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 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 construct tasks in three categories: retrieval, reasoning, and hybrid tasks. We then evaluate both popular LLMs and specialized methods to assess their ability to understand long-contexts in these tasks. Our results show that current methods exhibit strong task-specific preferences and perform poorly on highly redundant inputs, with no single method consistently outperforming others. We propose a new baseline that better handles redundancy in spoken text and achieves strong performance across tasks. Our findings highlight key limitations of current methods and suggest future directions for improving long-context understanding. Finally, our benchmark fills a gap in evaluating long-context spoken language understanding and provides a practical foundation for developing real-world e-commerce systems. The code and benchmark will be made publicly available after the review process.

1 Introduction

Spoken texts, common in scenarios such as dialogues and live streams, are increasingly used in conversational AI. 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. (Chen et al., 2020) While advanced LLMs support long-context lengths (Touvron et al., 2023) and current Key-Value (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., 2024; Zhang et al., 2024a) for long-context understanding predominantly focus on written texts, neglecting the informal characteristics of spoken language. Beyond these understanding challenges, LLMs often waste computation on filler words (*e.g.,* "um", "uh") and other redundant content. These tokens contribute little semantic value but significantly expand the KV cache, increasing memory and latency costs (Dinkar et al., 2020). This motivates the need for more aggressive and selective context compression strategies (Li et al., 2023b). To explore this, we raise two central 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 assess how well existing LLMs and context compression methods handle long-form spoken texts, we construct a new benchmark, **LiveLongBench**, tailored to the challenges of this setting. As a core component, we construct a novel dataset recorded from live streams, with sequences averaging approximately 97K tokens. To tackle the first

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

Dataset	Response Type		Attention Span		Language Style	Avg. Tokens
	Closed	Open	Global	Semantic	Spoken Texts	
LongBench	✓	✓	✓			~13k
∞Bench	✓				✓	~300k
Loong		✓		✓		~110k
Marathon	✓					~163k
L-Eval	✓	✓	✓	✓	✓	3k - 62k
M4LE	✓	✓	✓	✓		~4k
TCELongBench	✓	✓		✓	✓	~18k
FinTextQA		✓	✓	✓		~19k
LiveLongBench	✓	✓	✓	✓	✓	~97k

question, we follow the study (Wang et al., 2024) and design 6 distinct tasks that fall into three categories: *retrieval*, *reasoning*, and *hybrid tasks*. For each category, we synthesize multiple questions to evaluate various model capabilities, covering both open-domain and closed-domain settings to assess knowledge recall and generalization. Furthermore, according to the findings in (Kwan et al., 2024), the critical information required for task completion in long sequences can be categorized by relevance into single-span, multiple-span, or global. Specifically, global tasks involve reasoning over the entire context and can be viewed as generalizing over other span types. As an important extension, we introduce semantic multi-span, which focuses on semantically distributed spans rather than merely structurally separated paragraphs. This task type can be seen as an advanced form of multi-span reasoning, requiring models to integrate and infer over multiple conceptually related but dispersed segments. Built upon these designs, LiveLongBench offers a thorough evaluation framework for long-context understanding in spoken language, which remains underexplored in existing benchmarks.

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 LLMLingua (Pan et al., 2024). Interestingly, we discover that **certain method combinations can lead to improved performance while reducing memory consumption, outperforming individual approaches**. For example, using *MInference+LLMLingua-4×* outperforms each single method, while using *KIVI-4bit+MInference+LLMLingua-2×* achieves the lowest memory usage and still surpasses individual approaches such as KIVI or MInference. To further balance memory efficiency and performance, we apply a Data Envelopment Analysis (DEA) framework to evaluate the cost-effectiveness of all method combinations. The resulting ranking offers a practical reference for selecting the optimal combination under different deployment constraints.

Our contributions are summarized as follows:

• We construct and release LiveLongBench, the first 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.

2 Related Work

Long-context Understanding Benchmarks. Numerous benchmarks have been developed to evaluate long-text understanding, predominantly focusing on formal, written texts (Shaham et al., 2022). Datasets such as TCELongBench (Zhang et al., 2024b), Loong (Wang et al., 2024) 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 (Rajpurkar et al., 2018). 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 has introduced benchmarks such as DailyDialog (Li et al., 2017), PersonaChat (Zhang et al., 2018), and DSTC (Williams et al., 2016), which feature short, goal-oriented dialogues with little noise or redundancy. Traditional SLU datasets like ATIS (Hemphill et al., 1990) and SNIPS (Coucke et al., 2018) are also widely used but remain highly structured and domain-specific with limited context. In contrast, corpora such as Switchboard (Godfrey et al., 1992) and CallHome (Kingsbury et al., 1997) capture the irregular, fragmented nature of natural spoken language, though restricted to telephony. Live streaming platforms now provide diverse spoken data, yet systematic collection and analysis are scarce. Recent work in video summarization (Song et al., 2015) and e-commerce dialogue datasets further highlights the need for specialized methods, as comprehensive solutions for long-form spoken language understanding remain underdeveloped.

Spoken Long-Text Benchmarks: Gaps and Advances. Existing long-text benchmarks focus largely on formal written language, overlooking the redundancy, informality, and variability of spoken texts, and rarely test redundancy reduction or long-context processing on authentic spoken data (Chen and Chen, 2008). As summarized in Table 1, LongBench (Bai et al., 2024) provides rich content but with evidence often confined to specific paragraphs; ∞Bench (Zhang et al., 2024a), Marathon, and Loong (Wang et al., 2024) extend context length but offer limited question diversity; L-Eval (An et al., 2024) and M4LE (Kwan et al., 2024) include varied tasks but over shorter inputs; and domain-specific datasets such as TCELongBench (Zhang et al., 2024b) and FinTextQA (Chen et al., 2024) target news or finance. In contrast, LiveLongBench preserves extensive context, covers a broader range of question types, and integrates spoken linguistic features, making it more representative of real-world spoken language.

