Facilitating Long Context Understanding via Supervised Chain-of-Thought Reasoning

Jingyang Lin[1,2†] **Andy Wong**[2], **Tian Xia**[2], **Shenghua He**[2], **Hui Wei**[2,3†], **Mei Han**[2], **Jiebo Luo**[1]
[1] University of Rochester, NY, United States
[2] PAII Inc., CA, United States
[3] University of California, Merced, CA, United States
https://long-pai.github.io/

Abstract

Recent advances in Large Language Models (LLMs) have enabled them to process increasingly longer sequences, ranging from 2K to 2M tokens and even beyond. However, simply extending the input sequence length does not necessarily lead to effective long-context understanding. In this study, we integrate Chain-of-Thought (CoT) reasoning into LLMs in a *supervised* manner to facilitate effective long-context understanding. To achieve this, we introduce LongFinanceQA, a synthetic dataset in the financial domain designed to improve long-context reasoning. Unlike existing long-context synthetic data, LongFinanceQA includes intermediate CoT reasoning before the final conclusion, which encourages LLMs to perform explicit reasoning, improving accuracy and interpretability in long-context understanding. To generate synthetic CoT reasoning, we propose Property-based Agentic Inference (PAI), an agentic framework that simulates human-like reasoning steps, including property extraction, retrieval, and summarization. We evaluate PAI's reasoning capabilities by assessing GPT-4o-mini w/ PAI on the Loong benchmark, outperforming standard GPT-4o-mini by 20.0%. Furthermore, we fine-tune LLaMA-3.1-8B-Instruct on LongFinanceQA, achieving a 28.0% gain on Loong's financial subset.

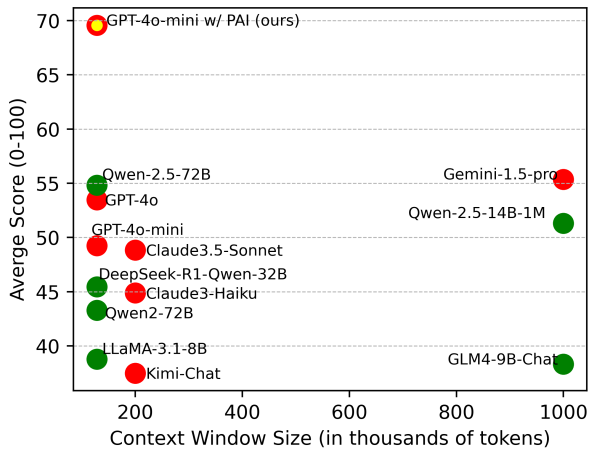

Figure 1: The hype of long-context large language models. The results shown are from the Loong benchmark, where green points refer to open-source LLMs and red points indicate closed-source LLMs. Our **GPT-4o-mini w/ PAI** stands out as the red open circle.

1 Introduction

Long context understanding remains an evolving challenge (Kočiský et al., 2018; Wu et al., 2021; Bai et al., 2024; Wang et al., 2024) in natural language processing (NLP). Achieving long-context understanding requires processing long-form textual information, thereby enhancing a model's ability to generate coherent, accurate, and contextually relevant responses (IBM Research, 2024). Practical long context understanding has a potential impact on numerous applications, such as private document analysis (Mukherjee et al., 2023), large codebase understanding (Nam et al., 2024), and multimodal content understanding (Chandrasegaran et al., 2024; Lin et al., 2023; Tang et al., 2023; Chen et al., 2024c). Recent advancements in large language models (Ouyang et al., 2022; Reid et al., 2024; Dubey et al., 2024) have significantly extended the input sequence length, ranging from 2K to 2M tokens, as shown in Figure 7 in Appendix A.

However, simply increasing the input sequence length does not necessarily improve the ability to comprehend the long content (Yang et al., 2024a; Goldman et al., 2024). In particular, Figure 1 presents the performance of various long-context LLMs on the Loong benchmark (Wang et al., 2024). The results suggest that, regardless of the maximum sequence length models can process, their performance remains similar, typically between 45% and 55%. This phenomenon exposes a hype that despite considerably enlarging context window size, the state-of-the-art LLMs still fail to perform satisfacto-

Corresponding author: ✉ jlin81@ur.rochester.edu.
†Work was done during the internship at PAII Inc.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 12232–12248
November 4-9, 2025

rily in practical long-context problem-solving tasks. Therefore, instead of merely increasing the input sequence length, achieving effective long-context understanding remains an open challenge.

In parallel with advanced long-context architectures and techniques (Peng et al., 2024; Liu et al., 2024c; Dao and Gu, 2024), constructing high-quality long-context training data remains essential yet underexplored. Given the scarcity and high annotation costs of long-context data, generating high-quality synthetic data for long-context modeling is valuable and urgent. Early attempts (Raffel et al., 2020; Fu et al., 2024) simply pack all short-length data into long-context chunks without considering document boundaries. Later works (Zhang et al., 2024a; He et al., 2024) have introduced long-context QA tasks from both single- and multi-source perspectives. Qwen-Agent (Yang et al., 2024a) develops an agentic system to further improve the quality of synthetic answers.

However, existing synthetic long-context data typically pair challenging questions with brief final answers for model training. This way overlooks a key difference between long-context and traditional QA tasks: practical long-context questions often require multi-step reasoning throughout the long content. Without intermediate reasoning, LLMs struggle to learn effective patterns from the paired complex questions and brief answers. We hypothesize that *directly guiding models to generate brief answers without intermediate reasoning steps for long-context modeling will lead to suboptimal training*. Instead, incorporating intermediate reasoning into synthetic data will help LLMs learn effective patterns and enhance training optimization.

To validate this hypothesis, we introduce *LongFinanceQA*, a novel long-context synthetic dataset constructed using financial data. Each sample in this dataset consists of a practical long-context question and the corresponding augmented answer along with intermediate chain-of-thought (CoT) reasoning steps. In particular, we first collect 6,911 bilingual financial annual reports (*i.e.*, English and Chinese) published before 2022. We then build a financial metric pool comprising key metrics commonly found in these reports, such as profit, cash flow, and debt. Based on these financial metrics, we generate 46,457 long-context questions that require single- or multi-source evidence. To generate reliable reasoning-augmented answers, we propose Property-based Agentic Inference (PAI), a comprehensive agentic framework. PAI leverages LLM-based agents to simulate human-like reasoning and operates in three steps: 1) **a property extraction agent** decomposes complex queries by identifying key properties, where each property consists of a metric and the corresponding subject; 2) **a property-based retrieval agent** retrieves relevant information from long documents for each identified property, and then generates intermediate findings by leveraging the property-based retrieved content; 3) **a summarization agent** synthesizes the intermediate findings into a coherent conclusion. The synthetic reasoning results are formed by integrating outputs from the property extraction and retrieval stages, while the conclusion is derived from the summarization stage. Finally, *LongFinanceQA* comprises 46,457 long-context QA pairs with CoT reasoning over 6,911 financial reports.

Although PAI performs human-like reasoning in long-context scenarios, it relies on human-crafted design and multi-step inference. To simplify this inference process, we intend to transfer PAI's long-context reasoning ability to a large language model, LLaMA-3.1, via supervised fine-tuning on LongFinanceQA. The enhanced model, LongPAI, leverages CoT reasoning to handle long-context problems in a single step. This fine-tuning procedure is termed as *Supervised CoT Reasoning*.

Empirically, we first evaluate the outcome of PAI on the Loong benchmark (Wang et al., 2024), involving challenging long-context tasks on three different domains. The results show that equipping GPT-4o-mini with PAI achieves a substantial 20% improvement on the Loong as shown in Figure 1. Moreover, the effectiveness of PAI guarantees the quality of the synthetic data in LongFinanceQA. Meanwhile, the enhanced LLaMA-3.1 model, LongPAI, achieves a 28.0% improvement on the *Financial* subset of Loong. Notably, in several scenarios, LongPAI even surpasses its teacher model PAI. This phenomenon emphasizes the importance of long-context modeling, contradicting recent arguments that *the long-context problem can be solved by short language models* (Qian et al., 2024; Chen et al., 2024b).

Our main contributions are three-fold: 1) we introduce *LongFinanceQA*, a long-context synthetic dataset for fine-tuning with 46,457 QA pairs featuring high-quality CoT reasoning from 6,911 bilingual financial annual reports; 2) to generate reasoning-augmented answers, we propose an agentic framework, *Property-based Agentic Inference*, to mimic human behaviors; and 3) empirical results validate the effectiveness of PAI and supervised CoT reasoning, the quality of LongFinanceQA, and the importance of long-context modeling.

2 Related Work

Long-context Synthetic Data. The scarcity of well-annotated long-context data makes high-quality synthetic data generation a valuable research direction. Early works (Raffel et al., 2020; Fu et al., 2024) concatenate short data into long-context fragments without considering the boundaries of the document. Large World Model (Liu et al., 2024b) addresses this boundary issue with masked sequence packing to keep attention within documents. Subsequent studies (Zhang et al., 2024a; He et al., 2024) construct single- and multi-source long-context QA pairs requiring evidence retrieval across multiple document positions. Qwen2-Agent (Yang et al., 2024a) leverages multiple agents to enhance answer quality. Beyond conventional QA pairs, our study augments answers with intermediate reasoning steps, explicitly guiding language models in learning reasoning abilities for practical long-context scenarios.

