LiveLongBench: Tackling Long-Context Understanding for Spoken Texts from Live Streams

Anonymous ACL submission

Abstract

Long-context understanding poses significant challenges in natural language processing, particularly for real-world dialogues characterized by speech-based elements, high redundancy, and uneven information density. Although large language models (LLMs) achieve impressive results on existing benchmarks, these datasets fail to reflect the complexities of such texts, limiting their applicability to practical scenarios. To bridge this gap, we construct the first spoken long-text dataset, derived from live streams, designed to reflect the redundancy-rich and conversational nature of real-world scenarios. We design tasks across three main categories—retrieval-dependent, reasoning-dependent, and hybrid—and evaluate both popular LLMs and specialized methods for their ability to understand long-contexts in these tasks. Our results reveal that current methods struggle to effectively process highly redundant texts, with clear preferences for specific task types but no single method excelling across all tasks. Based on our findings, we propose a simple yet strong baseline that addresses these challenges, achieving substantial improvements in performance. Our analysis offers valuable insights into the strengths and limitations of existing methods for processing spoken texts, laying the groundwork for advancing long-text understanding in real-world applications. As the first benchmark specifically designed for spoken long-text understanding, it not only tackles key challenges in this domain but also serves as a valuable resource for driving innovation in e-commerce applications.

1 Introduction

Spoken texts, prevalent in scenarios such as dialogues and live streams, are becoming increasingly common as conversational AI and real-time communication continue to expand. Existing studies have demonstrated that spoken text exhibits unique linguistic properties (Eisenstein, 2013), particularly *high redundancy* characterized by repetitive phrases and filler words. This redundancy imposes significant computational challenges, including increased processing overhead and difficulties in semantic understanding. While advanced LLMs support long context lengths (Touvron et al., 2023) and current KV cache compression methods (Liu et al., 2024b; Jiang et al., 2024; Pan et al., 2024) have been designed for written texts, their ability to handle the unique redundancy patterns of spoken texts remains unexplored. This gap underscores the need for specialized approaches tailored to the characteristics of spoken language.

Generally, long contexts pose challenges for both understanding and computation. LLMs often struggle with lengthy texts, such as the *lost in the middle* phenomenon (Liu et al., 2024a). However, existing benchmarks (Bai et al., 2023; Zhang et al., 2024a) for long-context understanding predominantly focus on written texts, neglecting the informal characteristics of spoken language. Compressing the KV cache can alleviate some of the computational burden on LLMs when handling long contexts, as many filler words (*e.g.,*, "um", "uh") contribute unnecessary redundancy. This requires higher compression rates and more efficient context compression. Therefore, we raise two questions:

Question (1): Can base models effectively process long spoken texts with informal language characteristics?

Question (2): Can existing methods achieve higher compression rates, for example, through the combination of multiple techniques?

To showcase the effectiveness of current foundation models and context compression methods to long-form spoken texts, we construct a new benchmark, **LiveLongBench**, as summarized in Table 1. We construct a novel dataset recorded from live streams, featuring both Chinese and English, with sequences averaging approximately $\tilde{9}$7K

Dataset	Response Type		Multi Span		Language Style		Avg. Words
	Closed	Open	Explicit	Semantic	Spoken Texts	Languages	
LongBench	✓	✓				En.&Zh.	~13k
∞Bench	✓				✓	En.	~300k
Loong		✓		✓		En.&Zh.	~110k
Marathon	✓					En.	~163k
L-Eval	✓	✓		✓	✓	En.&Zh.	3k - 62k
M4LE	✓	✓	✓	✓		En.&Zh.	~4k
TCELongBench	✓	✓		✓		En.	~18k
FinTextQA		✓	✓	✓		En.	~19k
LiveLongBench	✓	✓	✓	✓	✓	En.&Zh.	~97k

Table 1: Comparison of Different Long-context Benchmark Datasets.

tokens. To tackle the first question, we follow the study (Wang et al., 2024a) covering three perspectives: *retrieval*, *reasoning*, and *hybrid tasks*, and design a total of nine distinct tasks. For each category, we synthesize multiple task types to evaluate various model capabilities, including both open-domain and closed-domain tasks to measure knowledge recall and generalization. Additionally, to test the model's ability to understand information across different context lengths, we incorporate tasks with both explicit spans, which require literal matching, and semantic spans, which emphasize inferential comprehension. Together, these design choices ensure that LiveLongBench provides a comprehensive evaluation of long-context understanding, particularly in the challenging domain of spoken language.

To address the second question, we first evaluate individual KV cache compression methods, including KIVI (Liu et al., 2024b), MInference (Jiang et al., 2024), and Lingua (Pan et al., 2024). Interestingly, we discover that certain method combinations achieve better performance with lower memory consumption compared to single ones. For example, using *Minference+Lingua-4x* outperforms any single method, while using *KIVI-4bit+Minference+Lingua-2x* achieves the lowest memory usage and still surpasses individual approaches such as KIVI or M-Inference. To further balance memory efficiency and performance, we adopt a Data Envelopment Analysis (DEA) framework to rank all method combinations based on their cost-effectiveness. This analysis produces a practical strategy list, providing a convenient reference for selecting optimal compression combinations according to different performance-memory trade-offs.

Our contributions are summarized as follows:

- We construct and release LiveLongBench, the first bilingual benchmark derived from live-streaming spoken texts, designed to evaluate long-context understanding and reasoning, with sequences averaging approximately ∼97K tokens.
- We systematically evaluate current LLMs, uncovering significant performance degradation when processing lengthy spoken contexts and highlighting the unique challenges posed by informal language patterns.
- We propose a hybrid KV cache compression strategy, which combines multiple compression methods and achieves superior performance-memory trade-offs, as identified through a comprehensive DEA-based efficiency analysis.
- Our experimental results and analyses provide new insights into long-context compression and offer practical guidance for enhancing LLM performance in real-world spoken-language applications.

2 Related Work

This section reviews related work in three primary areas: benchmarks for long-text understanding, conversational and spoken text processing, and techniques for handling redundancy in NLP tasks.

Long-context Understanding Benchmarks. Numerous benchmarks have been developed to evaluate long-text understanding, predominantly focusing on formal, written texts. Datasets such as (Zhang et al., 2024b; Wang et al., 2024b) emphasize structured, coherent, and information-dense content, while tasks like document summarization, information retrieval, and long-form question answering have been extensively studied using datasets such as NarrativeQA (Kočiskỳ et al., 2018), MultiNews (Fabbri et al., 2019), and SQuAD 2.0 (Sulem et al., 2021). Although these benchmarks have driven progress in long-text understanding, their reliance on formal language

overlooks the challenges posed by spoken language—characterized by disfluencies, redundancy, and variability—which leads to models that often struggle with real-world applications such as live-stream transcripts and conversational logs.

Conversational and Spoken Text Processing. Research in conversational text processing—especially within dialogue systems and ASR—has produced benchmarks such as DailyDialog (Li et al., 2017), PersonaChat (Zhang, 2018), and DSTC that typically feature short, goal-oriented dialogues with minimal noise and redundancy, failing to capture the full complexity of spontaneous speech. In contrast, corpora like Switchboard (Godfrey et al., 1992) and CallHome reflect the irregular, fragmented nature of natural spoken language, albeit in limited domains like telephony. Emerging sources from live commerce and streaming platforms offer a more diverse range of spoken data, yet systematic collection and analysis remain sparse. Recent efforts in video summarization (e.g., TVSum) and e-commerce dialogue datasets highlight the need for specialized approaches, as comprehensive solutions for long-text understanding in spoken contexts are still lacking.