3 LiveLongBench

3.1 Basic Challenges

To study long-context understanding in realistic spoken scenarios, we construct a dataset that captures the core challenges encountered in practice, particularly *informal language* and *high redundancy* (Li et al., 2023a). Next, we will describe the construction process in detail.

Informal Language. Live streaming e-commerce data often involves conversational speech, contributing to the informality of the language. Unlike formal text, live streaming content typically consists of short, fragmented utterances, leading to a high occurrence of syntactic reduction (Gliwa et al., 2019). Additionally, interactive conversations with viewers frequently introduce topic drift, where discussions shift abruptly, making it difficult for models to maintain contextual coherence (Wang et al., 2020). 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 area."
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 streaming 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 (Dinkar et al., 2020). This high redundancy poses challenges for long-context processing, requiring models to efficiently filter out noise while retaining essential details (Li et al., 2023b).

3.2 Dataset Collection

To tackle the challenges of long-context compression in spoken language, we construct a dataset through the pipeline illustrated in Figure 6 (Appendix A.2), designed to capture real-world characteristics at scale and support systematic evaluation across diverse spoken tasks.

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?"

Data Source. We constructed the dataset from Douyin (the Chinese version of TikTok) live streaming e-commerce sessions, covering 11 major product categories and 32 subcategories (*e.g.,* apparel, electronics, household goods) (Chang et al., 2024). Each document primarily consists of host monologues featuring informal expressions, repetitive promotions, and frequent Q&A exchanges (see Section 3.1), thus capturing the linguistic challenges of real-world spoken language. Audio was transcribed with Whisper, retaining repetitions and fillers to preserve authenticity, while detailed ASR performance can be found in Appendix A.2. Sensitive identifiers were removed for privacy, and a light filtering step eliminated extreme noise (e.g., sentences repeated over ten times), ensuring both realism and usability for downstream analysis.

3.3 Task Construction

Motivated by the study (Wang et al., 2024), we define three primary task categories that align with the inherent characteristics of spoken language (see Figure 1): 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, such as identifying product policy (*i.e.,* task "`Policy`"), or extracting product specifications from a single document (*i.e.,* task "`Single`"). These tasks may require the model to find the listed price of a product mentioned during a live stream or to verify its attributes based on the host's verbal 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. This includes classifying a product into the correct category (*i.e.,* task "`Class`"), which often requires external knowledge about product types and market conventions, *e.g.,* recognizing that a niche electronic device is a type of wearable. It also includes summarizing key points from lengthy and informal conversations (*i.e.,* task "`Summary`"), where the model must identify and synthesize essential information despite redundancy and noise, such as distilling a coherent summary from a promotional session with repeated slogans and off-topic remarks (Francia et al., 2020).

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 streaming transcript (*i.e.,* task "`Multiple Document QA`"), where the model must retrieve dispersed details based on semantic cues, such as descriptions of shoe style, material, and weight, typically scattered across different parts of the input. These attributes need to be accurately extracted and contextually integrated, challenging the model's ability to maintain coherence and perform fine-grained semantic alignment (Chai et al., 2021). Another example is the comparing product prices task(*i.e.,* task "`Price`"), where the model must identify price points mentioned at different times or in varying contexts, reason about fluctuations, detect potential discounts, and distinguish between original and adjusted prices to support informed comparison.

4 Experiments

Next, to answer the questions metioned in Section 1, we first report the performance of existing large language models on LiveLongBench (Section 4.1) and then introduce our new baseline for context compression (Section 4.2).