Long-context Large Language Models. Two main approaches enhance long-context problem-solving: reduction-based and extension-based methods. Reduction-based methods compress input by preserving essential information, allowing LLMs to focus on relevant content. Techniques like Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) and task decomposition for book summarization(Wu et al., 2021) follow this approach. Extension-based methods expand the context window directly. RoPE (Su et al., 2024) introduces a foundational positional encoding widely used in long-context LLMs (Peng et al., 2024; Yang et al., 2024a; Dubey et al., 2024). Parallelism techniques (Ren et al., 2021; Liu et al., 2024c) scale context capacity, enabling fully fine-tuning for long-context LLMs. This study leverages parallelism techniques to achieve long-context training.

Chain-of-Thought Resoning. The CoT technique improves the reasoning abilities of language models by incorporating intermediate reasoning steps. Early works (Wei et al., 2022) present that prompting LLMs with step-by-step reasoning significantly improves performance on complex reasoning tasks. Following works (Zelikman et al., 2022; Yao et al., 2024) explore structured CoT approaches, such as tree-based and self-consistent reasoning. This work incorporates CoT reasoning steps into fine-tuning data instead of well-crafted prompts. This approach enables the augmented data to explicitly supervise language models in reasoning skills.

Agentic RAG. Agentic RAG frameworks can be grouped into two paradigms (Liang et al., 2025): predefined reasoning and agentic reasoning RAG. Our proposed PAI framework is a hybrid of the two paradigms. Like predefined RAG systems (Press et al., 2023; Sarthi et al., 2024), PAI follows a structured, multi-step reasoning workflow. However, the components of PAI employ the techniques used in agentic reasoning RAG. Concretely, Property Extraction Agent and Proper-based Retrieval Agent use prompt-based approaches (*i.e.*, function calling (Eleti et al., 2023)) to generate structured outputs. Agentic RAG systems (Li et al., 2024; Yao et al., 2023; Eleti et al., 2023) typically treat follow-up queries as free-form text, which invites query drift, irrelevant results, and hard-to-audit reasoning. Different from agentic RAG, PAI instead converts each sub-query into a well-formed property, enabling verifiable retrieval, interpretable reasoning traces, and compact supervised training. Beyond agentic RAG, this study further investigates transferring the capability of long-context reasoning into a lightweight language model via supervised CoT, resulting in the LongPAI model.

3 Methodology

In this section, we first introduce the problem formulation of the QA task enhanced by intermediate CoT reasoning in Section 3.1. Then, we present the procedure of LongFinanceQA dataset construction in Section 3.2. After that, we describe the fine-tuning details in Section 3.3.

3.1 Problem Formulation

Traditional QA tasks require language models to generate outputs $\mathbf{A}$ directly from a given query $\mathbf{Q}$ and the corresponding input content $\mathbf{X}$ by modeling the conditional probability:

$$p_\theta(\mathbf{A}|\mathbf{X}, \mathbf{Q}), \quad (1)$$

where θ is the parameter of the language models.

Compared to traditional QA tasks, practical long-context QA tasks often require intermediate reasoning steps to analyze multiple pieces of evidence across long documents, and then derive the final answer. Thus, we formalize long-context QA tasks as a joint conditional probability of intermediate

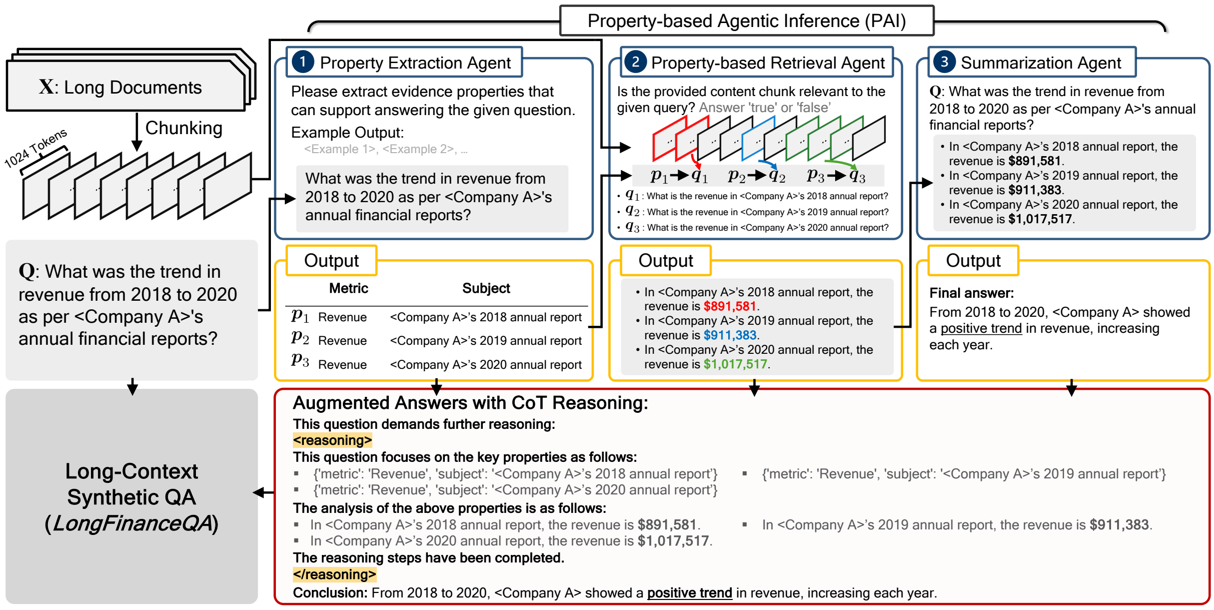

Figure 2: Overview of Property-based Agentic Inference (PAI), containing three stages. **A Property Extraction Agent** identifies key properties $\boldsymbol{p}_i$ from the given query $\mathbf{Q}$, where each property consists of a measurable metric and its corresponding subject. Given the selected properties, **a Property-based Retrieval Agent** first transforms each property into a sub-query $\boldsymbol{q}_i$ to retrieve relevant content chunks from long documents, yielding intermediate findings. **A Summarization Agent** integrates these intermediate findings to generate a comprehensive conclusion $\mathbf{A}$. After finishing PAI, we incorporate the output from the above three agents to produce reasoning-augmented answers. These augmented answers serve as the core contribution of the *LongFinanceQA*.

reasoning results $\mathbf{R}$ and the final answer $\mathbf{A}$:

$$p_\theta(\mathbf{R}, \mathbf{A}|\mathbf{X}, \mathbf{Q}). \tag{2}$$

As shown in Eq.(2), this study aims to facilitate long-context understanding by guiding LLMs to first predict intermediate CoT reasoning steps before generating the final answer. To achieve this, we construct a long-context synthetic dataset that explicitly supervises language models in learning intermediate reasoning. The data construction process can be factorized as follows:

$$p_\theta(\mathbf{R}, \mathbf{A}|\mathbf{X}, \mathbf{Q}) = p_\theta(\mathbf{R}|\mathbf{X}, \mathbf{Q}) \cdot p_\theta(\mathbf{A}|\mathbf{X}, \mathbf{Q}, \mathbf{R}), \tag{3}$$

where $p_\theta(\mathbf{R}|\mathbf{X}, \mathbf{Q})$ indicates the generation of intermediate reasoning steps $\mathbf{R}$ given a query $\mathbf{Q}$ and input content $\mathbf{X}$, and then $p_\theta(\mathbf{A}|\mathbf{X}, \mathbf{Q}, \mathbf{R})$ is the process of incorporating the query and summarizing reasoning steps to produce the answer $\mathbf{A}$. The data construction follows the principle of Eq.(3).

3.2 LongFinanceQA Dataset

LongFinanceQA dataset is designed to generate practical long-context QA pairs with reasoning steps to effectively analyze long content. The finance domain was chosen for several reasons. First, annual financial reports are readily accessible and present complex long-context reasoning challenges. Moreover, finance is a data-driven field where accurate insights can drive critical decisions, making advancements in AI particularly valuable for real-world applications. We will introduce the data construction pipeline of LongFinanceQA as follows.

Data Collection. To begin, we collect bilingual financial annual reports (*i.e*., English and Chinese) dated before 2022 from open-source official websites, specifically the SEC-10-K[1] and cninfo[2] platforms. Documents from companies included in the Loong benchmark are then excluded. Next, we filter reports based on a token length range of 20K to 80K, determined using the GPT-4o tokenizer. Moreover, we prioritize companies with consistent annual reports over the years. In the end, 6,911 bilingual financial reports are selected.

Diverse Query Generation. We first construct a financial metric pool containing key metrics commonly found in financial reports, such as profit, revenue, and cash flow. Then, we randomly pick several metrics in the metric pool and select a combination of financial reports. Given these financial metrics and the metadata from the selected documents (*e.g*., company name and year), we generate various long-context questions that require either

[1] https://www.sec.gov/
[2] http://www.cninfo.com.cn/

single- or multi-source evidence. Next, we filter out questions whose corresponding combined documents exceed 256K tokens, as this surpasses the maximum token limit of our model. Finally, we obtain 46,457 practical long-context questions. Please refer to Appendix A for more details.