Spoken Long-Text Benchmarks: Gaps and Advances Existing long-text benchmarks primarily target formal written language, overlooking the redundancy, informality, and variability of spoken texts and rarely evaluating methods for redundancy reduction or long-context processing on authentic spoken data. To address this gap, we introduce LiveLongBench—the first benchmark explicitly designed for long-text understanding in spoken contexts, focusing on live streams and dialogues. As shown in Table 1, while LongBench (Bai et al., 2023) offers rich content with key evidence often confined to specific paragraphs, and benchmarks such as ∞Bench (Zhang et al., 2024a), Marathon, and Loong (Wang et al., 2024b) provide ultra-long contexts with limited question diversity, and L-Eval (An et al., 2023) and M4LE (Kwan et al., 2023) feature varied question types over shorter contexts, and domain-specific benchmarks like TCELongBench (Zhang et al., 2024b) and FinTextQA (Chen et al., 2024) target the news and finance domains, LiveLongBench preserves extensive context, offers a broader range of question types, and incorporates spoken linguistic characteristics, making it more representative of real-world spoken language.

3 LiveLongBench

3.1 Basic Challenges

We aim to construct a dataset that captures the real-world challenges associated with long-context processing, particularly the issues of *informal language* and *high redundancy*. Live streams, with its spontaneous conversational style and repetitive content, exemplifies these challenges through numerous real-world instances, making it an ideal domain for studying long spoken text understanding.

Informal Language. Live-streaming e-commerce data often involves conversational speech, contributing to the informality of the language. Unlike formal text, live-stream content typically consists of short, fragmented utterances, leading to a high occurrence of syntactic reduction. Additionally, interactive conversations with viewers frequently introduce topic drift, where discussions shift abruptly, making it difficult for models to maintain contextual coherence. These characteristics significantly increase the complexity of document understanding compared to well-structured formal text.

Examples of the informal languages

▷ **Syntactic Reduction:**
"Big scarf The discount office."
Verbless
"This place, the focus of our vision."
Right-Dislocation

▷ **Topic Drift:**
"This handbag is made of genuine leather and comes in three colors. I bought one for my sister last week... Oh, by the way, did you see the movie I talked about yesterday?" From product details to unrelated personal topics

High Redundancy. Live-stream transcripts contain a substantial amount of filler words. To emphasize key product features, presenters often include repetitive content ,reiterating the same information multiple times. Furthermore, interactive dialogues introduce additional non-informative tokens, which inflate the overall length while lowering the density of useful information. This high redundancy poses challenges for long-context processing, requiring models to efficiently filter out noise while retaining essential details.

These inherent challenges highlight the need for

Examples of the redundant content

▷ **Filler Words:**
"Um, okay, so, yeah, you know, like, I mean, actually, basically..."

▷ **Repetitive Content:**
"This bag is beautiful, really beautiful, so beautiful! I mean, it's just beautiful!"

▷ **Non-informative Tokens:**
"This is really nice, you know? It's just so good. Like, really good, you know what I mean?"

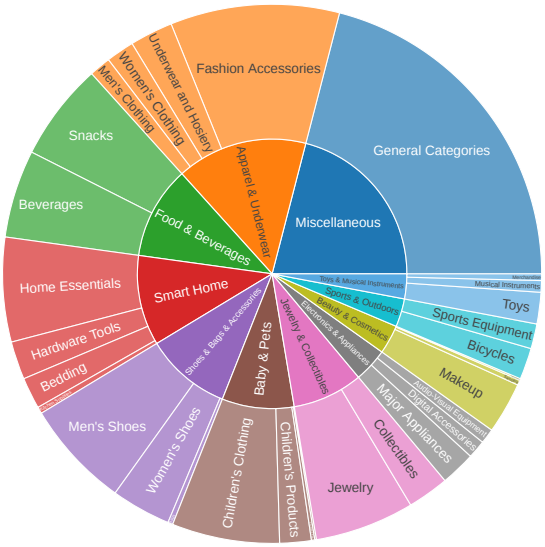


Figure 1: Distribution of Data Categories Across E-Commerce Domains

a benchmark that not only captures the complexity of long-context processing but also reflects the nuances of informal language and high redundancy features.

3.2 Dataset Collection

To tackle the challenges of long-context compression in spoken language, we introduce LiveLongBench, a novel dataset that captures the informal, repetitive, and dynamic nature of e-commerce live-stream discourse.

Data Source. The dataset is built from **Douyin** e-commerce live streams, known for their diverse and dynamic Live streaming styles. We collected and transcribed audio from live sessions spanning 11 major product categories and 32 subcategories, including apparel, electronics, beauty, and household goods. Figure 1 shows the distribution of product categories included in the dataset.

This dataset captures the spontaneous and repetitive nature of spoken language in live-stream settings, making it highly representative of real-world discourse. Each document mainly contains continuous host monologues, which characterized by informal expressions, repetitive promotional content, and frequent Q&A exchanges. This diversity ensures that this dataset accurately mirrors the linguistic challenges of real-world spoken language, providing a valuable benchmark for developing compression methods tailored to informal and redundant nature of spoken text.

Processing and Structuring. Our raw video data is sourced exclusively from publicly accessible Douyin live streams. To ensure complete preservation of all information, we utilize a pre-trained Whisper speech-to-text model [1] to transcribe the audio, retaining all repetitions and filler words so as to maintain the authenticity of the spoken context. Based on the transcribed text, the most critical step is to remove all sensitive information, such as personal identifiers, during the data processing phase, ensuring that the dataset is solely for research purposes. Subsequently, we apply a filtering process to remove non-informative noise, such as consecutive repetition of the same sentence more than ten times, thereby ensuring that the dataset remains faithful to the informal and redundant nature of live-stream discourse while enhancing its quality for subsequent analyses.

3.3 Task Construction

We define three primary task categories that align with the inherent characteristics of spoken language (Figure 2): 1) *retrieval-dependent* tasks, which challenge models to extract specific information from lengthy and often redundant spoken content, 2) *reasoning-dependent* tasks, which require models to navigate informal expressions, filler words, and fragmented structures to perform complex logical inference, and 3) *hybrid tasks*, which combine both retrieval and reasoning, reflecting real-world spoken scenarios where models must identify relevant details while simultaneously reasoning over loosely structured discourse.

Retrieval-Dependent Tasks. Retrieval in this context refers to a model's ability to locate specific information from spoken content, including identifying product policy (task `Policy`), and product specifications in the single document (task `Single`).

[1] Whisper pre-trained on a large-scale audio corpus.

