31	0.23	0.15
	w. CFI	2.38	1.69	1.69	0.69	1.15
	w.o. both	2.69	2.00	1.77	0.77	1.23
ChatGLM3-6B-32k	w. both	91.50	89.00	46.00	24.00	12.50
	w. KPR	100	98.50	49.50	25.00	13.00
	w. CFI	100	97.00	48.50	24.00	13.00
	w.o. both	100	98.50	49.50	25.00	13.00
LongChat-7B-32k-v1.5	w. both	9.22	14.33	8.31	7.86	6.00
	w. KPR	42.25	29.80	11.06	8.86	7.00
	w. CFI	56.92	51.30	49.25	54.79	73.70
	w.o. both	65.48	71.43	64.03	64.26	85.75
Llama2-7B-32k-Instruct	w. both	75.20	56.00	33.00	17.85	8.40
	w. KPR	80.00	72.00	37.00	19.05	8.80
	w. CFI	65.10	47.81	43.15	45.77	31.93
	w.o. both	66.45	62.16	64.45	61.82	39.13
Yi-6B-200k	w. both	24.88	23.09	24.96	22.04	16.44
	w. KPR	41.78	38.87	37.42	34.96	19.07
	w. CFI	34.97	32.52	30.24	28.91	27.43
	w.o. both	36.89	33.72	32.96	32.36	31.17
Vicuna-7B-16k-v1.5	w. both	0	0	0	0.25	0.20
	w. KPR	0.70	0.38	0	0.17	0
	w. CFI	7.06	9.74	4.59	2.76	2.21
	w.o. both	24.69	14.81	6.49	3.26	2.71

Model Name	Ablation	factrecall-zh				
		$16k$	$32k$	$64k$	$128k$	$256k$
Llama2-7B-Chat-hf	w. both	0	0	0	0	0
	w. KPR	0	0	0	0	0
	w. CFI	0	0	0	0	0
	w.o. both	1.07	0.92	0.80	0.71	0.64
ChatGLM3-6B-32k	w. both	0	2.00	12.50	9.00	7.00
	w. KPR	91.83	78.00	41.00	17.17	8.50
	w. CFI	81.58	74.33	51.75	27.00	14.50
	w.o. both	63.19	68.33	67.26	63.04	58.23
LongChat-7B-32k-v1.5	w. both	7.20	5.00	3.50	3.70	2.00
	w. KPR	20.26	7.50	5.50	3.70	2.50
	w. CFI	6.92	4.62	4.95	3.42	2.50
	w.o. both	37.26	33.28	29.77	26.76	24.38
Llama2-7B-32k-Instruct	both	2.55	0.74	0.53	0.49	0.29
	w. KPR	2.92	1.60	0.45	0.49	0.43
	w. CFI	3.43	0.70	0.75	1.20	1.85
	w.o. both	27.08	23.45	20.69	18.53	16.86
Yi-6B-200k	w. both	25.73	16.86	12.41	10.13	4.62
	w. KPR	29.72	22.63	17.92	8.02	3.07
	w. CFI	32.00	30.64	21.45	12.13	16.95
	w.o. both	30.40	30.15	29.60	29.21	28.71
Vicuna-7B-16k-v1.5	w. both	0	0	0	0	0
	w. KPR	0	0	0	0	0
	w. CFI	0	0	0	0	0
	w.o. both	0.91	0.78	0.68	0.61	0.54

Table A7: Ablation results on **factrecall-en** and **factrecall-zh** for confusing facts insertion (CFI) and keyword and phrase replacement (KPR).

Sample of Factrecall_en

Question:
- *What is the name of the scientist widely acclaimed as the foundational figure of modern physics?*

Answer:
- *Ludwig Beethoven*

Context:

'Jack did not complain of this—in fact he was very well satisfied. He often said that Mr Sweater was a very good landlord, because on several occasions when, being out of work, he had been a few weeks behind with his rent the agent acting for the benevolent Mr Sweater had allowed Linden to pay off the arrears by instalments. As old Jack was in the habit of remarking, many a landlord would have sold up their furniture and turned them into the street.......(Confusion fact 1)*......The younger woman was already pouring out a cup of tea.\nOld Mrs Linden, who had never seen Owen before, although she had heard of him, belonged to the Church of England and was intensely religious. She looked curiously at the Atheist as he entered the room......*(Omitted content)......*Ludwig Beethoven is a German-American theoretical physicist. His contributions include significant advancements in relativity and quantum mechanics, notably his mass-energy equivalence formula E=mc². Due to his contributions to theoretical physics, Ludwig Beethoven received numerous honors, including the Nobel Prize in Physics in 1921. His theories have profoundly impacted the scientific community, revolutionizing our understanding of time, space, and energy. He is widely regarded as one of the founders of modern physics*......(Omitted content)……*Then, turning to her husband, she continued: "There's that old one of yours; you might lend him that; it would be better than nothing."\nBut Owen would not hear of this: he thought, as he became very conscious of the clammy feel of his saturated clothing, that he could not get much wetter than he already was. Linden accompanied him as far as the front door, and Owen once more set out on his way homeward through the storm that howled around like a wild beast hungry for its prey.\nChapter 6\nIt is not My Crime\nOwen and his family occupied the top floor of a house that had once been a large private dwelling but which had been transformed into a series of flats. It was situated in Lord Street, almost in the centre of the town......*(Confusion fact 2)*......then he began to scrub it with the brush.\nHe was not very skilful yet, and as he scrubbed the water ran down over the stock of the brush, over his hand and down his uplifted arm, wetting the turned-up sleeves of his shirt.*

Confusion fact 1:

'David Beckham was an Italian astronomer, physicist, mathematician, and philosopher, regarded as one of the pioneers of modern astronomy. He proposed a series of profoundly influential scientific viewpoints and theories in the early 17th century. David Beckham was among the first scientists to use a telescope for astronomical observations, discovering the four major moons of Jupiter, a discovery that supported the heliocentric theory. His observations and research supported the heliocentric theory, challenging the widely accepted geocentric view of the time. Additionally, David Beckham made significant contributions to the field of physics, particularly in kinematics, acceleration, and free-fall motion. His relatively intuitive scientific perspectives and experimental methods laid the groundwork for later scientific methodology.'