Augmented Answer Generation with CoT Reasoning. Given long-context questions and their corresponding documents, our goal is to generate answers with CoT reasoning, following the principle of Eq. (3). In particular, we first generate step-by-step reasoning based on the question and documents, then integrate the reasoning steps into a conclusion. Generally, long-context questions are challenging as they require models to retrieve multiple pieces of evidence scattered throughout long content and then integrate them for a global understanding (Wang et al., 2024; Edge et al., 2024). Inspired by this phenomenon, we intend to first extract key evidence points that support answering the question, then retrieve relevant information based on these points, and finally aggregate them into the conclusion. We term these supporting evidence points as "*Properties*". Following this methodology, we propose the **Property-based Agentic Inference (PAI)** framework, illustrated in Figure 2, containing three steps: (1) property extraction, (2) property-based retrieval, and (3) summarization.

Step 1: Property Extraction. This step focuses on extracting a set of properties $\{\boldsymbol{p}_i\}_{i=1}^{N_p}$ from the given query $\mathbf{Q}$, where N_p denotes the number of properties. Each property consists of a metric (a measurable factor being analyzed) and its corresponding subject mentioned in the query $\mathbf{Q}$. For instance, given the query shown in Figure 2, the metric is "*revenue*" and subjects are "<Company A>'s annual reports from different years (*e.g.*, 2018, 2020, and 2022)", which serve as the sources of information to determine the trend.

Step 2: Property-based Retrieval. After extracting the properties $\{\boldsymbol{p}_i\}_{i=1}^{N_p}$ from the given query, this step aims to retrieve relevant information based on these properties. Specifically, each property is first transformed into a sub-query $\boldsymbol{q}_i$, which is then matched against relevant content chunks. These chunks are derived from the original long documents $\mathbf{X}$, with each chunk limited to 1,024 tokens. Based on the sub-queries and their corresponding retrieved chunks, we can derive several intermediate findings (*i.e.*, sub-answers).

Step 3: Summarization. In this final step, the original query $\mathbf{Q}$ and all intermediate findings from the second step are integrated to generate a comprehensive conclusion $\mathbf{A}$.

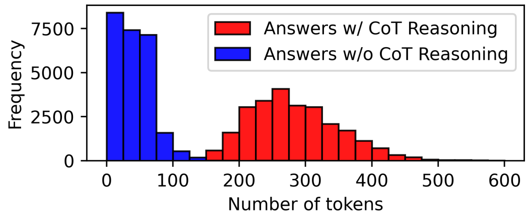

Figure 3: Token length distribution of answers with and without CoT reasoning from *Multi-Source* QA pairs in the proposed LongFinanceQA dataset.

The PAI serves as a low-cost annotator for the LongFinanceQA dataset. Using PAI, we produce augmented answers with CoT reasoning by combining the intermediate reasoning results from the first two steps with the conclusion. Figure 3 illustrates the token length difference between answers with and without CoT reasoning, showing that reasoning-augmented answers are nearly 200 tokens longer on average.[3] Please check out more data statistics in the Appendix A.

3.3 Supervised CoT Reasoning

Although PAI can behave like a human in long-context scenarios, it requires human-crafted design and multi-step inference. To address this limitation, we seek to transfer the long-context reasoning capability of PAI to existing language models, enabling them to analyze long content in a single-step inference. To this end, we fine-tune a language model on LongFinanceQA to explicitly guide it in learning CoT reasoning. This fine-tuning procedure is termed as *Supervised CoT Reasoning*. Specifically, we use LLaMA-3.1-8B-Instruct (Dubey et al., 2024) as the base language model. Following (Fu et al., 2024), we first extend the context window of LLaMA-3.1 from 128K to 262K through continued pretraining on 1.6B tokens, which consist of packed documents from Slimpajama (Soboleva et al., 2023). After extending context length, we fine-tune the extended LLaMA-3.1 model (parameterized by θ) on LongFinanceQA with reasoning-augmented answers.

Through this process, we obtain an enhanced LongPAI model by maximizing the log-likelihood of the predicted answer $\mathbf{Y}$ along with the CoT reasoning $\mathbf{R}$, conditioned on the given long documents

[3]The answers are tokenized by the LLaMA-3.1 tokenizer.

$\mathbf{X}$ and the query $\mathbf{Q}$. The objective is calculated as:

$$\mathcal{L}(\theta) = \sum_{i=1}^{N_r+N_y} \log p_\theta(\mathbf{R}, \mathbf{Y}|\mathbf{X}, \mathbf{Q}), \tag{4}$$

where N_r is the number of reasoning tokens and N_y indicates the answering token, including those in the properties, sub-answers, and final answers. We mask non-answer positions during the fine-tuning, allowing for more efficient learning. Although we only consider the LLaMA-3.1 model here, LongFinanceQA can also be used to fine-tune other language models (See Section 4.1 for details).

4 Experiments

4.1 Experimental Setup

Evaluation Benchmarks. We evaluate long-context understanding using two practical benchmarks: Loong (Wang et al., 2024) and ∞Bench (Zhang et al., 2024b). Loong focuses on real-world multi-document question answering and comprises 1,600 test samples across four categories, including Spotlight Locating, Comparison, Clustering, and Chain of Reasoning. These tasks assess distinct capabilities in handling long-context tasks. ∞Bench facilitates multilingual evaluation, assessing models on English (En.QA) and Chinese (Zh.QA) question-answering tasks that require long-range dependency and reasoning beyond short-passage retrieval.

Evaluation Metrics. For evaluation, Loong employs GPT-4-Turbo as a judge, scoring model responses based on accuracy, hallucinations, and completeness on a scale of 0 to 100. Meanwhile, it introduces the Perfect Rate, measuring the proportion of responses achieving a perfect score. In ∞Bench, model performance is measured by the F1 score (Zhang et al., 2024b).

Base Models. We adopt GPT-4o-mini (Achiam et al., 2023) and LLaMA-3.1-8B-Instruct (Dubey et al., 2024) as our base models. Specifically, GPT-4o-mini serves as the agent of the proposed Property-based Agentic Inference (PAI), while LLaMA-3.1-8B is used as the base model for LongPAI, which is fine-tuned on the LongFinanceQA.

Implementation Details. To enable training on long sequences (> 250K), we employ several optimization techniques, including flash-attention-2 (Dao, 2023) and ring-attention (Liu et al., 2024c). Furthermore, we adopt a zigzag sharding approach (Zhu, 2024) within ring attention for more effective load distribution across multiple GPUs. This training setup allows us to fine-tune the large language models fully. Using 8 A100 GPUs, the long-context training is completed in three days for fine-tuning. In addition, in training the long-context LLMs, we adopt the rotary base scaling approach (Liu et al., 2024d) and scale up the base value (*i.e.*, rotary base of 1,247,820) to adapt RoPE to a longer context. For optimization, we use a constant learning rate of 1e-5 for the entire training procedure. Following the common practice (Fu et al., 2024), we set the batch size to 16M tokens as mentioned in (Dubey et al., 2024). During inference, we set the temperature as zero to eliminate the randomness. We also increase the maximum output tokens to 1,024 since the CoT reasoning requires more output tokens.

4.2 Main Results

In this section, we first assess the quality of our long-context synthetic data (LongFinanceQA) by evaluating the performance of the proposed PAI on the Loong benchmark. Next, we compare the enhanced long-context language model (LongPAI) with its base LLaMA-3.1 model and other state-of-the-art LLMs on the *Finance* subset of Loong.

Data Quality Assessment. Reasoning-augmented answers in LongFinanceQA are automatically generated by the PAI framework. To assess the quality of these synthetic answers, we evaluate the annotator, PAI, on the Long benchmark, using GPT-4o-mini as the agent within the PAI, referred to as GPT-4o-mini w/ PAI. Table 1 shows that the GPT-4o-mini-based PAI framework significantly enhances the base model's performance, demonstrating the effectiveness of the PAI. In particular, compared to the standard GPT-4o-mini model, the overall average score improves by 20.3%, with substantial gains in key tasks such as spotlight locating (+20.2%), comparison (+15.7%), clustering (+29.2%), and chain of reasoning (+10.2%). Although the basic GPT-4o-mini is not the strongest model, GPT-4o-mini w/ PAI outperforms the state-of-the-art closed-source model, Geneni-1.5-pro (Reid et al., 2024), by over 15%. Moreover, the superior performance in the Spotlight Locating task (*i.e.*, single-source QA task) *highlights the quality of intermediate reasoning results generated by the PAI*, as predictions in this task serve as essential reasoning steps for the other three multi-source QA tasks. Consequently, the strong performance on the Loong demonstrates *the capability of PAI as a*

Table 1: Data quality assessment for the long-context synthetic dataset (*LongFinanceQA*) by measuring the performance of PAI on the Loong benchmark. *AS* denotes *Average Scores (0-100)*, and *PR* represents the *Perfect Rate* (0-1). Green highlights the remarkable improvements over the base model (GPT-4o-mini).

