NeedleBench: Can LLMs Do Retrieval and Reasoning in Information-Dense Context?

Anonymous authors
Paper under double-blind review

Abstract

The capability of large language models to handle long-context information plays a crucial role across various real-world applications. Existing methods for evaluating long-context abilities often rely either on real-world long texts, making it difficult to exclude the influence of models' inherent knowledge, or introduce large amounts of irrelevant filler content to artificially reach target lengths, reducing the relevance and effectiveness of assessments. To address these limitations, we introduce NeedleBench, a comprehensive synthetic framework designed to assess retrieval and reasoning performance in bilingual long-context tasks with adaptive context lengths (e.g., 32k, 128k, and beyond). NeedleBench systematically embeds key data points at varying depths to rigorously test models' capabilities in diverse settings. Tasks within NeedleBench are categorized into two distinct scenarios: information-sparse, characterized by minimal relevant details embedded within extensive irrelevant text to simulate simpler real-world retrieval tasks; and information-dense, implemented as the Ancestral Trace Challenge, where relevant information is continuously distributed throughout the context to simulate more complex real-world reasoning tasks. Our experiments show that, while recent reasoning models such as Deepseek-R1 and OpenAI's o3 have demonstrated strong performance on mathematical reasoning benchmarks, they still struggle to generalize their reasoning abilities and perform poorly on our information-dense tasks, frequently encountering difficulties with continuous retrieval and reasoning even at relatively shorter context lengths. Furthermore, we identify and characterize a phenomenon termed 'under-thinking', wherein models prematurely conclude their reasoning processes despite the availability of relevant information. NeedleBench thus provides critical insights and targeted evaluation tools essential for understanding and improving the long-context capabilities of LLMs. All codes and resources will be publicly available.

1 Introduction

The capability of LLMs to process long texts is particularly crucial across various situations (Mohtashami & Jaggi, 2023; Grattafiori et al., 2024; Yang et al., 2025; Team et al., 2024; Tunstall et al., 2023a; Team, 2025; Yang et al., 2024b; DeepSeek-AI, 2025; Lyu et al., 2025). LLMs can rapidly identify and summarize relevant information within lengthy documents, making them invaluable for legal document retrieval, academic research, and aggregating business intelligence, among other applications (Wang et al., 2024; Lee et al., 2024). To meet these needs, modern close-sourced LLMs have recently been developed to support longer context windows (OpenAI, 2023; Anthropic, 2024b; Gemini Team, 2024; Cai et al., 2024; GLM et al., 2024; Bai et al., 2023a). As models accommodate longer text lengths, verifying their comprehension of details within the text becomes increasingly essential.

A variety of approaches have been proposed to evaluate the long-context capabilities of LLMs, though assessing performance at extremely long contexts (e.g., around 1M tokens) remains challenging. Early methods embed crucial "passkeys" in repetitively structured texts to test information retrieval over long sequences (Mohtashami & Jaggi, 2023; Zhang et al., 2023). Building on this idea, the Needle In A Haystack (NIAH) test (Kamradt, 2023) introduces more realistic settings by using non-repetitive personal essays as filler material, increasing task complexity and extending context lengths up to 200K tokens. According to the results reported in Kamradt (2023), advanced models such as Claude 2.1 and GPT-4 Turbo generally perform well on these targeted extraction tasks (Anthropic, 2024b).

More recent benchmarks like LongBench v2 (Bai et al., 2024; 2023b) offer diverse comprehension tasks but typically fix task length and lack adaptability, while Ruler (Hsieh et al., 2024) attempts adaptive-length evaluations by inserting irrelevant text. However, inserting too much irrelevant content can make it easy for models to complete tasks by focusing only on a few key points, without truly reading the full context. This may fail to reflect how models perform in more information-dense tasks that require careful reading and integration of content. For instance, in legal case analysis, an LLM must extract relevant facts and legal provisions from case files and synthesize them to answer specific questions. In such cases, **even small details in the long context can affect the final judgment**, making the input highly information-dense with little room for irrelevant content. Furthermore, benchmarks derived from real-world texts often encounter the issue of models leveraging prior knowledge acquired during pre-training, thereby undermining a true assessment of long-context understanding.