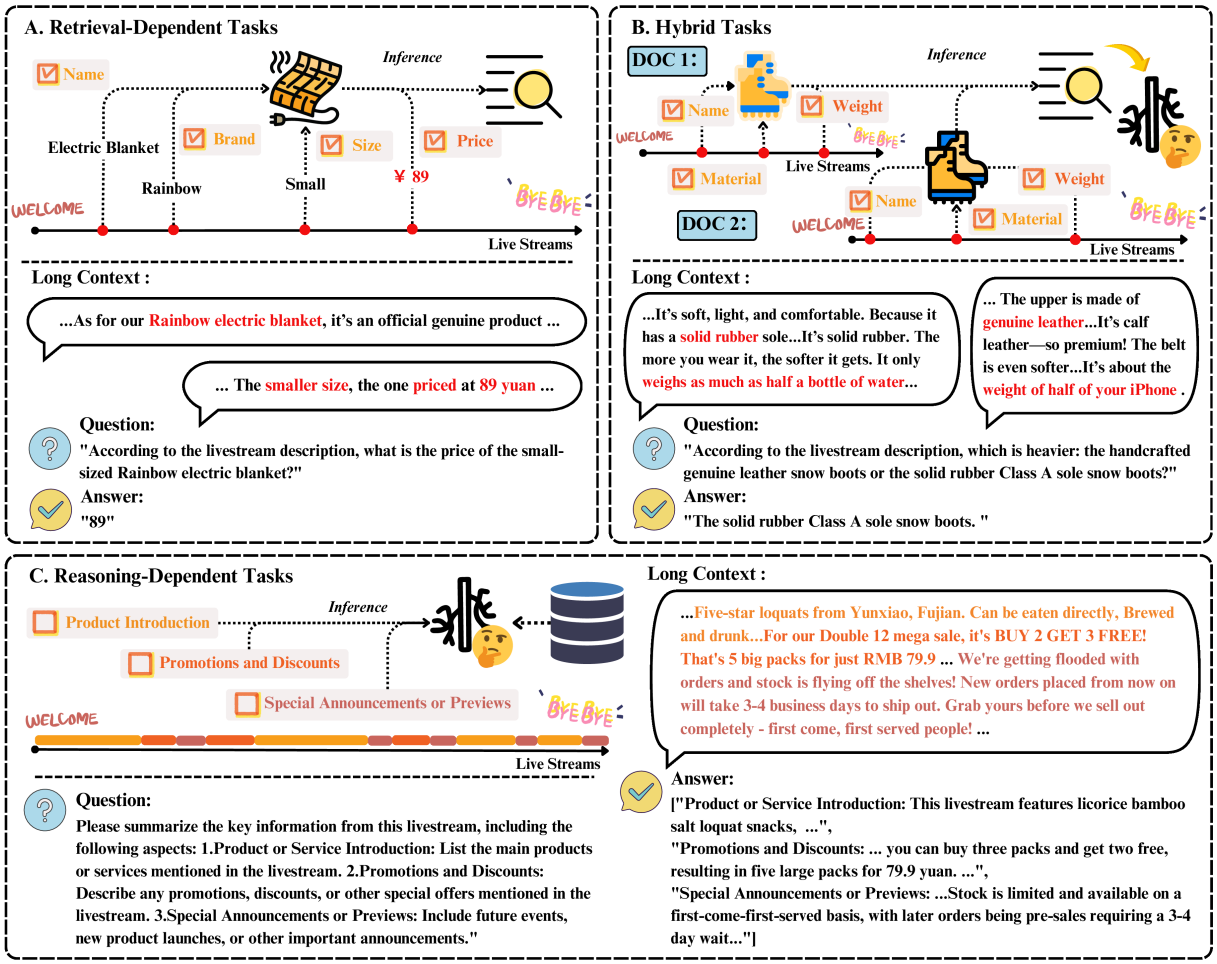


Figure 2: Showcase of Three Evaluation Tasks in LiveLongBench

For instance, a model may be asked to find the listed price of a product mentioned during a live-stream session or verify the product's specifications based on the host's descriptions.

Reasoning-Dependent Tasks. Reasoning refers to a model's ability to infer information not explicitly mentioned in the spoken content by leveraging internal knowledge within the LLMs. This includes classifying a product into the correct category (task `Class`), which often requires external knowledge about product types and market conventions, or summarizing key points from lengthy and informal conversations (task `Summary`), where the model must identify and synthesize essential information despite the presence of redundant and irrelevant content. For example, a model may need to categorize a niche electronic gadget that is not explicitly labeled during the live stream or produce a concise summary of a promotional session.

Hybrid Tasks. Hybrid tasks combine both retrieval and reasoning, requiring models to first extract multiple relevant pieces of information from spoken content and then synthesize them through reasoning to form a coherent response. This includes answering questions that span multiple segments of a live-stream transcript (task `Multiple Document QA`), where the model must retrieve dispersed details and integrate them to provide accurate answers, and comparing product prices (task `Price`), which involves locating price points mentioned at different times and reasoning about differences or promotions. For example, a model might need to compare two smartphones' battery lives and prices when the relevant information is scattered across various moments of the live streams.

4 Experiments

Next, we present the evaluation results of LiveLongBench from large language models and context compression methods separately.

	Retrieval			Hybrid			Reasoning			Overall
	Single	Policy	Avg.	Multi	Price	Avg.	Class	Sum.	Avg.	
Score										
Human	91.5	100.0	92.1	81.8	55.4	74.9	41.0	65.8	50.0	76.3
GLM4plus	**25.1**	75.0	**28.5**	34.1	16.5	29.5	36.2	**92.1**	**56.5**	**35.4**
Qwen2-7B	17.1	20.0	17.3	42.0	16.7	35.4	35.7	78.1	51.1	31.5
Mistral-7B	9.0	**80.0**	13.8	33.9	13.5	28.6	33.8	52.1	40.5	25.2
LLaMA-8B	19.2	74.6	23.0	30.9	**33.2**	31.5	**39.7**	64.1	48.6	31.9
eCeLLM	11.5	75.0	15.8	**48.4**	16.9	**40.2**	21.4	20.0	20.9	25.6
Exact Match (%)										
Human	89.1	100.0	89.8	51.4	15.4	42.0	4.8	8.3	6.1	53.5
GLM4plus	18.2	75.0	22.0	21.6	7.7	18.0	0.0	50.0	18.2	19.7
Qwen2-7B	10.9	0.0	10.2	16.2	6.7	13.7	0.0	30.8	11.2	11.7
Mistral-7B	3.6	75.0	8.5	16.2	0.0	12.0	0.0	0.0	0.0	7.8
LLaMA-8B	12.8	72.7	16.8	6.7	10.0	7.5	11.5	6.9	9.9	11.9
eCeLLM	5.5	75.0	10.2	35.1	7.7	28.0	0.0	0.0	0.0	14.1

Table 2: Performance of large language models including close sourced (GLM4) and popular open sourced (Qwen, LLaMA and Mistral models).

4.1 Large Language Models

Experimental Setup. We investigate whether foundation models can handle long and spoken queries using both closed-source models (GLM4plus) and open-source models (Qwen, LLaMA, and Mistral). Our experimental setup ensures that each model is evaluated under the same conditions, *e.g.,* max sequence length. To investigate the impacts of domain-specific fine-tuning, we also include eCeLLM (Peng et al., 2024), a model fine-tuned for e-commerce, alongside general-purpose LLMs.

For evaluation metrics, we use *Exact Match (EM)* (%) to measure the accuracy of model outputs against ground-truth answers, and a *Score* metric to provide a softer, more continuous measure of performance across different tasks.

Comparison of Foundation Models. As summarized in Table 2, closed-source models generally outperform open-source ones. It is notable that GLM4plus achieves the highest overall score (35.4). Among open-source models, LLaMA attains 31.9, closely approaching GLM4plus's performance. Notably, task-specific variations are observed: Qwen2-7B excels in reasoning tasks, LLaMA demonstrates strong retrieval performance, and eCeLLM performs well in integrated tasks.

Impacts of Domain-specific Fine-tuning. Models pre-trained or fine-tuned on specialized domains (*e.g.,* finance, e-commerce) often exhibit deeper knowledge in those areas, which can enhance reasoning or mitigate redundancy in domain-specific tasks. Notably, eCeLLM demonstrates superior performance in integrated tasks (40.2 in score and 28.0% exact match), likely due to its enhanced domain understanding. However, this specialization compromises its reasoning ability, resulting in the lowest reasoning score (20.9) and 0.0% exact match among all evaluated models.

4.2 Context Compression Methods

LLMs show varying capabilities in long-context scenarios but often face challenges due to memory usage and computational overhead. To address these limitations, we evaluate existing context compression methods and introduce a simple yet effective baseline for improving their performance.

Experimental Setup. We evaluate representative context compression methods on LiveLongBench to assess their utility for long-context understanding and their performance across retrieval, reasoning, and hybrid tasks. The evaluated methods fall into three categories:

- *Token pruning*, which directly removes tokens deemed less relevant, exemplified by LLMLingua (Pan et al., 2024).
- *Attention sparsification*, which reduces computational complexity by applying sparse attention mechanisms, represented by MInference (Jiang et al., 2024).