Model	**Context**	**Spotlight Locating**		**Comparison**		**Clustering**		**Chain of Reasoning**		**Overall**	
	Length	***AS***	***PR***	***AS***	***PR***	***AS***	***PR***	***AS***	***PR***	***AS***	***PR***
Open-Source Long-Context Large Language Models											
LLaMA-3.1-8B-Instruct	128K	62.42	0.52	39.13	0.21	25.96	0.01	44.20	0.22	38.79	0.18
DeepSeek-R1-Qwen-32B	128K	51.68	0.41	49.25	0.34	41.53	0.16	45.00	0.30	45.45	0.27
Qwen2-72B-Instruct	128K	54.17	0.36	42.38	0.20	36.71	0.04	47.76	0.18	43.29	0.15
Qwen2.5-72B-Instruct	128K	65.08	0.55	51.90	0.30	46.07	0.08	64.43	0.40	54.83	0.28
Qwen2.5-14B-Instruct-1M	1000K	67.50	0.58	55.12	0.35	39.05	0.04	57.81	0.31	51.30	0.25
Closed-Source Long-Context Large Language Models											
Kimi-Chat	200K	60.98	0.50	34.74	0.13	28.76	0.04	38.52	0.15	37.49	0.16
Claude3.5-Sonnet	200K	58.45	0.49	54.21	0.35	45.77	0.07	43.92	0.25	48.85	0.23
GPT-4o	128K	73.95	0.62	50.50	0.28	44.29	0.09	57.95	0.28	53.47	0.26
Gemini-1.5-pro	1000K	75.02	0.56	49.94	0.27	44.10	0.09	64.97	0.37	55.37	0.27
GPT-4o-mini (Base)	128K	59.46	0.49	51.90	0.27	34.55	0.04	64.28	0.39	49.25	0.24
GPT-4o-mini w/ PAI (*ours*)	128K	**79.74**	**0.67**	**67.60**	**0.46**	**62.80**	**0.27**	**75.46**	**0.57**	**69.58**	**0.44**

reliable annotator, ensuring high-quality synthetic data in long-context scenarios.

Human Evaluation on LongFinanceQA. To further quantify the data quality, we manually annotated 200 randomly selected LongFinanceQA questions, including both single-source and multi-source questions. Then, 10 volunteers score each response on a 1-to-5 scale (1 = completely incorrect, 5 = completely correct). Furthermore, volunteers rate the correctness of intermediate reasoning steps on a 1-to-5 scale for each multi-source question. The averaged scores above 4.0 can be interpreted as largely correct. Table 2 shows that both final answers and their supporting reasoning trajectories are consistently reliable. It provides concrete evidence that the performance improvements primarily arise from high-quality synthesized data, rather than from patterns in noisy samples.

Method Comparisons on Existing Benchmarks. To evaluate the effectiveness of supervised CoT reasoning on long-context modeling, we measure the performance of the enhanced LongPAI model on two well-known long-context understanding benchmarks, including Loong and ∞Bench. First, we evaluate the LongPAI on an in-domain benchmark, namely the *Finance* subset of Loong. Table 3 presents that supervised CoT reasoning enables LongPAI to outperform its base model, LLaMA-3.1-8B-Instruct, by 28.0% in overall results. Furthermore, LongPAI exhibits a 30% improvement on subsets with longer content (*i.e.*, 200K-250K). Also, LongPAI achieves a competitive performance against existing state-of-the-art language models. Remarkably, LongPAI is comparable to its teacher model, GPT-4o-mini w/ PAI (73.94% vs. 75.56%).

Table 2: Human evaluation on LongFinanceQA dataset.

Subset	**Score**
Single-source QA	4.45 ± 0.21
Multi-source QA	4.24 ± 0.23
Intermediate Reasoning Trajectories	4.30 ± 0.22
Overall	4.29 ± 0.19

In some settings, LongPAI even surpasses its teacher model. This finding highlights the significance of long-context modeling. Meanwhile, this finding strongly challenges the recent claim that *the long-context problem can be adequately addressed by short language models* (Qian et al., 2024; Chen et al., 2024b). In other words, certain long-context problems require long-context modeling, as short language models struggle to analyze and reason effectively over reduced or retrieved information. At the same time, we evaluate the LongPAI on ∞Bench. The results in Table 4 show that the LongPAI outperforms its base model even on the out-of-domain benchmark.

4.3 Discussion

Ablation Study on Supervised CoT Reasoning. To further analyze the impact of supervised CoT reasoning, we fine-tune the base LLaMA-3.1 model on LongFinanceQA while excluding CoT reasoning steps from the augmented answers, resulting in a new model, LongPAI§. Unlike the original LongPAI, LongPAI§ directly predicts the final answer, skipping intermediate reasoning steps. Table 5 shows that LongPAI significantly outperforms LongPAI§ over different input lengths. This result strongly supports the hypothesis mentioned in Section 1, which argues that *directly guiding models to generate brief answers without inter-*

Table 3: Performance on *Financial* subset of Loong benchmark. *AS* represents *Avg Scores (0~100)* and *PR* denotes *Perfect Rate* (0~1). Green highlights improvements over the base model. Full results are shown in Appendix A.

Model	**Context Length**	**Spotlight Locating** *AS*	*PR*	**Comparison** *AS*	*PR*	**Clustering** *AS*	*PR*	**Chain of Reasoning** *AS*	*PR*	**Overall** *AS*	*PR*
`All Set` **(10K-250K)**											
Qwen2-72B-Instruct	128K	59.80	0.47	61.12	0.43	34.32	0.06	74.68	0.50	53.20	0.32
Qwen2.5-72B-Instruct	128K	71.07	0.63	59.14	0.41	38.23	0.08	81.09	0.60	57.36	0.36
Qwen-2.5-14B-Instruct-1M	1000K	78.83	0.72	65.27	0.50	36.24	0.07	79.10	0.64	59.78	0.41
GPT-4o	128K	88.23	0.84	62.90	0.48	45.51	0.17	69.40	0.43	63.05	0.44
GPT-4o-mini (Base)	128K	70.90	0.59	59.37	0.38	36.33	0.06	79.58	0.58	56.50	0.34
GPT-4o-mini w/ PAI (*ours*)	128K	91.07	0.83	74.40	0.58	61.55	0.32	89.63	0.78	75.56	0.57
LLaMA-3.1-8B-Instruct (Base)	128K	67.84	0.56	47.12	0.30	24.62	0.02	63.63	0.34	45.88	0.26
LongPAI (*ours*)	262K	89.79	0.84	71.69	0.60	59.71	0.32	90.28	0.83	73.94	0.58
`Set3` **(100K-200K)**											
Qwen2-72B-Instruct	128K	47.00	0.33	48.07	0.27	25.79	0.00	69.37	0.34	42.98	0.20
Qwen2.5-72B-Instruct	128K	60.47	0.48	49.00	0.28	30.61	0.01	76.54	0.46	48.99	0.26
Qwen-2.5-14B-Instruct-1M	1000K	74.33	0.68	54.64	0.35	30.72	0.01	73.71	0.51	53.47	0.33
GPT-4o	128K	87.25	0.83	46.00	0.31	36.68	0.08	64.57	0.40	54.79	0.36
GPT-4o-mini (Base)	128K	63.05	0.53	53.48	0.24	29.80	0.01	72.37	0.46	50.03	0.26
GPT-4o-mini w/ PAI (*ours*)	128K	94.08	0.83	74.13	0.63	55.78	0.20	87.71	0.77	74.21	0.55
LLaMA-3.1-8B-Instruct (Base)	128K	63.38	0.47	36.04	0.19	20.28	0.00	62.49	0.26	40.45	0.20
LongPAI (*ours*)	262K	94.17	0.90	63.76	0.49	51.83	0.24	88.66	0.77	70.00	0.54
`Set4` **(200K-250K)**											
Qwen2-72B-Instruct	128K	41.85	0.19	39.75	0.15	29.17	0.03	41.67	0.07	37.23	0.11
Qwen2.5-72B-Instruct	128K	57.48	0.44	49.50	0.30	27.33	0.00	55.00	0.13	45.51	0.22
Qwen-2.5-14B-Instruct-1M	1000K	59.44	0.41	37.00	0.20	27.33	0.00	31.67	0.00	39.57	0.16
GPT-4o	128K	69.26	0.56	50.50	0.35	30.70	0.00	50.67	0.07	49.58	0.25
GPT-4o-mini (Base)	128K	48.37	0.26	50.00	0.30	28.70	0.00	48.33	0.07	42.30	0.15
GPT-4o-mini w/ PAI (*ours*)	128K	82.78	0.70	63.50	0.35	48.00	0.17	76.00	0.53	66.14	0.42
LLaMA-3.1-8B-Instruct (Base)	128K	40.74	0.30	35.85	0.20	19.77	0.00	28.73	0.00	30.88	0.13
LongPAI (*ours*)	262K	71.48	0.59	56.50	0.45	46.83	0.17	76.00	0.67	60.92	0.43

Table 4: Comparison on En.QA and Zh.QA of ∞Bench. * indicates results borrowed from (Zhang et al., 2024b).

	En.QA	**Zh.QA**
YaRN-Mistra*	9.55	16.98
Kimi-Chat*	16.52	18.62
Claude 2*	11.97	10.53
GPT-4*	22.22	23.06
LLaMA-3.1-8B-Instruct (Base)	27.11	29.77
LongPAI (ours)	**32.65**	**32.88**

mediate reasoning steps for long-context modeling will lead to suboptimal training (**finding 1**). Moreover, Table 5 presents several interesting findings. First, while LongPAI§ performs comparably to LongPAI on short content (10K–50K tokens), its performance declines significantly on longer content, which means *CoT reasoning is necessary for long-context modeling* (**finding 2**). Furthermore, LongPAI§ performs well in single-source questions (Spotlight Locating) but struggles with multi-source questions (Comparison, Clustering, and Chain of Reasoning) as input length increases, demonstrating that *CoT reasoning benefits complex long-context problem-solving* (**finding 3**). In sum, all these findings reaffirm the importance of the reasoning capability for long-context modeling.