Given these limitations, it is crucial to design benchmarks that not only support adaptive context lengths, but also **feature information-dense tasks** where relevant content is distributed throughout the input. Such tasks should minimize reliance on irrelevant filler used in Kamradt (2023); Hsieh et al. (2024), ensuring that models must engage with the entire context to perform well. This helps better evaluate a model's true capacity for long-context understanding, while also avoiding scenarios where models can simply rely on memorized knowledge from pre-training instead of actually processing the input text (Bai et al., 2023b; 2024).

To address the limitations of existing long-context evaluation methods, we present *NeedleBench*, a dataset framework that encompasses both information-sparse and information-dense tasks. *NeedleBench* is designed to provide a comprehensive, targeted assessment of models' abilities to extract, analyze, and reason over long texts. In particular, our benchmark supports flexible context lengths (4k, 8k, 32k, 128k, 200k, 1000k, and beyond), allowing strategic insertion of key data points at various depths to rigorously test retrieval and reasoning skills. Moreover, these tasks are largely synthetic, which helps mitigate the influence of prior internal knowledge and compels models to truly process the given context.

Within *NeedleBench*, we include tasks that continue the tradition of inserting irrelevant filler (e.g., Single-Needle Retrieval, Multi-Needle Retrieval, Multi-Needle Reasoning), providing a information-sparse baseline for evaluating how well models handle straightforward retrieval in extended texts. On the other hand, we propose the Ancestral Trace Challenge (ATC), an information-dense task designed to reflect more complex real-world scenarios that require continuous logical reasoning. Our findings reveal that, despite recent top-performing models such as OpenAI's o3(OpenAI, 2025b) and DeepSeek-R1 (DeepSeek-AI, 2025) achieving impressive results on mathematical reasoning benchmarks such as AIME (Di Zhang, 2025) and MATH500 (Hendrycks et al., 2021), they still struggle to **generalize their reasoning abilities** and perform poorly on our information-dense tasks. Our major contributions are as follows:

- **Comprehensive Bilingual Long-Context Benchmark:** We introduce *NeedleBench*, a customizable framework for evaluating bilingual long-context capabilities of LLMs across multiple length intervals, covering both information-sparse and information-dense tasks.
- **Long-Context Information-Dense Task:** We design the Ancestral Trace Challenge, simulating real-world information-dense tasks where models must track interdependent entities and constraints across evolving contexts. Our experiments demonstrate that current LLMs still struggle with complex long-context tasks.
- **Fine-Grained Evaluation and Analysis:** We offer an in-depth assessment of mainstream models' retrieval and reasoning performance under different context conditions. All reproducible scripts, code, and datasets will be made available upon publication.

2 Related Work

Long-Context Language Models. Recent large language models have rapidly increased their context window sizes, from early 4K-token models like Longformer and BigBird (Beltagy et al., 2020; Zaheer et al.,