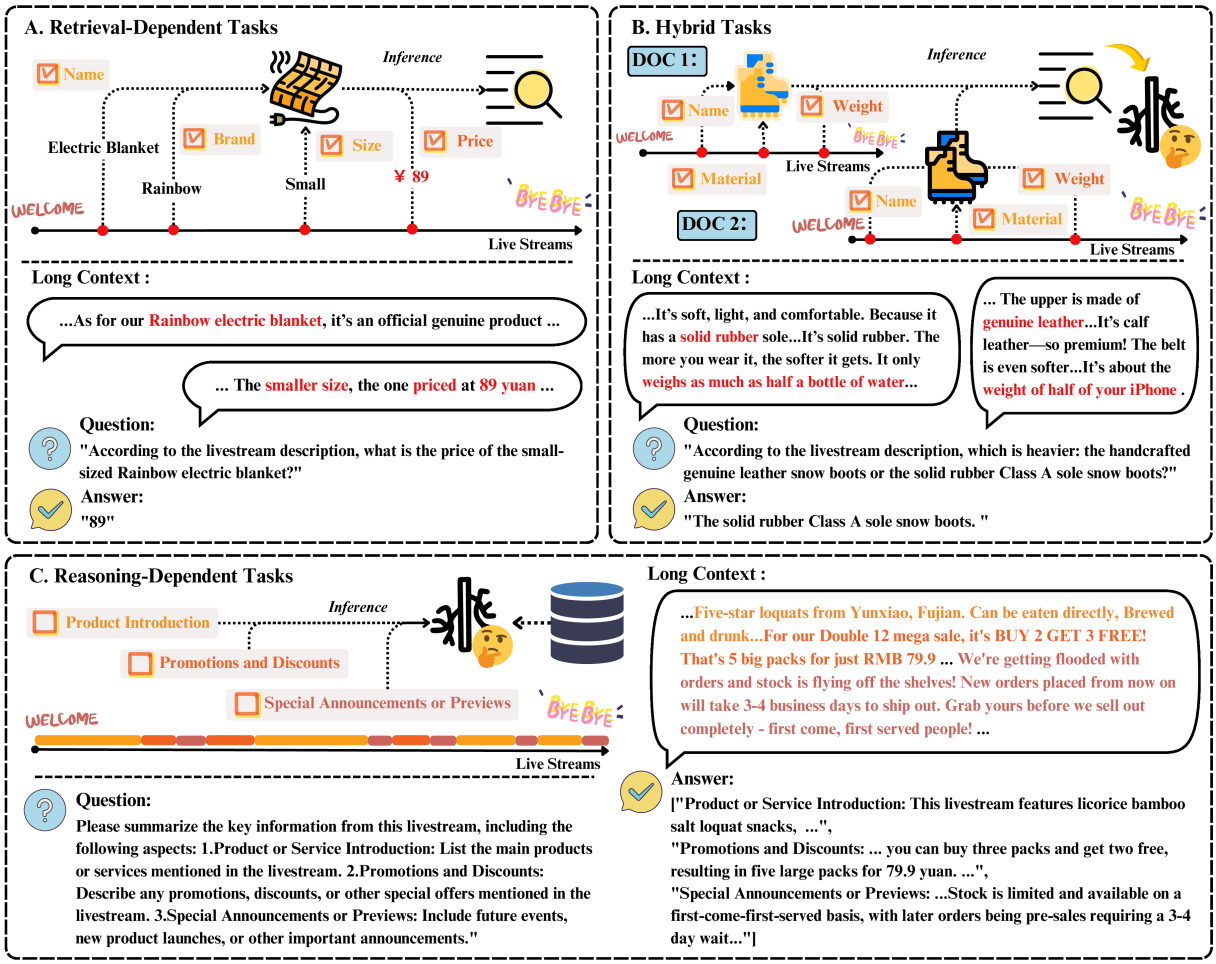


Figure 1: Showcase of Three Evaluation Tasks in LiveLongBench.

4.1 Large Language Models

Experimental Setup. We investigate whether foundation models can handle long and spoken queries using both closed-source models (GPT-4o, Gemini-1.5-pro, Claude-3.7-sonnet, GLM4plus) and open-source models (Qwen2.5-7B [1], LLaMA-3.1-8B [2], and Mistral-7B [3]). Our experimental setup ensures that each model is evaluated with identical decoding and prompting settings while using each model's native maximum context window. To investigate the impacts of domain-specific fine-tuning, we also include eCeLLM-M (Peng et al., 2024) [4], a model fine-tuned from the large base models Mistral-7B-Instruct for e-commerce, alongside general-purpose LLMs. For evaluation metrics, we employ *Exact Match* (%), which assesses whether the model's output exactly matches the ground-truth answer. This metric provides a strict evaluation of model correctness. In addition, we introduce a complementary metric, *Score*, which offers a softer and more fine-grained assessment by capturing partial correctness and enabling a continuous measure of model performance across tasks.

> **Findings on Research 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 long spoken texts.*

Comparison of Foundation Models. Table 2 presents the performance of various foundation models [5] across tasks. Overall, we find that closed-

[1] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct
[2] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
[3] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3
[4] https://huggingface.co/NingLab/eCeLLM-M
[5] For models with a maximum context length shorter than the benchmark input (e.g., Mistral-7B, 32K), we truncated inputs to the maximum supported length, ensuring consistency

		Score										
		Claimed Length	Retrieval Single	Retrieval Policy	Retrieval Avg.	Hybrid Multi	Hybrid Price	Hybrid Avg.	Reasoning Class	Reasoning Sum.	Reasoning Avg.	Overall
	Human	-	91.5	100.0	92.1	81.8	55.4	74.9	41.0	65.8	50.0	76.3
Closed	GPT-4o	128K	27.7	64.0	30.2	38.9	**62.3**	45.0	**87.0**	75.3	**82.7**	47.6
	Gemini-1.5-pro	1000K	**64.4**	85.0	**65.8**	**77.7**	27.9	**64.8**	59.8	91.4	71.3	**66.7**
	Claude-3.7-sonnet	200K	47.3	**100.0**	50.8	43.3	16.5	36.3	62.5	79.3	68.6	49.9
	GLM4plus	131K	25.1	75.0	28.5	34.1	16.5	29.5	36.2	**92.1**	56.5	35.4
Open	Qwen2.5-7B	131K	17.1	20.0	17.3	42.0	16.7	35.4	35.7	78.1	51.1	31.5
	LLaMA-3.1-8B	128K	19.2	74.6	23.0	30.9	33.2	31.5	39.7	64.1	48.6	31.9
	Mistral-7B	32K	9.0	80.0	13.8	33.9	13.5	28.6	33.8	52.1	40.5	25.2
	eCeLLM-M	32K	11.5	75.0	15.8	48.4	16.9	40.2	21.4	20.0	20.9	25.6
		Exact Match (%)										
		Claimed Length	Retrieval Single	Retrieval Policy	Retrieval Avg.	Hybrid Multi	Hybrid Price	Hybrid Avg.	Reasoning Class	Reasoning Sum.	Reasoning Avg.	Overall
	Human	-	89.1	100.0	89.8	51.4	15.4	42.0	4.8	8.3	6.1	53.5
Closed	GPT-4o	128K	20.6	40.0	21.9	**54.1**	**61.5**	**56.0**	56.0	58.6	**57.0**	**42.1**
	Gemini-1.5-pro	1000K	**52.3**	75.0	**53.8**	11.1	9.6	10.7	**59.2**	44.2	53.7	38.6
	Claude-3.7-sonnet	200K	34.6	**100.0**	39.0	28.4	11.8	24.0	12.5	**66.7**	32.2	32.1
	GLM4plus	131K	18.2	75.0	22.0	21.6	7.7	18.0	0.0	50.0	18.2	19.7
Open	Qwen2.5-7B	131K	10.9	0.0	10.2	16.2	6.7	13.7	0.0	30.8	11.2	11.7
	LLaMA-3.1-8B	128K	12.8	72.7	16.8	6.7	10.0	7.5	11.5	6.9	9.9	11.9
	Mistral-7B	32K	3.6	75.0	8.5	16.2	0.0	12.0	0.0	0.0	0.0	7.8
	eCeLLM-M	32K	5.5	75.0	10.2	35.1	7.7	28.0	0.0	0.0	0.0	14.1

Table 2: Performance comparison of large language models, including closed-source models (GPT-4o (128k), Gemini 1.5 Pro (1000k), Claude 3.7 Sonnet (200k), and GLM4plus (131k)) and widely used open-source models (Qwen, LLaMA, and Mistral). eCeLLM-M is a domain-specific model fine-tuned from Mistral-7B-Instruct.

source models generally outperform open-source ones. For example, Gemini-1.5-pro achieves the highest overall score (66.7). However, a notable performance gap remains between LLMs and human annotators, that is even with a 1M-token context window, Gemini-1.5-pro still falls short of human-level performance (76.3), indicating the benchmark remains challenging. In addition, we observe that longer context windows substantially improve performance on retrieval-related tasks. Gemini-1.5-pro significantly outperforms GLM4plus (131k context) in both retrieval (65.8 vs. 28.5) and hybrid tasks (64.8 vs. 29.5), demonstrating the benefit of extended context. By zooming in on open-source models, we observe that Qwen2.5-7B and LLaMA-3.1-8B reach overall scores of 31.5 and 31.9, respectively, approaching those of GLM4plus. In contrast, other models tend to perform well only on a limited subset of tasks. Specifically, Qwen2.5-7B performs well in reasoning tasks, while LLaMA3.1-8B excels at retrieval.