Comparison of Various Inference Frameworks. We compare PAI with four relevant inference frameworks: PAI$^-$, Self-Route (Li et al., 2024), Self-Ask (Press et al., 2023), and RAG (Lewis et al., 2020). PAI$^-$ is a variant of PAI that generates subquestions directly, rather than first extracting properties and then generating sub-questions. Table 6 presents that PAI$^-$ generally outperforms the base GPT-4o-mini, except on `Set1`. However, there is a gap between PAI$^-$ and PAI, highlighting the superiority of the Property Extraction Agent. RAG follows a two-step process: it first retrieves the top K chunks relevant to a given query, and then uses the retrieved chunks to generate an answer. Following Loong (Wang et al., 2024), we adopt the *BGE* (Chen et al., 2024a) as the embedding choice and set K to 50, selecting from 5, 10, 30, and 50. Table 6 shows that RAG struggles with long-context problems, a conclusion also reached by previous work (Wang et al., 2024). Furthermore, we compare the proposed PAI framework with both predefined reasoning RAG (Self-Route) and agentic reasoning RAG (Self-Ask). The results demonstrate that PAI consistently outperforms the other two agentic RAG methods, especially on longer

Table 5: Ablation study on *Supervised CoT Reasoning*. Symbol § represents LongPAI without supervised CoT reasoning during fine-tuning. Average Scores (0-100) are evaluated by GPT-4-Turbo. Green highlights the remarkable improvements over the base LLaMA-3.1-8B, while Red indicates a decline. Abbreviations: **S.L.** (Spotlight Locating), **Comp.** (Comparison), **Clust.** (Clustering), and **Chain.** (Chain of Reasoning).

Method	S.L.	Comp.	Clust.	Chain.	Overall
			`Set1` **(10K-50K)**		
LLaMA-3.1-8B	89.13	72.33	31.77	74.0	60.50
LongPAI§	93.39	68.67	79.38	83.80	79.82
LongPAI	97.30	90.17	72.88	94.00	85.42
			`Set2` **(50-100K)**		
LLaMA-3.1-8B	80.58	51.11	27.39	75.12	51.13
LongPAI§	82.12	40.40	11.24	50.25	38.11
LongPAI	91.25	76.27	66.02	96.12	78.19
			`Set3` **(100K-200K)**		
LLaMA-3.1-8B	63.38	36.04	20.28	62.49	40.45
LongPAI§	59.58	30.43	5.47	37.43	29.46
LongPAI	94.17	63.76	51.83	88.66	70.00
			`Set4` **(200K-250K)**		
LLaMA-3.1-8B	40.74	35.85	19.77	28.73	30.88
LongPAI§	43.37	22.00	5.60	17.20	22.14
LongPAI	71.48	56.50	46.83	76.00	60.92

contexts (Set3 and Set4), highlighting a significant superiority of the PAI framework under the scenario of long-context understanding.

Efficiency Analysis on PAI and LongPAI. The PAI framework relies on multi-step inference, whereas LongPAI achieves results with a single inference step. We adopt the number of input tokens processed on the Loong benchmark (*i.e.*, 1,600 samples) as a metric to compare the efficiency between PAI and LongPAI. In this comparison, *PAI processes 3.53B tokens in total, determined by the GPT-4o tokenizer, whereas LongPAI requires only 112M tokens,* ***less than 3% of PAI's total***. Despite PAI delivering the highest overall performance, LongPAI stands out as the far more efficient approach, demonstrating a dramatic reduction in computational cost.

5 Conclusion

In this work, we introduce LongFinanceQA, a novel long-context synthetic dataset featuring reasoning-augmented answers for practical long-context questions. To generate these enriched answers, we develop the Property-based Agentic Inference (PAI) framework, which simulates human-like reasoning through property extraction, retrieval, and summarization. Empirical analysis demonstrates the effectiveness of PAI and ensures the high quality of the LongFinanceQA dataset.

Table 6: Comparison of inference frameworks on the Loong Benchmark. Performance is evaluated by GPT-4-Turbo across four sets with different context sizes: `Set1` (10–50K), `Set2` (50–100K), `Set3` (100–200K), and `Set4` (200–250K). Green indicates an improvement over base GPT-4o-mini. Red denotes a decline.

Method	`Set1`	`Set2`	`Set3`	`Set4`	Overall
GPT-4o-mini	65.05	50.63	45.41	31.61	49.25
w/ PAI	78.69	69.62	68.95	58.12	69.58
w/ PAI^{-}	60.52	57.69	54.89	37.99	54.56
w/ Self-Ask	62.44	52.37	46.91	34.53	50.17
w/ Self-Route	68.79	55.24	49.16	34.44	53.13
w/ RAG	64.67	41.21	31.80	22.08	40.35

Beyond PAI, we explicitly guide a lightweight language model in learning CoT reasoning via fine-tuning on LongFinanceQA. A series of empirical results underscores the importance of reasoning-augmented long-context modeling.

Limitations

In this study, we investigate the effectiveness of supervised CoT reasoning using reasoning-augmented long-context synthetic data (LongFinanceQA). While effective, our approach has certain limitations. First, although the reasoning-enhanced LongPAI model demonstrates significant improvement in the financial domain, its ability to generalize to broader long-context scenarios remains uncertain since overfitting is an inherent issue that stems from supervised fine-tuning (Li et al., 2025). To address this, our future work will explore the impact of diverse data sources and different data scales. Also, we might explore potential techniques to solve this issue, such as regularization (Li et al., 2025) and reinforcement learning (Schulman et al., 2017; Shao et al., 2024; Rafailov et al., 2023) Second, while the PAI framework achieves strong performance on the Loong benchmark, it still relies on domain-specific human-crafted prompts to guide agents in maintaining structured reasoning. For example, prompts may explicitly direct the model to focus on concepts such as "profit, revenue, or cash flow" when processing financial data, or "reference" and "citation" when handling scientific texts. Therefore, we plan to explore inference methods that enable autonomous reasoning with long-context inputs, minimizing reliance on extensive human-crafted prompts.

Despite the current limitations of PAI and LongPAI, their ability to achieve significant performance gains through minimal domain-specific guidance highlights their strong practical potential for specialized applications.

References

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

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. 2023. The falcon series of open language models. *arXiv:2311.16867*.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv:2305.10403*.

Anthropic Team. 2024. Introducing the next generation of claude. https://www.anthropic.com/news/claude-3-family.

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

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 *ACL*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *NeurIPS*.

Keshigeyan Chandrasegaran, Agrim Gupta, Lea M Hadzic, Taran Kota, Jimming He, Cristóbal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Li Fei-Fei. 2024. Hourvideo: 1-hour video-language understanding. In *NeurIPS*.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *ACL*.

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

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv:2107.03374*.

Wenting Chen, Linlin Shen, Jingyang Lin, Jiebo Luo, Xiang Li, and Yixuan Yuan. 2024c. Fine-grained image-text alignment in medical imaging enables explainable cyclic image-report generation. In *ACL*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*.

Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv:2307.08691*.

Tri Dao and Albert Gu. 2024. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. In *ICML*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv:2407.21783*.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv:2404.16130*.

Atty Eleti, Jeff Harris, and Logan Kilpatrick. 2023. Function calling and other api updates.

Yao Fu, Rameswar Panda, Xinyao Niu, Xiang Yue, Hannaneh Hajishirzi, Yoon Kim, and Hao Peng. 2024. Data engineering for scaling language models to 128k context. *arXiv:2402.10171*.

Omer Goldman, Alon Jacovi, Aviv Slobodkin, Aviya Maimon, Ido Dagan, and Reut Tsarfaty. 2024. Is it really long context if all you need is retrieval? towards genuinely difficult long context nlp. *arXiv:2407.00402*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv:2501.12948*.

Junqing He, Kunhao Pan, Xiaoqun Dong, Zhuoyang Song, LiuYiBo LiuYiBo, Qianguosun Qianguosun, Yuxin Liang, Hao Wang, Enming Zhang, and Jiaxing Zhang. 2024. Never lost in the middle: Mastering long-context question answering with position-agnostic decompositional training. In *ACL*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv:2410.21276*.

IBM Research. 2024. What's an llm context window and why is it getting larger? https://research.ibm.com/blog/larger-context-window.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv:2401.04088*.

Tomáš Kočiský, 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*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks.

Zhuowan Li, Cheng Li, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. 2024. Retrieval augmented generation or long-context LLMs? a comprehensive study and hybrid approach. In *EMNLP: Industry Track*.

Ziniu Li, Congliang Chen, Tian Xu, Zeyu Qin, Jiancong Xiao, Zhi-Quan Luo, and Ruoyu Sun. 2025. Preserving diversity in supervised fine-tuning of large language models. In *ICLR*.

Jintao Liang, Gang Su, Huifeng Lin, You Wu, Rui Zhao, and Ziyue Li. 2025. Reasoning rag via system 1 or system 2: A survey on reasoning agentic retrieval-augmented generation for industry challenges. *arXiv:2506.10408*.