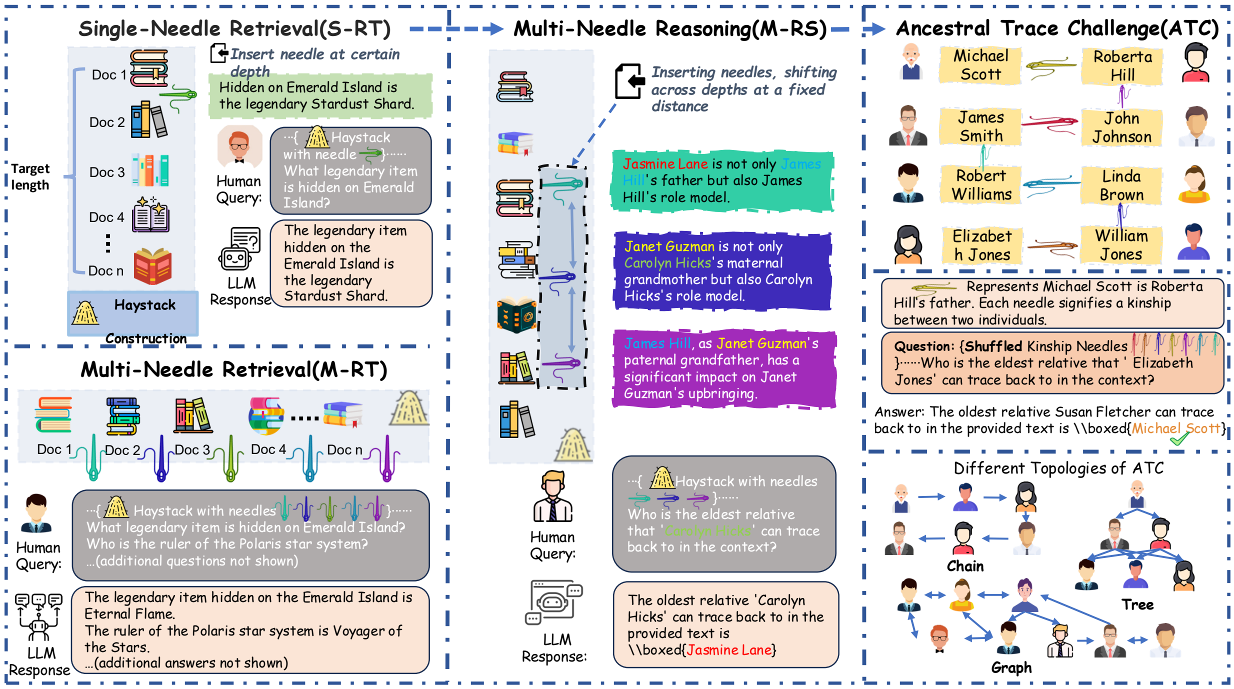


Figure 1: *NeedleBench* Framework. Our benchmark consists of two main categories: Information-Sparse Tasks (left two columns), which include Single-Needle Retrieval, Multi-Needle Retrieval, and Multi-Needle Reasoning with irrelevant filler content; and Information-Dense Tasks (rightmost column), specifically the Ancestral Trace Challenge, designed to eliminate irrelevant filler and require comprehensive understanding of all content.

2021) to commercial models such as GPT-4 Turbo (128K), Claude 3 (200K), and Gemini 1.5 Pro (1M) (OpenAI, 2023; Anthropic, 2024a; Gemini Team, 2024). While many models achieve near-perfect scores on NIAH test, these results do not guarantee strong reasoning or comprehension in information-dense settings. Recent work (Hsieh et al., 2024) shows that passing retrieval tests does not always mean robust understanding, underscoring the need for more challenging benchmarks to truly assess long-context reasoning.

Long-Context benchmarks. Existing long-context benchmarks mainly fall into two categories: those using repetitive or irrelevant filler to test retrieval over long sequences (Mohtashami & Jaggi, 2023; Zhang et al., 2023; Kamradt, 2023; Hsieh et al., 2024), and those based on real-world texts (Bai et al., 2024; 2023b). While methods like the Needle In A Haystack(NIAH) test (Kamradt, 2023) and Ruler (Hsieh et al., 2024) extend context length, they often rely on information-sparse settings, making it possible for models to succeed by focusing on a few key points. Real-world benchmarks, on the other hand, may be affected by models' prior knowledge (Bai et al., 2023b; 2024). There remains a need for adaptive, information-dense benchmarks that require models to process and reason over all content, minimizing irrelevant filler and prior knowledge effects.

3 Tasks and Datasets

We categorize *NeedleBench* tasks into two types: **Information-Sparse Tasks**, where only a small fraction of the input text contains relevant information for answering the question; and **Information-Dense Tasks**, where every sentence contains essential content and the model must fully comprehend all details to succeed. The overall structure of these tasks are illustrated in Fig. 1.