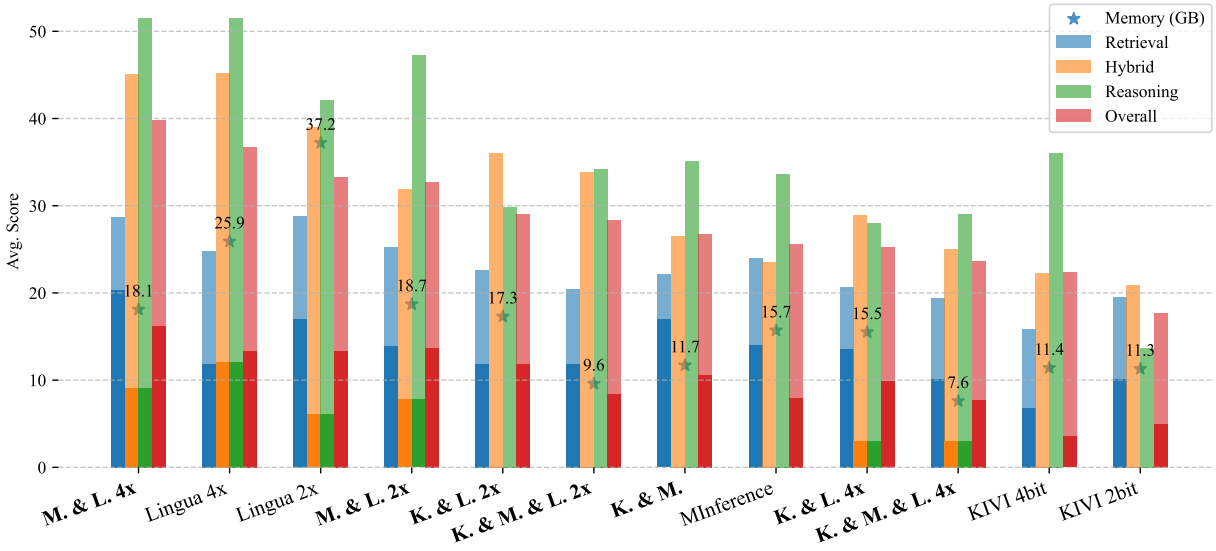


Figure 3: Performance of Context Compression Methods on LLaMA-3.1-8B-Instruct. "K." denotes KIVI, "M." denotes MInference, and "L." denotes LLMLingua, while "2x" and "4x" refer to compression ratios. Methods shown in bold along the x-axis represent multi-methods. From left to right, these methods are arranged in descending order of their *Overall* average score. To visually convey each method's exact match rate (%) on different tasks, the darker segment of each bar is computed by "Avg. Score × Exact Match". Details are shown at Table 4 in the Appendix.

Answer the Question (1):
While closed-source models remain the strongest, there is a clear gap compared to humans, with retrieval tasks being the most challenging for current models when processing spoken texts.

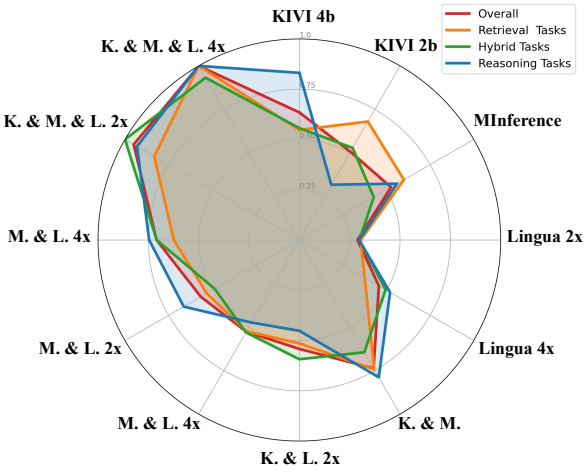


Figure 4: Efficiency Scores Based on DEA Analysis

• *Quantization*, which compresses internal key-value caches into lower-precision formats, as implemented by KIVI (Liu et al., 2024b).

Additionally, we report the performance and resource usage of each model when applying compression methods, ensuring a comprehensive assessment of both accuracy and efficiency. The results are summarized in Table 4.

Single-Method Analysis. Our analysis reveals that different compression methods exhibit distinct preferences across tasks. 1) *Low-bit quantization, by preserving all information, performs better in retrieval tasks, where retaining comprehensive details is critical.* For example, KIVI, even under ultra-low-bit quantization, achieves the highest retrieval accuracy of 80% in the `policy` task while maintaining the lowest memory usage. However, its performance declines in other tasks, likely due to excessive compression leading to information loss. The advantage of KIVI in retrieval is further validated by our experiments on the "Needle in the Haystack" task (See Section B), underscoring the critical role of information retention in achieving accurate retrieval. 2) *In contrast, sparsification and token pruning methods, which discard portions of the input, struggle with retrieval due to incomplete information but demonstrate superior reasoning performance.* For instance, LLMLingua, with a 4x compression rate, significantly outperforms other single methods in reasoning tasks. This improvement is likely due to the removal of redundant content, which serves as a form of noise reduction, enabling models to focus on essential semantic information. An empirical case study

Examples of Denoising Effects of Lingua4x

▷ **Oringinal Text:**
"Let me show you this pair of gloves..."
<long noisy text>
"...rabbit wool thermal gloves, just 9.9 yuan per pair! Item No. 1, available for two days. 9.9 yuan per pair, 9.9 yuan per pair!... "
<long noisy text>
▷ **Compressed Text by Lingua4x:**
"...The Rabbit wool thermal gloves, just 9.9 yuan per pair! 9.9 yuan per pair!..."

▷ **Question:**
"What is the price of the rabbit wool thermal gloves?"
▷ **w/o Lingua4x Answer:**
"8.8 yuan"
▷ **w Lingua4x Answer:**
"The price of the rabbit wool thermal gloves is 9.9 yuan per pair."

demonstrates how Lingua4x, by eliminating redundancy, enhances the clarity of key information (*i.e.,* price).

Notably, while compression methods are typically used to reduce computational costs in formal text, our findings reveal that in high-redundancy contexts, they also offer significant denoising effects, improving both model accuracy and overall performance.

Multi-Methods Analysis. Our analysis highlights that combining different compression strategies can achieve extreme sparsity without compromising performance. As shown in Table 4, the **MInference+Lingua4x** combination achieves the highest overall performance by balancing retrieval accuracy and reasoning capabilities. Its strength likely comes from efficient memory utilization and selective token retention. In comparison, **MInference+Lingua2x** excels in reasoning tasks, particularly logical inference, due to its prioritization of critical tokens and attention heads, though with slightly lower retrieval scores. Integrating KIVI with Lingua and MInference maintains competitive retrieval performance but shows weaker reasoning abilities, possibly due to excessive compression affecting long-range coherence.

Optimal Combination of Balancing Performance and Memory. To better understand the trade-offs between performance and memory efficiency, we apply *Data Envelopment Analysis (DEA)*, a robust method for evaluating the relative efficiency of different context compression strategies. DEA is a non-parametric approach that treats each method as a Decision-Making Unit (DMU), where memory consumption is considered the input and performance (measured by average score) is the output. By constructing a linear programming model, we assess the efficiency of each compression method, considering both their computational cost and ability to maintain performance across tasks. As shown in Figure 4, the Efficiency Scores reveal crucial insights: **hybrid approaches, notably the combination of KIVI, Minference, and LLMLingua2x, emerge as the most efficient configuration overall.** This hybrid strategy strikes the best balance, effectively improving performance while minimizing memory usage. The results highlight that hybrid methods outperform individual techniques by integrating complementary strengths, making them an ideal choice for applications like LiveLongBench, where both performance and resource constraints are critical.

Answer the Question (2):
The combination of Minference and Lingua achieves the best overall performance, while integrating all three methods (KIVI, Minference, and Lingua) strikes the most balanced trade-off between performance and memory efficiency.

5 Conclusion

In this work, we introduce LiveLongBench, the first benchmark for evaluating long-context understanding in live-stream spoken texts, featuring sequences of up to 97K tokens. Our evaluation shows that current LLMs suffer notable performance degradation when processing lengthy, informal speech due to redundancy, colloquial expressions, and complex discourse structures. To address these challenges, we found that a hybrid compression strategy that integrates multiple techniques can improve both performance and memory efficiency. Using DEA-based efficiency analysis, we determine the optimal balance among context length, computational cost, and performance. Overall, our study offers new insights into long-context compression and provides practical guidelines for enhancing LLM efficiency in real-world spoken-language applications.