Jingyang Lin, Hang Hua, Ming Chen, Yikang Li, Jenhao Hsiao, Chiuman Ho, and Jiebo Luo. 2023. Videoxum: Cross-modal visual and textural summarization of videos. *IEEE Transactions on Multimedia*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024a. Deepseek-v3 technical report. *arXiv:2412.19437*.

Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. 2024b. World model on million-length video and language with blockwise ringattention. *arXiv:2402.08268*.

Hao Liu, Matei Zaharia, and Pieter Abbeel. 2024c. Ringattention with blockwise transformers for near-infinite context. In *ICLR*.

Xiaoran Liu, Hang Yan, Chenxin An, Xipeng Qiu, and Dahua Lin. 2024d. Scaling laws of roPE-based extrapolation. In *ICLR*.

Mistral AI team. 2024a. Large enough. https://mistral.ai/en/news/mistral-large-2407.

Mistral AI team. 2024b. Mistral nemo. https://mistral.ai/en/news/mistral-nemo.

Pritam Mukherjee, Benjamin Hou, Ricardo B Lanfredi, and Ronald M Summers. 2023. Feasibility of using the privacy-preserving large language model vicuna for labeling radiology reports. *Radiology*.

Daye Nam, Andrew Macvean, Vincent Hellendoorn, Bogdan Vasilescu, and Brad Myers. 2024. Using an llm to help with code understanding. In *ICSE*.

Erik Nijkamp, Hiroaki Hayashi, Caiming Xiong, Silvio Savarese, and Yingbo Zhou. 2023a. Codegen2: Lessons for training llms on programming and natural languages. *arXiv:2305.02309*.

Erik Nijkamp, Bo Pang, Hiroaki Hayashi, Lifu Tu, Huan Wang, Yingbo Zhou, Silvio Savarese, and Caiming Xiong. 2023b. Codegen: An open large language model for code with multi-turn program synthesis. In *ICLR*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2024. YaRN: Efficient context window extension of large language models. In *ICLR*.

Sundar Pichai, Demis Hassabis, and Koray Kavukcuoglu. 2024. Introducing gemini 2.0: our new ai model for the agentic era. https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of EMNLP*.

Aman Priyanshu, Yash Maurya, and Zuofei Hong. 2024. Ai governance and accountability: An analysis of anthropic's claude. *arXiv preprint arXiv:2407.01557*.

Hongjin Qian, Zheng Liu, Peitian Zhang, Kelong Mao, Yujia Zhou, Xu Chen, and Zhicheng Dou. 2024. Are long-llms a necessity for long-context tasks? *arXiv:2405.15318*.

Qwen Team. 2024. Introducing qwen1.5. https://qwenlm.github.io/blog/qwen1.5/.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*.

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:2403.05530*.

Jie Ren, Samyam Rajbhandari, Reza Yazdani Aminabadi, Olatunji Ruwase, Shuangyan Yang, Minjia Zhang, Dong Li, and Yuxiong He. 2021. Zero-offload: Democratizing billion-scale model training. In *USENIX ATC*.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, et al. 2022. Multitask prompted training enables zero-shot task generalization. In *ICLR*.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D Manning. 2024. Raptor: Recursive abstractive processing for tree-organized retrieval. In *ICLR*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv:1707.06347*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv:2402.03300*.

Daria Soboleva, Faisal Al-Khateeb, Robert Myers, Jacob R Steeves, Joel Hestness, and Nolan Dey. 2023. SlimPajama: A 627B token cleaned and deduplicated version of RedPajama. https://huggingface.co/datasets/cerebras/SlimPajama-627B.

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

Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, et al. 2023. Video understanding with large language models: A survey. *arXiv preprint arXiv:2312.17432*.

Yi Tay, Jason Wei, Hyung Chung, Vinh Tran, David So, Siamak Shakeri, Xavier Garcia, et al. 2023. Transcending scaling laws with 0.1% extra compute. In *EMNLP*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv:2312.11805*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv: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. 2024. Leave no document behind: Benchmarking long-context LLMs with extended multi-doc QA. In *EMNLP*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

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

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024a. Qwen2 technical report. *arXiv:2407.10671*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024b. Qwen2.5 technical report. *arXiv:2412.15115*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. In *NeurIPS*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *ICLR*.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv:2403.04652*.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. Star: Bootstrapping reasoning with reasoning. In *NeurIPS*.

Peitian Zhang, Ninglu Shao, Zheng Liu, Shitao Xiao, Hongjin Qian, Qiwei Ye, and Zhicheng Dou. 2024a. Extending llama-3's context ten-fold overnight. *arXiv preprint arXiv:2404.19553*.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Hao, Xu Han, Zhen Thai, Shuo Wang, Zhiyuan Liu, et al. 2024b. ∞Bench: Extending long context evaluation beyond 100K tokens. In *ACL*.

Zilin Zhu. 2024. Ring flash attention. https://github.com/zhuzilin/ring-flash-attention.

A Appendix

A.1 Trend of Long-Context LLMs

In this section, we present the context window sizes of closed-source and open-source LLMs in Figure 7. It indicates the gap between closed-source and open-source LLMs becomes smaller. In particular, the LLMs listed in Figure 7 include T5 (Raffel et al., 2020), GPT-3 (Brown et al., 2020), Codex (Chen et al., 2021), T0 (Sanh et al., 2022), Anthropic (Priyanshu et al., 2024), InstructGPT (Ouyang et al., 2022), CodeGen (Nijkamp et al., 2023b), PaLM (Chowdhery et al., 2023), U-PaLM (Tay et al., 2023), Flan-T5 (Chung et al., 2024), GPT-3.5 Turbo, LLaMA (Touvron et al., 2023), GPT-4 (Achiam et al., 2023), Claude 1.3, CodeGen2 (Nijkamp et al., 2023a), PaLM2 (Anil et al., 2023), Claude 2, LLaMA-2 (Touvron et al., 2023), Qwen (Bai et al., 2023), Kimi-Chat-200K, Yi-34B (Young et al., 2024), Falcon 180B (Almazrouei et al., 2023), Gemini Ultra (Team et al., 2023), Mixtral (Jiang et al., 2024), Qwen1.5 (Qwen Team, 2024), Gemini 1.5 Pro, Claude 3.0 (Anthropic Team, 2024), Kimi-Chat-2M, GPT-4 Turbo (Achiam et al., 2023), LLaMA-3 (Dubey et al., 2024), GPT-4o (Hurst et al., 2024), Qwen2 (Yang et al., 2024a), GPT-4o-mini (Hurst et al., 2024), Mistral NeMo (Mistral AI team, 2024b), LLaMA-3.1 (Dubey et al., 2024), Mistral Large 2 (Mistral AI team, 2024a), o1, Qwen-2.5 (Yang et al., 2024b), Gemini 2.0 (Pichai et al., 2024), DeepSeek-V3 (Liu et al., 2024a), DeepSeek-R1 (Guo et al., 2025), Qwen-2.5-1M (Yang et al., 2024b), and o3-mini.

A.2 Prompts of Property-based Agentic Inference (PAI)

In this section, we present the prompts for the three agents in the PAI framework: the property extraction agent, the property-based retrieval agent, and the summarization agent.

Property Extraction Agent. To extract properties, we employ the function-calling API of GPT-4o-mini to selectively retrieve the relevant metric and its corresponding subject from a given query. Figure 4 illustrates the property extraction process using function calling. In addition, we incorporate domain-specific examples within the function call to enhance accuracy across different domains. For instance, in the finance domain, we use “profit”, “revenue”, and “debt” as metric examples, while the “financial document title” serves as a subject example. In the legal domain, “verdict” represents the metric, and “legal judgment” serves as the subject. Similarly, in the academic domain, “reference” and “citation” are used as metric examples, with “paper title” as the subject.

Property-based Retrieval Agent. After obtaining the metric and its corresponding subject, we generate a sub-question in the format: “What was the <metric> of the <subject>?”. Then, the long-context input document is divided into a list of 1024-token chunks. Each chunk is evaluated to determine its relevance to the sub-questions, as shown in Figure 4 (*bottom*). After that, each sub-question is assigned relevant chunks. Next, we pack these relevant chunks and generate a sub-answer to the

Function Calling: Property Extraction

```
{
  "type": "function",
  "function": {
    "name": "Query",
    "strict": true,
    "parameters": {
      "properties": {
        "metric": {
          "description":
             "Specifies the metric for each of the
              given queries. The metric refers to a
              measurable factor being analyzed.
              Examples include {metric_example}. Use
              precise and unambiguous terms that
              clearly represent the desired data
              point.",
          "title": "Metric",
          "type": "string"
        },
        "subject": {
          "description":
            "Indicates the subject towards the metric,
             such as {subject_example}. Ensure the
             subject is clearly defined for context.",
          "title": "Subject",
          "type": "string"
        }
      },
      "required": ["metric", "subject"],
      "title": "Query Parameters",
      "type": "object",
      "additionalProperties": false
    }
  }
}
```

Function Calling: Property-based Retrieval

```
{
  "type": "function",
  "function": {
    "name": "Query",
    "strict": true,
    "parameters": {
      "properties": {
        "keep_context": {
          "description":
             "Does the provided context contain
              information needed to answer the query?
              Answer 'true' or 'false'",
          "title": "Keep Context",
          "type": "boolean"
        },
      },
      "required": ["keep_context"],
      "title": "Query",
      "type": "object",
      "additionalProperties": false
    }
  }
}
```

Figure 4: Function calling details of property extraction agent (*top*) and property-based retrieval (*bottom*).