Impacts of Domain-specific Fine-tuning. Models pre-trained or fine-tuned in 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. As shown in Table 2, although Mistral-7B is not explicitly designed for understanding long contexts, domain-specific fine-tuning (eCeLLM-M) still improves its effectiveness in certain long input tasks. This improvement is likely attributed to its adaptation to domain-specific patterns and structures through fine-tuning. Notably, eCeLLM-M demonstrates superior performance in hybrid 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 the exact match of 0.0% among all the models evaluated.

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 to improve their performance.

with their native inference limits.

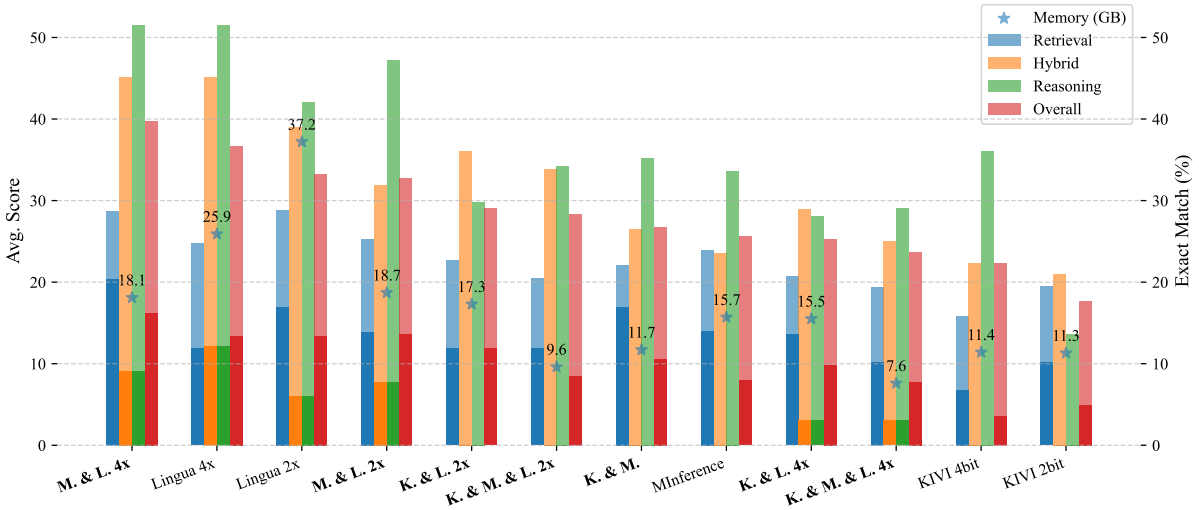


Figure 2: Performance of Context Compression Methods on LLaMA-3.1-8B-Instruct. "K." denotes KIVI, "M." denotes MInference, and "L." denotes LLMLingua, while "2x" and "4x" refer to compression ratios. Methods shown in bold along the x-axis represent multi-methods. From left to right, the methods are arranged in descending order of their Overall average scores. For each bar, the darker segment represents the "Exact Match (%)" score of the corresponding method. Detailed results are provided in Table 6 in the Appendix.

Experimental Setup. We evaluated representative context compression methods on LiveLongBench to assess their utility for long-context understanding and their performance on 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).
- *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. To illustrate the trade-offs introduced by different strategies, Figure 2 visualizes the performance of LLaMA-3.1-8B-Instruct across multiple compression settings. A detailed summary of the quantitative results is provided in Table 6 in the Appendix.

Single-Method Analysis. Our analysis reveals that different compression methods exhibit distinct preferences across tasks. 1) *Low-bit quantization excels in retrieval tasks by retaining complete information, which is essential for such tasks.* For example, Even under ultra-low-bit settings, KIVI achieves the highest retrieval score (80.0) in the `policy` task with minimal memory usage, but its performance drops in other tasks due to information loss from over-compression. The "Needle in the Haystack" task (See Appendix B.5) further validates KIVI's retrieval advantage, highlighting the role of information retention in accurate retrieval. 2) *In contrast, sparsification and token pruning hinder retrieval by discarding information but enhance reasoning performance.* For instance, LLMLingua, with a 4× 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. A case study shows that LLMLingua-4× enhances key information clarity (*e.g.,* price) by eliminating redundancy. While often used to cut computational costs in formal text, our findings reveal that compression in high-redundancy contexts also provides strong denoising benefits, improving accuracy and performance. For instance, LLMLingua-4× surprisingly outperforms 2× (36.7 vs. 33.3), indicating that stronger compression in high-redundancy contexts more effectively suppresses noise and boosts performance.

Multi-Methods Analysis. Our analysis highlights that combining different compression strate-

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."