6 Limitations

Our work has several limitations. First, LiveLongBench is primarily based on live-streaming content, which may not fully represent the variety of spoken language found in other domains, such as academic lectures or news broadcasts. However, this focus was chosen to capture the dynamic and informal nature of live communication. Second, the evaluation process involves substantial annotation effort, as assessing long-context understanding requires bilingual experts to review extensive documents. Future work should explore automated solutions to reduce this cost while maintaining high evaluation quality.

References

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

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*.

Jian Chen, Peilin Zhou, Yining Hua, Yingxin Loh, Kehui Chen, Ziyuan Li, Bing Zhu, and Junwei Liang. 2024. Fintextqa: A dataset for long-form financial question answering. *arXiv preprint arXiv:2405.09980*.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pages 359–369. The Association for Computational Linguistics.

Alexander R Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir R Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. *arXiv preprint arXiv:1906.01749*.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, speech, and signal processing, ieee international conference on*, volume 1, pages 517–520. IEEE Computer Society.

Huiqiang Jiang, Yucheng Li, Chengruidong Zhang, Qianhui Wu, Xufang Luo, Surin Ahn, Zhenhua Han, Amir H Abdi, Dongsheng Li, Chin-Yew Lin, et al. 2024. Minference 1.0: Accelerating pre-filling for long-context llms via dynamic sparse attention. *arXiv preprint arXiv:2407.02490*.

Hongye Jin, Xiaotian Han, Jingfeng Yang, Zhimeng Jiang, Zirui Liu, Chia-Yuan Chang, Huiyuan Chen, and Xia Hu. 2024. Llm maybe longlm: Self-extend llm context window without tuning. *arXiv preprint arXiv:2401.01325*.

Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Wai-Chung Kwan, Xingshan Zeng, Yufei Wang, Yusen Sun, Liangyou Li, Lifeng Shang, Qun Liu, and Kam-Fai Wong. 2023. M4le: A multi-ability multi-range multi-task multi-domain long-context evaluation benchmark for large language models. *arXiv preprint arXiv:2310.19240*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024a. Lost in the middle: How language models use long contexts. *Trans. Assoc. Comput. Linguistics*, 12:157–173.

Zirui Liu, Jiayi Yuan, Hongye Jin, Shaochen Zhong, Zhaozhuo Xu, Vladimir Braverman, Beidi Chen, and Xia Hu. 2024b. Kivi: A tuning-free asymmetric 2bit quantization for kv cache. *arXiv preprint arXiv:2402.02750*.

Amirkeivan Mohtashami and Martin Jaggi. 2023. Landmark attention: Random-access infinite context length for transformers. *arXiv preprint arXiv:2305.16300*.

Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Menglin Xia, Xufang Luo, Jue Zhang, Qingwei Lin, Victor Rühle, Yuqing Yang, Chin-Yew Lin, et al. 2024. Llmlingua-2: Data distillation for efficient and faithful task-agnostic prompt compression. *arXiv preprint arXiv:2403.12968*.

Bo Peng, Xinyi Ling, Ziru Chen, Huan Sun, and Xia Ning. 2024. ecellm: Generalizing large language models for e-commerce from large-scale, high-quality instruction data. *arXiv preprint arXiv:2402.08831*.

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

Elior Sulem, Jamaal Hay, and Dan Roth. 2021. Do we know what we don't know? studying unanswerable questions beyond squad 2.0. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4543–4548.

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

Minzheng Wang, Longze Chen, Fu Cheng, Shengyi Liao, Xinghua Zhang, Bingli Wu, Haiyang Yu, Nan Xu, Lei Zhang, Run Luo, Yunshui Li, Min Yang, Fei Huang, and Yongbin Li. 2024a. Leave no document behind: Benchmarking long-context llms with extended multi-doc QA. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 5627–5646. Association for Computational Linguistics.

Minzheng Wang, Longze Chen, Fu Cheng, Shengyi Liao, Xinghua Zhang, Bingli Wu, Haiyang Yu, Nan Xu, Lei Zhang, Run Luo, et al. 2024b. Leave no document behind: Benchmarking long-context llms with extended multi-doc qa. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5627–5646.

Saizheng Zhang. 2018. Personalizing dialogue agents: I have a dog, do you have pets too. *arXiv preprint arXiv:1801.07243*.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Hao, Xu Han, Zhen Thai, Shuo Wang, Zhiyuan Liu, et al. 2024a. ∞ bench: Extending long context evaluation beyond 100k tokens. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15262–15277.

Zhihan Zhang, Yixin Cao, Chenchen Ye, Yunshan Ma, Lizi Liao, and Tat-Seng Chua. 2024b. Analyzing temporal complex events with large language models? a benchmark towards temporal, long context understanding. *arXiv preprint arXiv:2406.02472*.

A Appendix

A.1 Human Annotators.

To facilitate the evaluation of LLMs, we employed a group of students as human annotators to provide gold-standard labels for the datasets used in our study. These human-generated scores serve as a reference point for comparing the performance of various LLMs. Next, we will introduce the annotation process in detail.

Selection of Annotators. We selected five students with relevant background knowledge for the task. The annotators have been trained to ensure consistency and accuracy in their labeling, with a focus on the specific requirements of our dataset.

Cost of the Annotation. The annotation task was carried out by five full-time students over two days. With each student receiving a monthly salary of 800 RMB, the total cost for this annotation effort amounted to around 400 RMB.

Quality Control. To maintain high annotation quality, we conducted regular quality checks throughout the process. This included cross-checking annotations from different annotators and resolving discrepancies through consensus or review by senior researchers.

A.2 Further Analysis of the Data

LiveLongBench is constructed through a systematic data collection and processing pipeline, as illustrated in Figure 7. The benchmark integrates multiple task types relevant to long-context understanding in the live-streaming e-commerce domain, ensuring a comprehensive evaluation of large language models. The detailed statistics of each task within LongLiveBench are presented in Table 3, outlining key dataset characteristics such as the number of instances, average context length, and task-specific attributes. These details provide a quantitative overview of the dataset composition, highlighting its suitability for assessing KV cache optimization techniques in long-context scenarios.

Length of the Data. We present the statistics on the length of LifelongBench. Table 3 illustrates the average number of tokens, languages, and test instances across major categories (retrieval, reasoning, hybrid) and their fine-grained subcategories. In addition, we use a bar plot (see Figure 6) to illustrate the distribution of data lengths in LifelongBench. As shown, the data follows a power-law

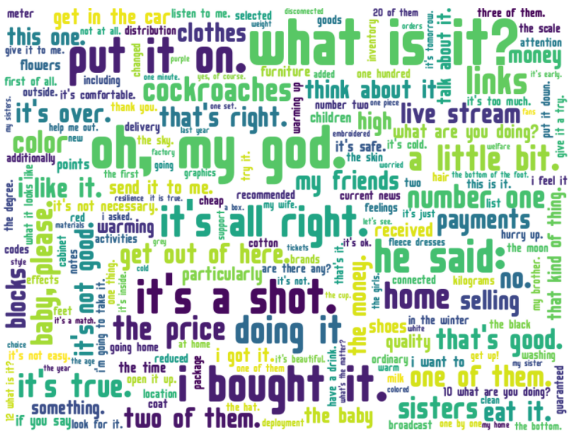


Figure 5: Wordcloud

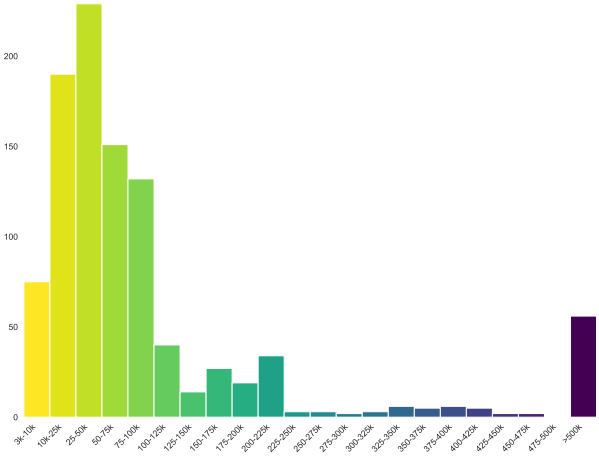


Figure 6: Distributions of the length in LiveLongBench

distribution, with the majority of instances concentrated below 220K tokens, while the overall distribution extends beyond 500K tokens.