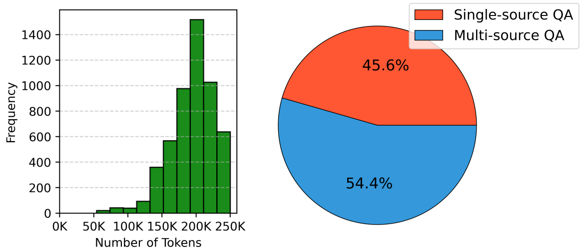

Figure 5: The histogram of input token length in the LongFinanceQA (*left*). The proportion of single-source and multi-source QA tasks in the LongFinanceQA (*right*).

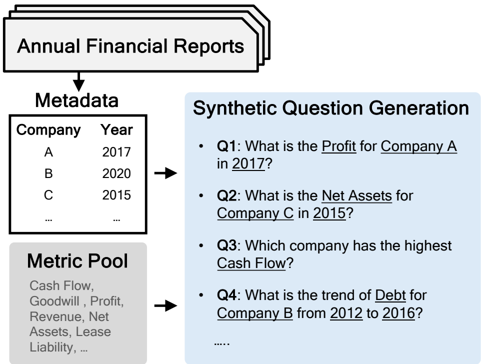

Figure 6: The pipeline of long-context synthetic question generation for the LongFinanaceQA dataset.

corresponding sub-question. Thus, this agent functions similarly to RAG (Lewis et al., 2020).

Summarization Agent. Given the original query, the summarization agent summarizes a conclusion based on the sub-answers generated by the property-based retrieval agent.

A.3 LongFinanceQA Dataset Statistics

Figure 5 presents a histogram of input token lengths in the LongFinanceQA dataset, which generally range from 50K to 250K tokens. Most input documents contain approximately 200K tokens. At the same time, we provide the proportion of single-source and multi-source QA pairs in the proposed LongFinanceQA dataset, where single-source QA pairs make up 45.6%, while multi-source QA pairs account for 54.4%.

A.4 Synthetic Question Generation

Figure 6 illustrates the procedure of synthetic Question generation. First, We create a financial metric pool with key metrics like profit, revenue, and cash flow (represented as the grey box). Then, we randomly select metrics and metadata of financial reports (*e.g.*, company name and year) to generate long-context questions requiring single- or multi-source evidence (represented as the blue box). In addition, the types of multi-source questions are inspired by Loong (Wang et al., 2024) and include comparison, clustering, trend analysis, etc.

A.5 More Empirical Results

Case Study on Practical Long-Context Problems. As shown in Figure 8, we present two representative long-context QA examples. Both questions are particularly challenging, as they require models to consider multiple pieces of evidence from the input text. Specifically, the case study includes the user query (white box), ground truth (yellow box), predictions from the base LLaMA-3.1-8B-Instruct model (Dubey et al., 2024) (white box), and predictions from our LongPAI model (green box). In addition, incorrect predictions are highlighted in red, while correct ones are marked in green. The results demonstrate the superiority of the proposed LongPAI model, which benefits from supervised CoT reasoning. Furthermore, Figure 8 illustrates that LongPAI offers significantly greater interpretability than its base model.

A.6 Full Performances

Table 7 presents the full results of the Table 3 in the main manuscript. Also, we involve more popular LLMs (Yang et al., 2024b; Guo et al., 2025) into the Table 7.

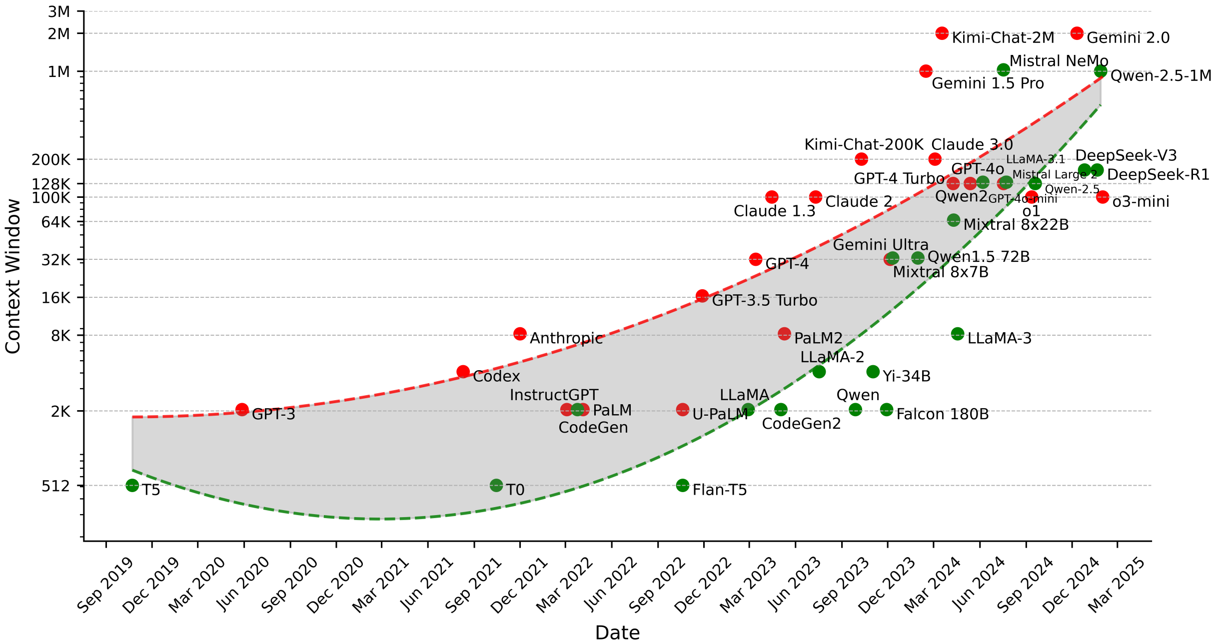

Figure 7: The trend of the context window size of recent large language models. The green dots refer to open-source LLMs, while the red dots denote closed-source LLMs.

User:
Which company has the highest 'Administrative Expenses'?

Ground Truth:
Ault Alliance, Inc.

LLaMA-3.1-8B-Instruct (Base):
Cross Timbers Royalty Trust

LongPAI (*ours*):
This question demands further reasoning:
<reasoning>
This question focuses on the key properties as follows:

- {'metric': 'Administrative Expenses', 'subject': 'CROSS TIMBERS ROYALTY TRUST's 2024 annual report'}
- {'metric': 'Administrative Expenses', 'subject': 'AULT ALLIANCE, INC.'s 2024 annual report'}
- {'metric': 'Administrative Expenses', 'subject': 'CLEARONE, INC.'s 2024 annual report'}

The analysis of the above properties is as follows:

- In CROSS TIMBERS ROYALTY TRUST's 2024 annual report, the Administrative Expenses is $361,500.
- In AULT ALLIANCE, INC.'s 2024 annual report, the Administrative Expenses is $13,380,000.
- In CLEARONE, INC.'s 2024 annual report, the Administrative Expenses is $1,023,000.

The reasoning steps have been completed.
</reasoning>
Conclusion: **AULT ALLIANCE, INC.**

User:
What is the trend observed in the 'Accounts Payable' figures for AMERICAN BATTERY MATERIALS, INC. from 2019 to 2023?

Ground Truth:
The trend observed in the **'Accounts Payable'** figures for **AMERICAN BATTERY MATERIALS, INC**. from 2019 to 2023 is as follows: it **increased** from $278,188 in 2019 to $314,533 in 2020, then slightly **decreased** to $303,248 in 2021, **rose** sharply to $438,667 in 2022, and sharply **decreased** to $164,948 in 2023.

LLaMA-3.1-8B-Instruct (Base):
The trend observed in the '**Accounts Payable**' figures for **AMERICAN BATTERY MATERIALS, INC**. from 2019 to 2023 is a **decrease**. The accounts payable decreased from **$438,667 in 2019** to **$164,948 in 2023**.

LongPAI (*ours*):
This question demands further reasoning:
<reasoning>

- This question focuses on the key properties as follows:
- {'metric': '**Accounts Payable**', 'subject': '**AMERICAN BATTERY MATERIALS, INC.**'s 2020 annual report'}
- {'metric': 'Accounts Payable', 'subject': 'AMERICAN BATTERY MATERIALS, INC.'s 2022 annual report'}
- {'metric': 'Accounts Payable', 'subject': 'AMERICAN BATTERY MATERIALS, INC.'s 2021 annual report'}
- {'metric': 'Accounts Payable', 'subject': 'AMERICAN BATTERY MATERIALS, INC.'s 2023 annual report'}
- {'metric': 'Accounts Payable', 'subject': 'AMERICAN BATTERY MATERIALS, INC.'s 2019 annual report'}

The analysis of the above properties is as follows:

- In AMERICAN BATTERY MATERIALS, INC.'s 2020 annual report, the Accounts Payable is **$314,533**.
- In AMERICAN BATTERY MATERIALS, INC.'s 2022 annual report, the Accounts Payable is **$438,667**.
- In AMERICAN BATTERY MATERIALS, INC.'s 2021 annual report, the Accounts Payable is **$303,248**.
- In AMERICAN BATTERY MATERIALS, INC.'s 2023 annual report, the Accounts Payable is **$164,948**.
- In AMERICAN BATTERY MATERIALS, INC.'s 2019 annual report, the Accounts Payable is **$278,188**.