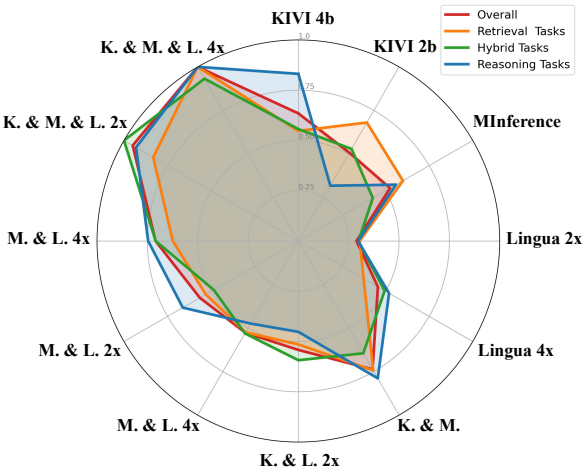


Figure 3: Efficiency Scores Based on DEA Analysis

gies can achieve extreme sparsity without compromising performance. Figure 2 shows that *MInference+LLMLingua-4×* delivers the best overall performance, balancing retrieval and reasoning. This strength likely stems from efficient memory use and selective token retention. In comparison, *MInference+LLMLingua-2×* excels in reasoning tasks, particularly logical inference, due to its prioritization of critical tokens and attention heads, though with slightly lower retrieval scores. KIVI combined with LLMLingua and MInference sustains strong retrieval but weakens reasoning, possibly due to excessive compression affecting long-range coherence.

Optimal Combination of Balancing Performance and Memory. As shown in Figure 3, 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. The resulting efficiency scores, illustrated in the figure, 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.

Findings on Research Question (2)

The combination of MInference and LLMLingua achieves the best overall performance, while integrating all three methods (KIVI, MInference, and LLMLingua) strikes the most balanced trade-off between performance and memory efficiency.

5 Conclusion

We present LiveLongBench, the first benchmark for evaluating long-context understanding in live-stream spoken data, with sequences averaging ∼ 97K tokens and extending beyond 500K. 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 mitigate these challenges, we show that a hybrid compression strategy combining multiple techniques enhances both performance and memory efficiency. Using DEA-based efficiency analysis, we determine the optimal balance among context length, computational cost, and performance. Overall, the study offers insights into long-context compression and practical guidelines for improving LLM efficiency in real-world spoken-language applications.

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.

Ethical Considerations

All live streaming data were collected from publicly accessible sessions in accordance with the platform's Terms of Service. Personally identifiable information (PII) was systematically removed during preprocessing to ensure user privacy, and only de-identified transcripts are retained for research purposes. The study protocol was reviewed for compliance with institutional research ethics standards, and no direct interaction with human subjects occurred beyond the use of publicly available data. The released benchmark is distributed under a research-only license, containing only metadata and de-identified text to mitigate potential risks of misuse.

References

Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. 2024. L-eval: Instituting standardized evaluation for long context language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14388–14411, Bangkok, Thailand. Association for Computational Linguistics.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3119–3137, Bangkok, Thailand. Association for Computational Linguistics.

Chunlei Chai, Guoliang Lu, Ruyun Wang, Chen Lyu, Lei Lyu, Peng Zhang, and Hong Liu. 2021. Graph-based structural difference analysis for video summarization. 577:483–509.

Jian-Peng Chang, Yan Su, Mirosław J. Skibniewski, and Zhen-Song Chen. 2024. Evaluating potential quality of e-commerce order fulfillment service: A collective intelligence-driven approach. 666:120425.

Berlin Chen and Yi-Ting Chen. 2008. Extractive spoken document summarization for information retrieval. 29(4):426–437.

Jian Chen, Peilin Zhou, Yining Hua, Loh Xin, Kehui Chen, Ziyuan Li, Bing Zhu, and Junwei Liang. 2024. FinTextQA: A dataset for long-form financial question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6025–6047, Bangkok, Thailand. Association for Computational Linguistics.

Luefeng Chen, Wanjuan Su, Yu Feng, Min Wu, Jinhua She, and Kaoru Hirota. 2020. Two-layer fuzzy multiple random forest for speech emotion recognition in human-robot interaction. 509:150–163.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*.

Tanvi Dinkar, Pierre Colombo, Matthieu Labeau, and Chloé Clavel. 2020. The importance of fillers for text representations of speech transcripts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7985–7993, Online. Association for Computational Linguistics.

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 Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

Matteo Francia, Matteo Golfarelli, and Stefano Rizzi. 2020. Summarization and visualization of multi-level and multi-dimensional itemsets. 520:63–85.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

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.

Charles T Hemphill, John J Godfrey, and George R Doddington. 1990. The atis spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*.

Huiqiang Jiang, Yucheng Li, Chengruidong Zhang, Qianhui Wu, Xufang Luo, Surin Ahn, Zhenhua Han, Amir H. Abdi, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2024. Minference 1.0: accelerating pre-filling for long-context llms via dynamic sparse attention. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Hongye Jin, Xiaotian Han, Jingfeng Yang, Zhimeng Jiang, Zirui Liu, Chia-Yuan Chang, Huiyuan Chen, and Xia Hu. 2024. Llm maybe longlm: Selfextend llm context window without tuning. ICML'24. JMLR.org.

Paul Kingsbury, David Graff, and George Zipperlen. 1997. Callhome american english transcripts. LDC97T14.

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, Yuxin Jiang, Lifeng Shang, Qun Liu, and Kam-Fai Wong. 2024. M4LE: A multi-ability multi-range multi-task multi-domain long-context evaluation benchmark for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15568–15592, Bangkok, Thailand. Association for Computational Linguistics.

Juan Li, Xueying Zhang, Fenglian Li, and Lixia Huang. 2023a. Speech emotion recognition based on optimized deep features of dual-channel complementary spectrogram. 649:119649.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Yucheng Li, Bo Dong, Frank Guerin, and Chenghua Lin. 2023b. Compressing context to enhance inference efficiency of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6342–6353, Singapore. Association for Computational Linguistics.

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 (Henry) Zhong, Zhaozhuo Xu, Vladimir Braverman, Beidi Chen, and Xia Hu. 2024b. Kivi: a tuning-free asymmetric 2bit quantization for kv cache.

Amirkeivan Mohtashami and Martin Jaggi. 2023. Random-access infinite context length for transformers. NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Menglin Xia, Xufang Luo, Jue Zhang, Qingwei Lin, Victor Rühle, Yuqing Yang, Chin-Yew Lin, H. Vicky Zhao, Lili Qiu, and Dongmei Zhang. 2024. LLMLingua-2: Data distillation for efficient and faithful task-agnostic prompt compression. pages 963–981.