World Cloud. To further explore the dataset, we generate a word cloud representation in Figure 5 that highlights the most frequent terms across the various categories and subcategories of LifelongBench. From this result, we observe a high degree of redundancy in the content, with frequent terms mostly consisting of discourse markers or exclamatory phrases, rather than being closely related to specific content. This observation aligns with the main challenges discussed in Section 3.1.

B Needle-in-a-Haystack Test

Experimental Setup. We follow the work (Mohtashami and Jaggi, 2023) to execute the Needle-in-a-Haystack Test. The corpus comprises live stream transcripts characterized by high redundancy and

Task Category	Avg Token	Language	#Test Instance
	Task		
Retrieval	132107.62	EN, ZH	443
Reasoning	20797.54	EN, ZH	129
Hybrid	85067.58	EN, ZH	434
	Sub Task		
Single Product Retrieval	147893.78	EN, ZH	351
Logistics Policy	71879.97	EN, ZH	92
Multiple Product comparison	101471.92	EN, ZH	349
Price Comparison	17713.27	EN, ZH	85
Product Classification	20531.62	EN, ZH	21
Live stream Summary	24380.45	EN, ZH	69

Table 3: Data statistics of LongLiveBench.

informal, spoken language. The results are presented in the Figure 8.

Results. Our results highlight the unique advantage of low-bit quantization in preserving retrieval performance, aligning with previous findings that retaining more information is critical for accurate retrieval. KIVI effectively reduces memory usage while maintaining retrieval accuracy, reinforcing the importance of information retention in long-context tasks. In addition, we also observe that the combination of MInference+KIVI consistently achieves strong retrieval performance, validating the effectiveness of hybrid compression methods in balancing efficiency and accuracy.

B.1 Needle-in-a-Haystack Test Details

Needle-in-a-Haystack (NIAH) a style of synthetically generated stress test designed to assess a language model's ability to retrieve specific information embedded within a large volume of unrelated background text. The core task involves inserting a critical piece of information at varying positions within different lengths of irrelevant content and then querying the model to recall this information accurately. Specifically,Mohtashami and Jaggi (2023) introduced a standardized passkey retrieval task, in which a key phrase formatted as "The pass key is <PASS KEY>. Remember it. <PASS KEY> is the pass key" is inserted into background text composed of repetitive generic sentences such as "The grass is green. The sky is blue. The sun is yellow. Here we go. There and back again." This formulation ensures that the task is purely focused on retrieval rather than inference. A variation of NIAH proposed by Greg Kamradt replaces the passkey with a more natural sentence, such as "The best thing to do in San Francisco is eat a switch and sit in Dolores Park on a sunny day," which serves as the retrievable target. In both formulations, the objective for large language models (LLMs) remains the same: they must successfully extract the inserted key information from an overwhelming amount of distractor text. Our implementation of the NIAH task closely follows the passkey retrieval template proposed by Mohtashami and Jaggi (2023). However, we introduce two key modifications: (1) the use of a 7-digit passkey instead of a generic phrase, and (2) the replacement of artificially structured background text with colloquial multi-domain live-streaming transcript fragments. This adjustment more closely reflects real-world applications where models must filter out irrelevant conversational noise while preserving and retrieving critical embedded information. As described in Arize-ai and Reid et al. (2024), the general retrieval prompt structure follows: "There is an important piece of information hidden inside a large volume of irrelevant text. Your task is to find and memorize it. I will later quiz you about this information." A standard filler, such as excerpts from Paul Graham's essays, precedes the inserted passkey phrase: `"The pass key is <7-DIGIT PASS KEY>. Remember it. <7-DIGIT PASS KEY> is the pass key." A suffix filler follows, after which the model is prompted with: "What is the pass key?"`

C Optimal Combination of Compression Methods with the Effect of SelfExtend

To evaluate the effectiveness of various KV cache compression methods and their combinations, we conduct experiments on LiveLongBench using LLaMA-3.1-8B-Instruct. The results, presented in Table 4, illustrate the performance of individ-