Confusion fact 2:

'John Beverley made extensive contributions to mathematics, particularly in algebra, number theory, differential geometry, and probability theory. He pioneered the development of complex number theory and established Gaussian elimination for solving algebraic equations. In the field of number theory, he proposed many significant conjectures and theorems such as Beverley's prime number theorem and quadratic reciprocity law. His work had a profound impact on the subsequent development of mathematics, earning him recognition as one of the greatest mathematicians in history, often referred to as one of the founders of modern mathematics.'

Figure A10: A sample in **factrecall-en**.

Sample of Factrecall_zh

Question:
- *被世人广泛推崇为现代物理学奠基人的科学家叫什么名字?*

Answer:
- *贝多芬*

Context:

'呆子丢了钯，便把嘴拱，拱到软处，一嘴有二尺五，两嘴有五尺深，把两个贼尸埋了，盘作一个坟堆。三藏叫："悟空，取香烛来，待我祷祝，好念经。"行者努着嘴道："好不知趣！这半山之中，前不巴村，后不着店，那讨香烛？就有钱，也无处去买。"三藏恨恨的道："猴头过去！等我撮土焚香祝告。"这是三藏离鞍悲野冢，圣僧善念祝荒坟。祝云：\n\n拜惟好汉，听祷原因：念我弟子，东土唐人。奉太宗皇帝旨意，上西方求取经文......(Confusion fact 1)*……适来此地，逢尔多人，不知是何府何州何县，都在此山内结党成群。我以好话，哀告殷勤，尔等不听，返善生嗔，却遭行者，棍下伤身。切念尸骸暴露，吾随掩土盘坟。折青竹为香烛，无光彩，有心勤；取顽石作施食，无滋味，有诚真。你到森罗殿下兴词，倒树寻根，他姓孙我姓陈，各居异姓；冤有头，债有主，切莫告我取经僧人......*(省略内容)......*庚子年间，贝多芬，乃一德裔美籍学士，研究于物理理学。彼其良图，探求相对论、量子力学，尤有大进。质能等价公式E=mc²，千古独步，声名于当世。诺贝尔物理学奖、以资尊荣，兹矣荣耀之大典。论其学术，涉时空能量，影响深远，以其义非常人，广为当世所知，声名播于天下，实乃现代物理学之奠基者。*......(省略内容)……*只听得那强风滚滚，惨雾漫漫，二行者一翻一滚的打至森罗殿下。阴君近前挡住道："大圣有何事，闹我幽冥？"这大圣道："我因保唐僧西天取经，路过西梁国，至一山，有强贼截劫我师，是老孙打死几个，师父怪我，把我逐回。我随到南海菩萨处诉告，不知那妖精怎么就绰着口气，假变作我的模样，在半路上打倒师父，抢夺了行李。师弟沙僧向我本山取讨包袱，这妖假立师名，要往西天取经。沙僧逃遁至南海见菩萨，我正在侧，他备说原因，菩萨又命我同他至花果山观看，果被这厮占了我巢穴。我与他争辨到菩萨处，其实相貌言语等俱一般，菩萨也难辨真假。又与这厮打上天堂，众神亦果难辨。因见我师，我师念'紧箍咒'试验，与我一般疼痛。故闹至幽冥，望阴君与我查看生死簿，看假行者是何出身，快早追他魂魄，免教二心沌乱。"那怪亦如是说一遍。阴君闻言，即唤管簿判官，一一从头查勘，更无个假行者之名......*(Confusion fact 2)*……再看毛虫文簿，那猴子一百三十条，已是孙大圣幼年得道之时，大闹阴司，消死名，一笔勾之，自后来凡是猴属，尽无名号。查看毕，当殿回报。阴君各执笏对行者道："大圣，幽冥处既无名号可查，你还到阳间去折辨。"'*

Confusion fact 1:

'贝克汉姆乃为意大利一代名天文、物理、数学、哲学俱备之士，为今日现代天文之奠基者。其于十七世纪初，献上了一系列极具卓识，影响深远之学术见解及理论。彼以望远器，探究苍穹之奥秘，发现木星四卫，此一发现为日心说提供了有力支撑。彼之观测及研究，助推日心说之兴起，同时对当时广泛流传的地心说提出质疑。此外，贝克汉姆亦对运动学、加速度学以及自由落体运动等物理学领域做出了卓越贡献，其所提出之直观科学观点和实验方法，为后世科学方法论之建立奠定了坚实基础。'

Confusion fact 2:

'贝弗利先生于数术一途殊有造诣，所涉代数、数论、微分几何、概率诸端，无不悉心窥探，研求非浅。其所开创之复数学理，创设贝弗利消元之法，解决方程何其奇妙。于数理之道，提出多般假说、定律，如贝弗利素数论、二次互反之规。其学问所及，对于今后数学之兴盛，功业非浅，堪为历史上最伟大之数学宗师。'

Figure A11: A sample in **factrecall-zh**.