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. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

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. Technical report.

Uri Shaham, Elad Segal, Maor Ivgi, Avia Efrat, Ori Yoran, Adi Haviv, Ankit Gupta, Wenhan Xiong, Mor Geva, Jonathan Berant, and Omer Levy. 2022. SCROLLS: Standardized CompaRison over long language sequences. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 12007–12021, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. 2015. Tvsum: Summarizing web videos using titles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5179–5187.

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.

Hei-Chia Wang, Wei-Fan Chen, and Chen-Yu Lin. 2020. NoteSum: An integrated note summarization system by using text mining algorithms. 513:536–552.

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. 2024. 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.

Jason D Williams, Antoine Raux, and Matthew Henderson. 2016. The dialog state tracking challenge series: A review. *Dialogue & Discourse*, 7(3):4–33.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

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. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1588–1606, Bangkok, Thailand. Association for Computational Linguistics.

Appendix A Data

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 involved five student assistants, each contributing two full days of work. Each annotator was compensated with 800 CNY, bringing the total cost of the annotation effort to 4,000 CNY.

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. Specifically, we applied 6 steps for the quanlity control. 1) Detailed Annotation Guidelines and Training, comprehensive guidelines were developed and training was conducted to ensure clear understanding of the annotation criteria. 2) Pilot Annotation Round, a pilot round was performed on a data subset to refine guidelines and address potential ambiguities. 3) Cross-Checking annotations, each annotation was independently verified by at least two annotators, with discrepancies flagged for review. 4) Consensus-Based resolution, conflicts were resolved through discussion; if consensus could not be reached, senior researchers provided the final decision. 5) Random Sample Review. A random subset of annotations was regularly re-evaluated by independent reviewers to ensure accuracy. 6) Continuous Feedback Loop, regular team meetings provide an ongoing channel for raising concerns and implementing improvements.

Appendix A.2 Further Analysis of the Data

LiveLongBench is constructed through a systematic data collection and processing pipeline, as illustrated in Figure 6. 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 benchmark is multilingual, consisting primarily of Mandarin live streaming transcripts along with a smaller set of English sessions. All data are analyzed in their original languages to preserve linguistic authenticity. The detailed statistics of each task within LongLiveBench are presented in To provide a clearer picture of the dataset composition, we report additional statistics in this

Table 3: Subcategory distribution (counts sum to their category total).

Category	Subcategory	Count	Share (%)
Baby, Kids & Pets	Pet Supplies	1	0.10
Baby, Kids & Pets	Children's Products	18	1.89
Baby, Kids & Pets	Children's Clothing	63	6.46
Baby, Kids & Pets	Children's Food	2	0.20
Digital & Electronics	Digital Accessories	11	1.09
Digital & Electronics	Audio & Video Equipment	9	0.89
Digital & Electronics	Home Appliances	19	2.09
Miscellaneous	General Categories	203	20.97
Jewelry & Collectibles	Jewelry	58	5.96
Jewelry & Collectibles	Collectibles	24	2.49
Food & Beverages	Beverages	51	5.27
Food & Beverages	Snacks	57	5.86
Smart Home	Home Essentials	61	6.26
Smart Home	Kitchenware	4	0.40
Smart Home	Bedding	18	1.89
Smart Home	Hardware Tools	22	2.29
Toys & Musical Instruments	Musical Instruments	7	0.70
Toys & Musical Instruments	Merchandise	4	0.40
Toys & Musical Instruments	Toys	18	1.89
Beauty & Personal Care	Makeup	30	3.08
Beauty & Personal Care	Skincare	3	0.30
Beauty & Personal Care	Beauty Devices	1	0.10
Shoes & Bags	Men's Shoes	63	6.46
Shoes & Bags	Bags	3	0.30
Shoes & Bags	Women's Shoes	34	3.58
Apparel & Underwear	Fashion Accessories	98	10.14
Apparel & Underwear	Men's Clothing	12	1.29
Apparel & Underwear	Women's Clothing	16	1.69
Apparel & Underwear	Underwear & Hosiery	26	2.58
Sports & Outdoor	Sports Equipment	14	1.49
Sports & Outdoor	Bicycles	19	1.89

subsection. First, Table 3 summarizes the distribution of product categories and subcategories across the e-commerce domain, reflecting the breadth of coverage in our benchmark. Next, Table 4 presents the detailed statistics of each task in LiveLongBench, offering a quantitative overview of task categories, instance counts, and average input lengths. Together, these tables demonstrate both the diversity of product coverage and the comprehensiveness of task design in LiveLongBench.

ASR performance. We evaluated several ASR systems, including Whisper, iFLYTEK [6], and Paraformer-zh [7], to obtain accurate results. Using the JiWER package, we compared the transcriptions with human-generated references and calculated the Word Error Rate (WER) and Character Error Rate (CER) for the long live streaming texts (see Table 5). We computed WER/CER on 48 hours of manually proof-read transcripts. Two annotators independently corrected outputs following a written protocol; disagreements were adjudicated by a senior reviewer. Metrics were computed with jiwer. Although iFLYTEK achieved high accuracy, it supports only 6-hour audio segments. For efficiency, we used Whisper for transcription followed

[6] iFLYTEK is a Chinese company known for its high-accuracy speech recognition system (see https://www.iflyrec.com/home/).

[7] Paraformer-zh is an automatic speech recognition model developed by Alibaba DAMO Academy (see https://huggingface.co/funasr/paraformer-zh).

Table 4: Data statistics of LiveLongBench.

Task Category	Sub Task Category	Task		Sub Task	
		Avg. Token	#Inst.	Avg. Token	#Inst.
Retrieval	Single Product Retrieval	132.11 K	443	147.89 K	351
	Logistics Policy Recognition			71.88 K	92
Reasoning	Product Classification	20.80 K	129	20.53 K	21
	Live Stream Summary			24.38 K	69
Hybrid	Price Comparison	85.07 K	395	17.71 K	85
	Multi Product Comparison			101.47 K	349

by manual proofreading. After proofreading, Whisper's transcription yielded a WER of 0.53% and a CER of 0.31%, confirming the high accuracy of our transcriptions.