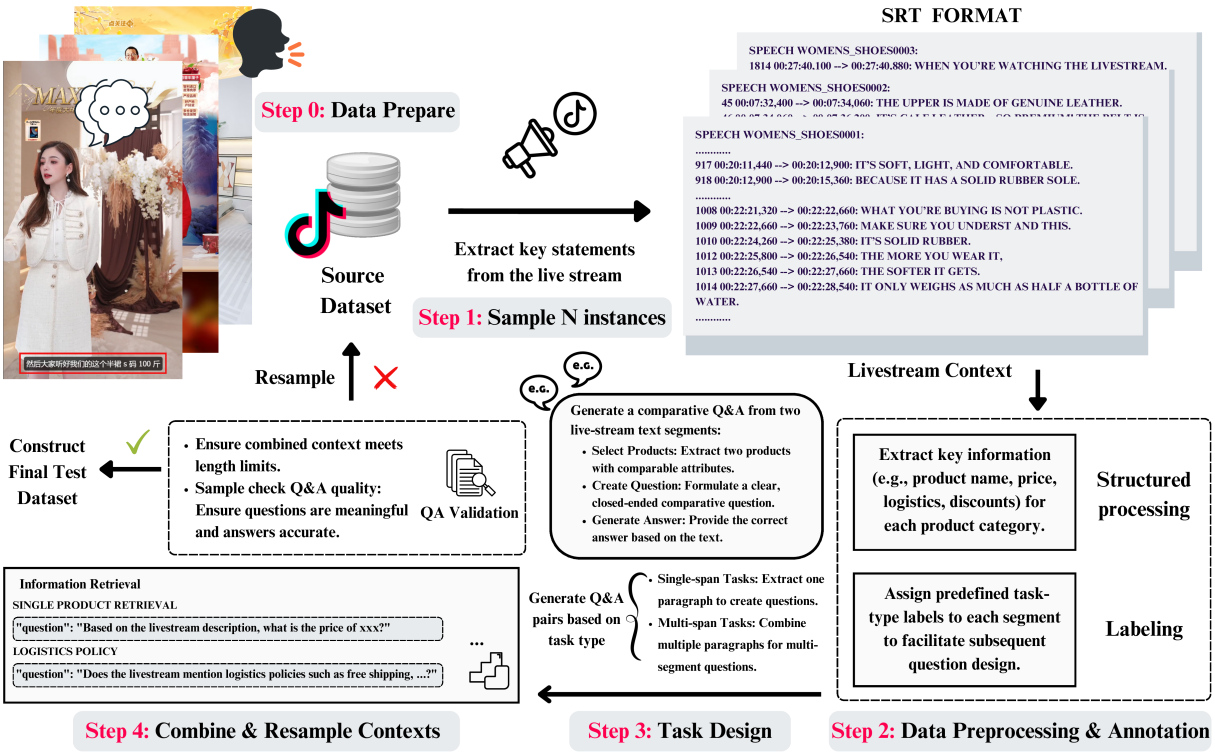


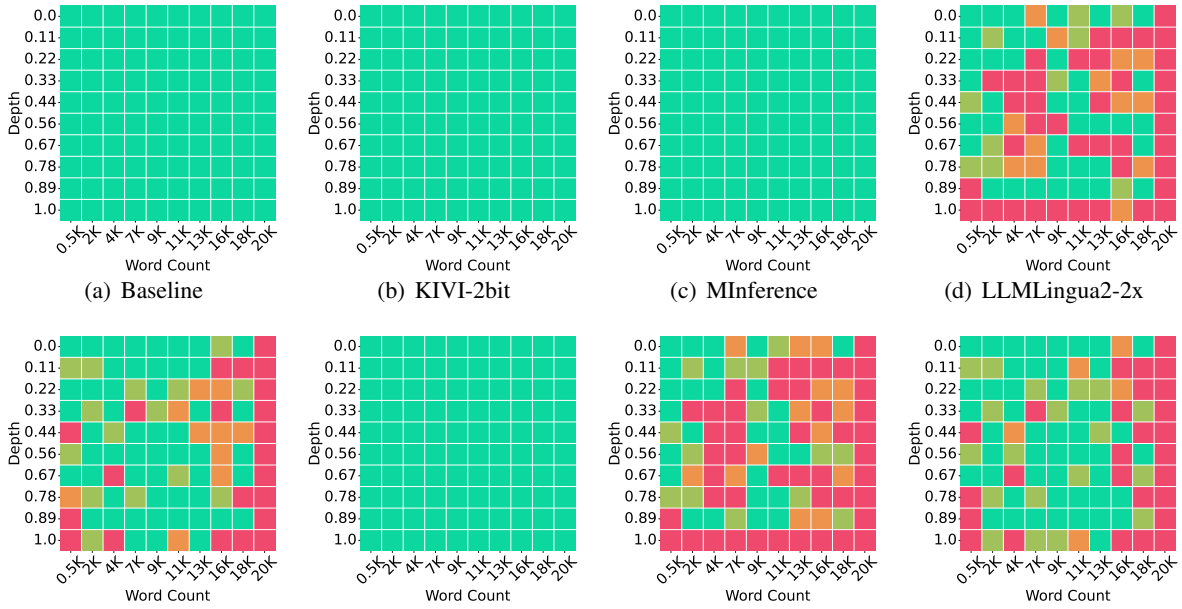
Figure 7: Illustrations of the Construction of LiveLongBench.

ual compression techniques as well as hybrid approaches, providing insights into their impact on long-context processing. The table details the accuracy and overall scores achieved under different configurations, highlighting the trade-offs between compression efficiency and model performance.

Building upon our evaluation of KV cache compression methods, we further explore the integration of Self-Extend (Jin et al., 2024), a self-regressive extension technique designed to enhance inference by expanding the context window of existing LLMs. As shown in Table 5, we incorporate Self-Extend into two compression method combinations: (1) the performance-optimal configuration, "MInference (③) + LLMLingua 4× (⑤)", and (2) the resource-performance balanced configuration, "KIVI 4-bit (①) + MInference (③) + LLMLingua 4× (⑤)", identified using the DEA method. In the table, different compression methods are denoted as follows: ① for KIVI, ③ for MInference, ④ for LLMLingua 2×, ⑤ for LLMLingua 4×, and ⑥ for Self-Extend. Experimental results demonstrate that incorporating Self-Extend (⑥) into the resource-optimal method further enhances inference performance, reinforcing the model's ability to process long-context inputs effectively.

D Case Study on the Performance of Different Compression Methods.

To help readers better understand the impact of KV cache compression methods on predictions, we provide several case studies in Figure 9 and Figure 10.



(a) Baseline (b) KIVI-2bit (c) MInference (d) LLMLingua2-2x

(e) LLMLingua + KIVI (f) MInference + KIVI (g) Lingua + MInference (h) Lingua + MInf + KIVI

Figure 8: Needle-in-a-Haystack results for each method on Llama-3-8B-Instruct. where a 20k words length input is converted to approximately 28k tokens.

		Score									
	Mem. (GB)	Retrieval			Hybrid			Reasoning			Overall
		Single	Policy	Avg.	Multi	Price	Avg.	Class	Sum.	Avg.	
Full	OOM	-	-	-	-	-	-	-	-	-	-
① KIVI 4bit	11.4	11.2	**80.0**	15.8	22.7	16.2	21.0	23.1	58.8	36.1	22.4
② KIVI 2bit	11.2	15.1	**80.0**	19.5	21.9	7.7	18.2	10.5	19.2	13.6	17.7
③ Minference	15.7	21.6	57.1	24.0	24.1	17.0	22.2	19.5	58.3	33.6	25.6
④ Lingua 2x	37.2	**26.0**	67.5	**28.8**	41.2	8.5	32.7	28.3	66.3	42.1	33.3
⑤ Lingua 4x	25.9	22.7	52.5	24.7	**46.6**	**25.0**	**41.0**	**39.5**	**72.5**	**51.5**	**36.7**
①+③	11.7	18.3	**75.0**	22.1	29.3	18.5	26.5	22.9	56.7	35.2	26.7
①+④	17.3	19.9	60.0	22.6	36.1	35.9	36.0	15.4	55.0	29.8	29.0
①+⑤	15.5	17.8	60.0	20.7	35.0	11.5	28.9	12.1	55.8	28.0	24.7
③+④	18.7	22.6	61.7	25.2	34.6	24.1	31.9	28.1	80.7	47.2	32.7
③+⑤	18.1	**26.4**	61.3	**28.7**	**46.1**	**42.3**	**45.1**	**34.5**	**81.3**	**51.5**	**39.8**
①+③+④	9.6	17.6	**60.0**	**20.4**	**34.7**	**31.5**	**33.9**	**18.6**	**61.7**	**34.2**	**28.4**
①+③+⑤	7.6	**17.9**	40.0	19.4	28.6	14.6	25.0	12.1	58.8	29.1	23.6
		Exact Match (%)									
	Mem. (GB)	Retrieval			Hybrid			Reasoning			Overall
		Single	Policy	Avg.	Multi	Price	Avg.	Class	Sum.	Avg.	
Full	OOM	-	-	-	-	-	-	-	-	-	-
① KIVI 4bit	11.4	1.8	75.0	6.8	2.7	0.0	2.0	0.0	0.0	0.0	3.5
② KIVI 2bit	11.2	5.5	75.0	10.2	2.7	0.0	2.0	0.0	0.0	0.0	4.9
③ Minference	15.7	10.9	57.1	14.0	8.1	0.0	6.0	0.0	0.0	0.0	8.0
④ Lingua 2x	37.2	14.6	50.0	17.0	18.9	0.0	14.0	0.0	16.7	6.1	13.4
⑤ Lingua 4x	25.9	10.9	25.0	11.9	18.9	7.7	16.0	14.3	8.3	12.1	13.4
①+③	11.7	12.7	75.0	17.0	10.8	7.7	10.0	0.0	0.0	0.0	10.6
①+④	17.3	9.1	50.0	11.9	16.2	29.4	19.7	0.0	0.0	0.0	11.9
①+⑤	15.5	10.9	50.0	13.6	13.5	0.0	10.0	0.0	8.3	3.0	9.9
③+④	18.7	12.5	33.3	13.9	20.0	9.5	17.3	4.8	13.0	7.8	13.7
③+⑤	18.1	20.0	25.0	20.3	16.2	15.4	16.0	9.5	8.3	9.1	16.2
①+③+④	9.6	9.1	50.0	11.9	10.8	7.7	10.0	0.0	0.0	0.0	8.5
①+③+⑤	7.6	9.1	25.0	10.2	10.8	0.0	8.0	12.5	8.3	4.9	7.8

Table 4: Performance of context compression methods on LLaMA-3.1-8B-Instruct.