The reasoning steps have been completed.
</reasoning>
Conclusion: The trend shows an initial **increase** from 2019 to 2020, a **decrease** in 2021, a **significant increase** in 2022, and a **substantial decrease** in 2023.

Figure 8: Case study on two representative long-context QA problems.

Table 7: Full results of Table 3. *AS* represents *Avg Scores (0~100)* and *PR* denotes *Perfect Rate* (0~1). Green indicates improvements compared to the base model; Red denotes a decrease compared to the base model.

Model	Context	Spotlight Locating		Comparison		Clustering		Chain of Reasoning		Overall	
	Length	*AS*	*PR*	*AS*	*PR*	*AS*	*PR*	*AS*	*PR*	*AS*	*PR*
All Set (10K-250K)											
DeepSeek-R1-Qwen-32B	128K	53.66	0.46	52.19	0.39	39.76	0.17	65.15	0.51	49.92	0.34
Qwen2-72B-Instruct	128K	59.80	0.47	61.12	0.43	34.32	0.06	74.68	0.50	53.20	0.32
Qwen2.5-72B-Instruct	128K	71.07	0.63	59.14	0.41	38.23	0.08	81.09	0.60	57.36	0.36
LLaMA-3-8B-Instruct-262K	262K	58.60	0.41	33.12	0.16	20.04	0.01	35.10	0.09	34.41	0.15
GLM4-9B-Chat	1000K	72.69	0.60	49.31	0.32	23.41	0.02	60.77	0.28	46.71	0.27
Qwen-2.5-14B-Instruct-1M	1000K	78.83	0.72	65.27	0.50	36.24	0.07	79.10	0.64	59.78	0.41
GPT-4o	128K	88.23	0.84	62.90	0.48	45.51	0.17	69.40	0.43	63.05	0.44
GPT-4o-mini (Base)	128K	70.90	0.59	59.37	0.38	36.33	0.06	79.58	0.58	56.50	0.34
GPT-4o-mini w/ PAI (*ours*)	128K	91.07	0.83	74.40	0.58	61.55	0.32	89.63	0.78	75.56	0.57
LLaMA-3.1-8B-Instruct (Base)	128K	67.84	0.56	47.12	0.30	24.62	0.02	63.63	0.34	45.88	0.26
LongPAI (*ours*)	262K	89.79	0.84	71.69	0.60	59.71	0.32	90.28	0.83	73.94	0.58
Set1 (10K-50K)											
DeepSeek-R1-Qwen-32B	128K	41.73	0.33	39.56	0.23	25.67	0.06	55.14	0.37	37.35	0.21
Qwen2-72B-Instruct	128K	88.04	0.83	89.33	0.83	43.00	0.17	93.50	0.80	71.46	0.57
Qwen2.5-72B-Instruct	128K	88.70	0.87	84.67	0.80	43.92	0.10	88.00	0.80	70.07	0.54
LLaMA-3-8B-Instruct-262K	262K	75.82	0.64	40.83	0.23	20.68	0.03	68.00	0.40	43.14	0.25
GLM4-9B-Chat	1000K	88.26	0.83	73.17	0.57	24.65	0.00	87.30	0.50	59.07	0.40
Qwen-2.5-14B-Instruct-1M	1000K	96.09	0.96	88.67	0.80	49.00	0.20	100.00	1.00	76.02	0.62
GPT-4o	128K	100.00	1.00	88.50	0.80	55.25	0.28	98.50	0.90	79.13	0.65
GPT-4o-mini (Base)	128K	97.39	0.96	81.83	0.67	46.50	0.15	100.00	1.00	73.35	0.56
GPT-4o-mini w/ PAI (*ours*)	128K	95.87	0.91	91.50	0.83	78.38	0.60	96.80	0.70	87.89	0.75
LLaMA-3.1-8B-Instruct (Base)	128K	89.13	0.87	72.33	0.60	31.77	0.05	74.00	0.60	60.50	0.45
LongPAI (*ours*)	262K	97.30	0.91	90.17	0.80	72.88	0.47	94.00	0.90	85.42	0.71
Set2 (50K-100K)											
DeepSeek-R1-Qwen-32B	128K	41.73	0.33	39.56	0.23	25.67	0.06	55.14	0.37	37.35	0.21
Qwen2-72B-Instruct	128K	74.88	0.65	68.60	0.51	40.70	0.09	87.00	0.72	62.38	0.41
Qwen2.5-72B-Instruct	128K	86.00	0.82	61.64	0.41	46.94	0.16	93.12	0.85	65.36	0.46
LLaMA-3-8B-Instruct-262K	262K	72.44	0.54	44.03	0.26	24.11	0.02	37.25	0.12	40.40	0.20
GLM4-9B-Chat	1000K	82.12	0.68	52.73	0.36	26.04	0.04	76.83	0.42	51.66	0.31
Qwen-2.5-14B-Instruct-1M	1000K	88.75	0.85	74.07	0.61	39.06	0.09	96.38	0.90	67.24	0.51
GPT-4o	128K	95.75	0.95	72.87	0.57	54.94	0.27	73.38	0.47	70.10	0.51
GPT-4o-mini (Base)	128K	82.65	0.70	58.77	0.41	40.89	0.08	92.50	0.78	61.61	0.40
GPT-4o-mini w/ PAI (*ours*)	128K	89.38	0.85	70.73	0.51	64.37	0.36	94.62	0.90	75.34	0.57
LLaMA-3.1-8B-Instruct (Base)	128K	80.58	0.70	51.11	0.33	27.39	0.02	75.12	0.47	51.13	0.30
LongPAI (*ours*)	262K	91.25	0.88	76.27	0.67	66.02	0.37	96.12	0.93	78.19	0.63
Set3 (100K-200K)											
DeepSeek-R1-Qwen-32B	128K	41.73	0.33	39.56	0.23	25.67	0.06	55.14	0.37	37.35	0.21
Qwen2-72B-Instruct	128K	47.00	0.33	48.07	0.27	25.79	0.00	69.37	0.34	42.98	0.20
Qwen2.5-72B-Instruct	128K	60.47	0.48	49.00	0.28	30.61	0.01	76.54	0.46	48.99	0.26
LLaMA-3-8B-Instruct-262K	262K	54.14	0.40	22.93	0.05	15.43	0.00	29.00	0.00	28.58	0.11
GLM4-9B-Chat	1000K	74.75	0.65	41.63	0.24	21.99	0.01	49.86	0.17	43.58	0.25
Qwen-2.5-14B-Instruct-1M	1000K	74.33	0.68	54.64	0.35	30.72	0.01	73.71	0.51	53.47	0.33
GPT-4o	128K	87.25	0.83	46.00	0.31	36.68	0.08	64.57	0.40	54.79	0.36
GPT-4o-mini (Base)	128K	63.05	0.53	53.48	0.24	29.80	0.01	72.37	0.46	50.03	0.26
GPT-4o-mini w/ PAI (*ours*)	128K	94.08	0.83	74.13	0.63	55.78	0.20	87.71	0.77	74.21	0.55
LLaMA-3.1-8B-Instruct (Base)	128K	63.38	0.47	36.04	0.19	20.28	0.00	62.49	0.26	40.45	0.20
LongPAI (*ours*)	262K	94.17	0.90	63.76	0.49	51.83	0.24	88.66	0.77	70.00	0.54
Set4 (200K-250K)											
DeepSeek-R1-Qwen-32B	128K	18.33	0.11	10.25	0.05	11.27	0.00	26.00	0.07	15.52	0.05
Qwen2-72B-Instruct	128K	41.85	0.19	39.75	0.15	29.17	0.03	41.67	0.07	37.23	0.11
Qwen2.5-72B-Instruct	128K	57.48	0.44	49.50	0.30	27.33	0.00	55.00	0.13	45.51	0.22
LLaMA-3-8B-Instruct-262K	262K	34.19	0.07	20.00	0.05	20.31	0.00	20.71	0.00	24.61	0.03
GLM4-9B-Chat	1000K	40.85	0.19	29.50	0.05	18.13	0.00	25.73	0.00	28.51	0.07
Qwen-2.5-14B-Instruct-1M	1000K	59.44	0.41	37.00	0.20	27.33	0.00	31.67	0.00	39.57	0.16
GPT-4o	128K	69.26	0.56	50.50	0.35	30.70	0.00	50.67	0.07	49.58	0.25
GPT-4o-mini (Base)	128K	48.37	0.26	50.00	0.30	28.70	0.00	48.33	0.07	42.30	0.15
GPT-4o-mini w/ PAI (*ours*)	128K	82.78	0.70	63.50	0.35	48.00	0.17	76.00	0.53	66.14	0.42
LLaMA-3.1-8B-Instruct (Base)	128K	40.74	0.30	35.85	0.20	19.77	0.00	28.73	0.00	30.88	0.13
LongPAI (*ours*)	262K	71.48	0.59	56.50	0.45	46.83	0.17	76.00	0.67	60.92	0.43