ASR	WER	CER	Dur.
Whisper + Manual	**0.53%**	**0.31%**	∞
Whisper	12.74%	10.20%	∞
iFLYTEK	3.13%	2.35%	6h
Paraformer-zh	36.17%	26.69%	∞

Table 5: Word Error Rate (WER) and Character Error Rate (CER) of different models

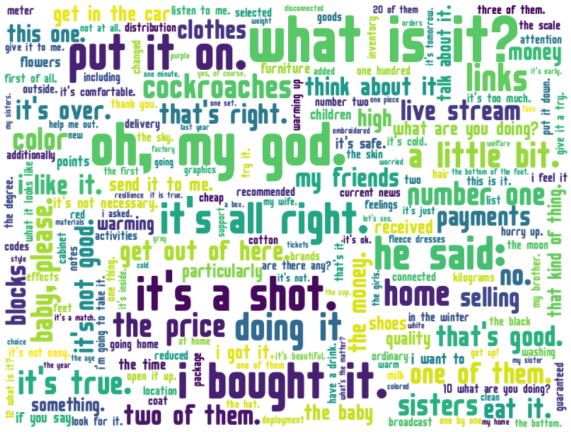


Figure 4: Wordcloud

Length of the Data. We present the statistics on the length of LifelongBench. Table 4 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 5) to illustrate the distribution of data lengths in LifelongBench. As shown, the data follows a power-law 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 4 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.

Appendix B Additional Experimental Details and Results

Appendix B.1 Evaluation Metrics.

We adopt two complementary metrics: (1) Exact Match (EM), the percentage of predictions exactly matching the ground truth, providing a strict measure of correctness; and (2) Score, a softer metric that assigns partial credit for answers with correct spans or elements, averaged across tasks. Both are reported throughout to give a comprehensive evaluation.

Appendix B.2 DEA Model Specification.

To analyze performance–efficiency trade-offs, we use an output-oriented CCR DEA model with memory usage (GB) as input and overall Score as output. Each compression method is treated as a decision-making unit, and efficiency is assessed against the DEA frontier. The linear program is solved with Pulp in Python. Robustness checks with $\pm 5\%$ memory perturbation showed stable efficiency rankings.

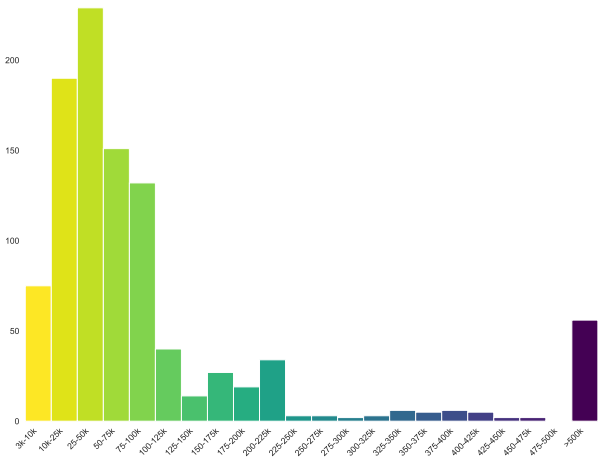


Figure 5: Distributions of the length in LiveLongBench

Appendix B.3 Detailed Results of Compression Methods.

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 detailed results, presented in Table 6, illustrate the performance of individual 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.

Appendix B.4 Optimal Combination of Compression Methods with the Effect of Self-Extend

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 7, 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.

Appendix B.5 Needle-in-a-Haystack Test

We follow the work (Mohtashami and Jaggi, 2023) to execute the Needle-in-a-Haystack Test. Needle-in-a-Haystack (NIAH) is 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. The corpus comprises live stream transcripts characterized by high redundancy and informal, spoken language. The results are presented in Figure 7. 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 informa-