	Mem. (GB)	***Score***									Overall
		Retrieval			Hybrid			Reasoning			
		Single	Policy	Avg.	Multi	Price	Avg.	Class	Sum.	Avg.	
③+⑤	18.1	**26.4**	61.3	**28.7**	**46.1**	42.3	**45.1**	34.5	**81.3**	**51.5**	**39.8**
③+⑤+⑥	18.1	18.7	**68.8**	22.1	39.7	**43.5**	40.7	**36.7**	72.1	50.0	35.0
①+③+⑤	7.6	**17.9**	40.0	**19.4**	28.6	14.6	25.0	12.1	58.8	29.1	23.6
①+③+⑤+⑥	7.6	14.6	**52.5**	17.2	**29.7**	**21.5**	**27.6**	**21.4**	**60.8**	**35.8**	**25.2**
		Exact Match (%)									
	Mem. (GB)	Retrieval			Hybrid			Reasoning			Overall
		Single	Policy	Avg.	Multi	Price	Avg.	Class	Sum.	Avg.	
③+⑤	18.1	20.0	25.0	20.3	16.2	15.4	16.0	9.5	8.3	9.1	16.2
③+⑤+⑥	18.1	12.7	25.0	13.6	16.2	7.7	14.0	14.3	25.0	18.2	14.8
①+③+⑤	7.6	9.1	25.0	10.2	10.8	0.0	8.0	12.5	8.3	4.9	7.8
①+③+⑤+⑥	7.6	3.6	25.0	5.1	10.8	7.7	10.0	9.5	9.1	7.3	7.8

Table 5: Optimal Combination of Compression Methods with the Effect of SelfExtend

(a) RETRIEVAL-DEPENDENT TASKS

LONG CONTEXT:

... You can wear this kind of vest underneath, and then layer it with other pieces, like how I did with my traditional costume.
I added some hoodies and lightweight down jackets inside it. It's all about staying close to the body.
If you have one of these, you'll stay warm on the front and back.
Today, the price is really good.
The sleeveless version is only $49.9, and the one with sleeves is $59.9.
You can choose between them. And if you pick any two pieces, you'll get a $10 discount. So, this is a great deal to consider!...

QUESTION: ACCORDING TO THE DESCRIPTION IN THE LIVESTREAM, WHAT IS THE PRICE OF THE THERMAL VEST?
Ground Truth: **49.9**

ANSWER:
Full: According to the description in the live streams, the price of the thermal vest was not mentioned.
KIVI 4 bit: The price of the thermal vest was not mentioned.
KIVI 2 bit: The answer is the pants priced at 86 yuan.
LLMLingua 2x: 49.9.
LLMLingua 4x: The price of the thermal vest is 49.9 yuan.
MInference: 129
KIVI 4 bit & MInference: 49.9
MInference & LLMLingua 4x: The price of the thermal vest is 49.9 yuan.
KIVI 4 bit & LLMLingua 4x: 49.9 yuan.
MInference & LLMLingua 4x & Self-Extend: The price of the thermal vest is 49.9 yuan.
KIVI 4 bit & MInference & LLMLingua 4x & Self-Extend: "This is a live streams content, where the host is introducing and displaying products. The main content includes:1. The host introduced the snow boots they were wearing, describing its material, design, and price...Based on the content, the price information is as follows: Snow Boots: 86 yuan and 132 yuan, ...Thermal Vest: 49.9, ..."

(b) RETRIEVAL-DEPENDENT TASKS

LONG CONTEXT:

... Today, we're introducing our new digital camera, the WD08Pro edition. ...
It has 4400K high definition and 64 million pixels.
In my live stream today, we support turning on the camera for a trial.
There's a 10-day trial period, shipping insurance, a three-year warranty, an upgraded version, a 5K lens, and 75 million high-definition pixels.
It's not suitable for concerts; it has an 18x zoom, which might be a bit challenging for concerts....

If you find it cheaper elsewhere, we'll cover the price difference.
Absolutely.
Get ready, everyone.
The link will be up in 3, 2, 1.
Hurry up and make your payments...

QUESTION: WHAT IS THE MOST EXPENSIVE PRODUCT?
Ground Truth: WD08Pro Camera

ANSWER:
Full: WD08Pro
KIVI 4 bit: WD08. WD08Pro version. The price of WD08Pro is over 800.
KIVI 2 bit: WD08. The price of WD08 is over 800...
LLMLingua 2x: Unable to determine
LLMLingua 4x: Unable to determine
MInference: Unable to determine
KIVI 4 bit & MInference: Unable to determine
MInference & LLMLingua 4x: Unable to determine
KIVI 4 bit & LLMLingua 4x: Unable to determine
MInference & LLMLingua 4x & Self-Extend: Unable to determine
KIVI 4 bit & MInference & LLMLingua 4x & Self-Extend: Unable to determine

Figure 9: Case Study of Retrieval-Dependent & Hybrid Tasks.

(c) RETRIEVAL-DEPENDENT TASKS

LONG CONTEXT:

... Let's talk about something else, like snow boots, huh? Snow boots. Which one are we talking about? What's the price? Friends, it's 86 yuan... You can wear these snow boots instead. Now I feel like a daughter of nature. Wow, it's like I just came back from skiing, isn't it? Wow, the color is so beautiful. Xiao just wants to get them right away. So, what's the price again? It's 86 yuan ...

QUESTION: PLEASE SUMMARIZE THE MAIN INFORMATION OF THE LIVESTREAM, INCLUDING THE FOLLOWING ASPECTS...

Ground Truth: ["Product or Service Introduction: Introduced products such as snow boots, hoodies, silver bracelets, red strings, necklaces, down jackets, and pants.", "Promotions and Discounts: Some products are on offer, such as 20 yuan off for two down jackets, and hoodies and pants are available in different colors.", "Important Announcements or Previews: Long down jackets are out of stock, some products are being restocked, and there are inventory alerts for some items."]

ANSWER:

KIVI 4 bit: "Product or Service Introduction: The products mentioned in the live broadcast include snow boots (86 yuan), Da Dong snow boots (86 yuan), ... , etc.", "Promotions and Discounts: The live broadcast mentioned a discount offer for two sweaters, with a 20% discount for purchasing two items.", "Special Announcements or Previews: No relevant information."

KIVI 2 bit: "Product or Service Introduction:" 1. Da Dong's Snow Boots 2. Lei Wen's Snow Boots 3. Ya Qin's Pants 4. Ya Qin's Pants 5. Ya Qin's Pants ...

LLMLingua 2x: "Product or Service Introduction: 1. Snow Boots: 86 yuan, with elastic laces, wide-leg fit, available in cowhide and synthetic leather, the material determines the price and warmth...Promotions and Discounts: 1. Down Jacket: 20 yuan off for two items, total 199 yuan.....Special Announcements or Previews: 1. Snow Boots: Not suitable for snowy and icy roads..."

LLMLingua 4x: "Product or Service Introduction:Snow Boots, their warmth, anti-slip properties, suitable for snowy conditions, priced at 86 yuan; ... Promotions and Discounts:Products from NetEase Select are relatively low-priced; discounts for two items, 20 yuan off for two; ... Special Announcements or Previews: ...if you have any questions, please leave a message, and we will respond; ..."

MInference: "Product or Service Introduction: Snow boots, Da Dong snow boots, Reebok snow boots, hoodies, ...", "Promotions and Discounts: Snow boots 88 yuan, hoodies 319 yuan, ...", "Special Announcements or Previews: No relevant information."

KIVI 4 bit & MInference: ["Pears", "Cheese Bag", "Pants", "Sweater", "Down Jacket", "Base Layer", "Pants", "Pants", "Pants", "Pants", "Pants",.....]

MInference & LLMLingua 4x: "Product or Service Introduction: Da Dong Snow Boots, priced at 86 yuan, suitable for snowy days, ...Promotions and Discounts: Discount for two items, 20 yuan off for two. ... Special Announcements or Previews: Peace and Joy Silver Bracelet, priced at 240 yuan, solid, with weight, and customizable engraving..."

KIVI 4 bit & LLMLingua 4x: "Product or Service Introduction" Snow Boots: 86 yuan, with a feeling of wearing snow boots,..., Promotions and Discounts" No relevant information, "Special Announcements or Previews" No relevant information

MInference & LLMLingua 4x & Self-Extend: "Product or Service Introduction: Snow Boots, priced at 86 yuan, suitable for snowy days...Promotions and DiscountsDiscount offer: 20 yuan off for two items...Announcements or Previews: Some products have limited stock, it is recommended that everyone purchase as soon as possible..."

KIVI 4 bit & MInference & LLMLingua 4x & Self-Extend: "Product or Service Introduction" 1. Snow Boots: 86 yuan, suitable for snowy and icy conditions, slip-resistant and waterproof. ...Promotions and Discounts: The livestream did not explicitly mention any promotions or discounts. Special Announcements or Previews: The livestream did not have any special announcements or previews.

Figure 10: Case Study of Reasoning-Dependent Tasks.