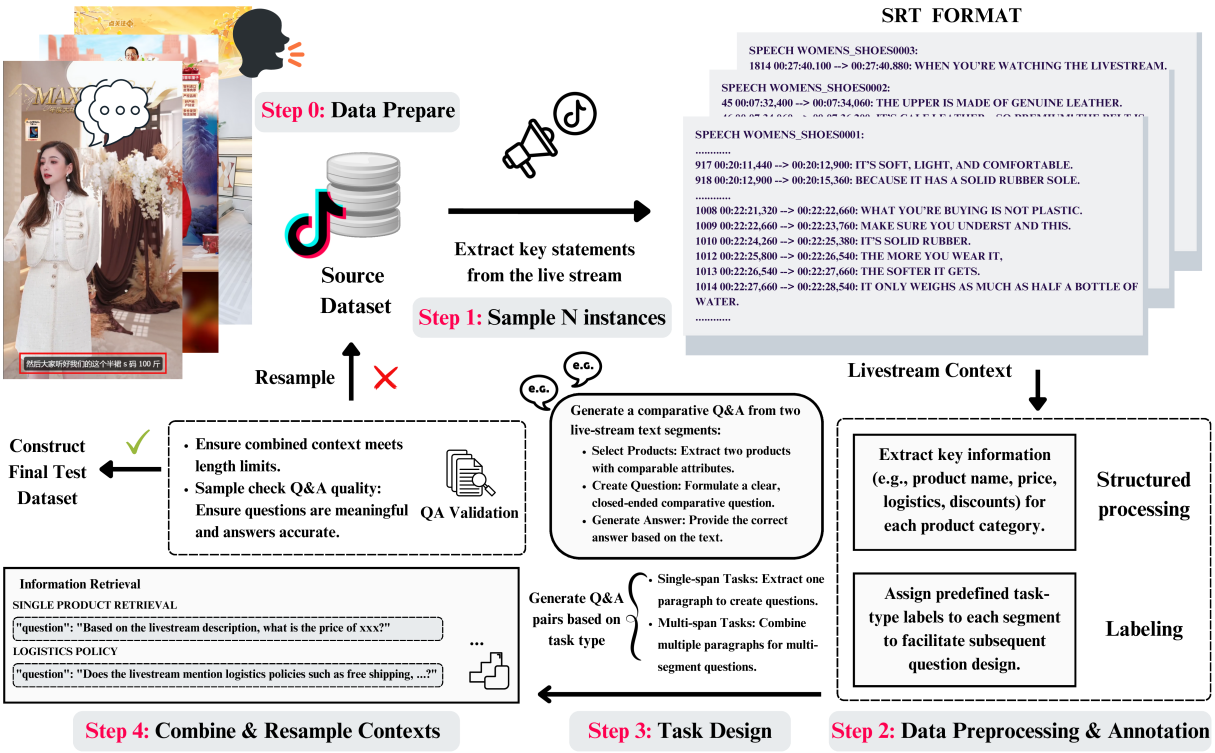


Figure 6: Illustrations of the Construction of LiveLongBench.

tion." 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?"` 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.

Appendix C 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 8 and Figure 9. The presented examples have been translated into English for ease of understanding and presentation. All analyses and evaluations, however, were performed on the original-language transcripts to maintain the authenticity of the spoken data.

Appendix D LLM Usage

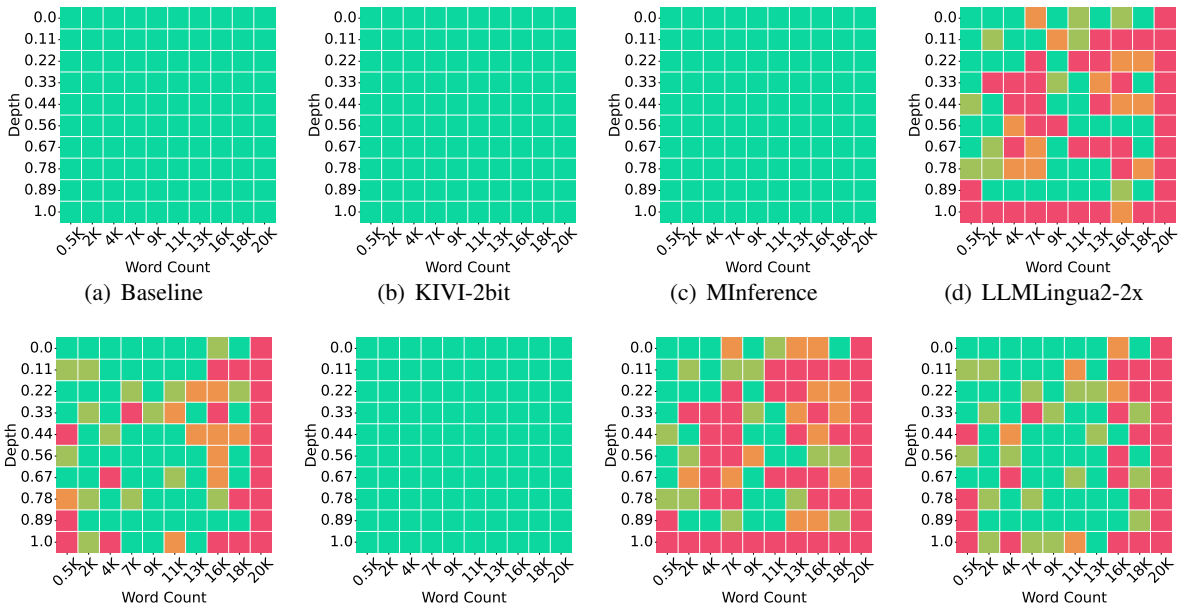
We use LLMs solely to check and correct grammatical errors in our paper.

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

		Score									
	Mem. (GB)	Retrieval			Hybrid			Reasoning			Overall
		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 7: Optimal Combination of Compression Methods with the Effect of SelfExtend



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

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

Figure 7: Needle-in-a-Haystack results for each method on Llama-3-8B-Instruct. A 20k-word input corresponds to approximately 28k tokens.

(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) HYBRID TASKS

LONG CONTEXT:

... This one here is our pink color, and this is our Peach Pink. It's part of our color-block series, with a light pink Peach Pink that really complements each other well. So, it's quite suitable for everyone to create a great outfit with, especially in pink. What else do we have?

QUESTION: IN THE LIVE BROADCAST, WHICH COLOR IS DARKER, THE PEACH PINK PAJAMAS OR THE VITALITY PINK PAJAMAS?
Ground Truth: Peach Pink Pajamas

Let's take a look at this color...Moving on, this is our light pink, which we call "Vitality Pink"—a soft, delicate pink. And then, we have this one with a small cuff at the bottom, where the purple and pink really complement each other. Next, we have our blue, a bubble blue, which is a bit deeper and a brighter shade of blue.Will it shed? Not really, so everyone can rest easy...

ANSWER:
Full: Unable to determine
KIVI 4 bit: Unable to determine
KIVI 2 bit: Unable to determine
LLMLingua 2x: Unable to determine
LLMLingua 4x: Unable to determine
MInference: Unable to determine
KIVI 4 bit & MInference: Unable to determine
MInference & LLMLingua 4x: Peach Pink Pajamas
KIVI 4 bit & LLMLingua 4x: Unable to determine
MInference & LLMLingua 4x & Self-Extend: Peach Pink Pajamas "Vitality Pink Pajamas" Unable to determine.
KIVI 4 bit & MInference & LLMLingua 4x & Self-Extend: "The host is showcasing and recommending some products, including clothing, shoes, hats, scarves, gloves, etc. Here is a brief description of these products:1. Clothing:H-fit down jacket (749 yuan)..."

Figure 8: Case Study of Retrieval-Dependent & Hybrid Tasks.

(c) REASONING-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: "Product or Service Introduction: Snow boots, down jackets, pants, shoes, etc. Promotions and Discount Activities: Snow boots for 86 yuan, down jackets for 749 yuan, pants for 79.9 yuan, shoes for 219 yuan, etc. Special Announcements or Previews: No relevant information available."
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 9: Case Study of Reasoning-Dependent Tasks.