3.1 *NeedleBench* Information-Sparse Tasks

The information-sparse tasks in *NeedleBench* is composed of the following three tasks, each targeting a specific capability of long-context processing:

- **Single-Needle Retrieval Task (S-RT):** Tests LLMs' ability to recall a **single key** information inserted at various positions in a long text, highlighting their precision in navigating and recalling single detail within extensive texts.
- **Multi-Needle Retrieval Task (M-RT):** Explores LLMs' ability to retrieve **multiple pieces** of related information scattered across a lengthy text, simulating complex real-world queries that require extracting several data points from comprehensive documents.
- **Multi-Needle Reasoning Task (M-RS):** Evaluates LLMs' ability for complex reasoning by extracting **multiple pieces** of information (ranging from 2 to 5 key facts) from long texts and using them to logically answer questions that demand an integrated understanding and reasoning of various text segments.

In *NeedleBench* tasks, both the "needles" (key information units) and the "haystack" (background or filler content) are carefully constructed to ensure a fair and challenging evaluation. The needles are **synthetic, abstract, and fictional** statements or relational facts, deliberately designed to avoid overlap with any real-world knowledge or pretraining data. For retrieval tasks, these may be unique fabricated facts (e.g., "Hidden on Emerald Island is the legendary Stardust Shard"), while for reasoning tasks, they are synthetic kinship needles, which are the same as those used in the information-dense task; see Sec. 3.2 for details. The haystack for English tasks is built by extending the prompt with passages from the PaulGrahamEssays dataset (Kamradt, 2023), and for Chinese tasks, we use the ChineseDomainModelingEval dataset (Wei et al., 2023) to ensure linguistic diversity and high-quality filler content. This design ensures that models cannot rely on memorized knowledge and must genuinely process the provided context to succeed.

Evaluation Metrics. To quantitatively evaluate model performance on the information-sparse tasks, we adopt a keyword-aware scoring approach that emphasizes the successful retrieval of core information from long texts. For each instance, a predefined set of core keywords is used to determine whether the model prediction sufficiently captures the essential content.

We employ a keyword-aware variant of the Exact Match (EM) metric in Eq. (1). For each test, a predefined set of core keywords $W_{c,d,r}$ is compared with the model's prediction $P_{c,d,r}$, where c represents the context length, d represents the needle depth or position, and r represents the index of repetition. Full credit is awarded if any keyword from $W_{c,d,r}$ appears in $P_{c,d,r}$, and zero otherwise:

$$\text{Score}_{c,d,r} = \begin{cases} 100, & \text{if } P_{c,d,r} \cap W_{c,d,r} \neq \emptyset, \\ \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

For each task, we first compute a task-specific score by averaging across different experimental dimensions:

$$\text{Task Score} = \frac{1}{|C| \cdot |D| \cdot R} \sum_{c \in C} \sum_{d \in D} \sum_{r=1}^{R} \text{Score}_{c,d,r} \tag{2}$$

where C represents the set of context lengths, D represents the set of needle depths or positions, and R is the number of repetitions for each configuration. The overall benchmark score is then calculated as a weighted average across all information-sparse tasks:

$$\text{Overall Score} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \text{Task Score}_t, \quad \mathcal{T} = \{\text{S-RT}, \text{M-RT}, \text{M-RS}\} \tag{3}$$

By evaluating performance across different context lengths, needle positions, and task types, *NeedleBench* provides a detailed assessment of a model's long-context abilities. For each configuration, we repeat the test $R = 10$ times to enhance result stability. Token lengths are measured using the GPT-4 tokenizer[1]. To reduce the risk of instruction truncation caused by tokenizer discrepancies between models, we subtract a small buffer from the target context length for each input, thereby minimizing the impact of tokenizer differences.

3.2 *NeedleBench* Information-Dense Tasks

Unlike information-sparse tasks that often include large portions of irrelevant filler content, the Ancestral Trace Challenge is explicitly designed to be information-dense: every sentence in the input context contains critical information directly related to the target question. There is no irrelevant text—each piece of content is essential and contributes to determining the correct answer. Each question in ATC is constructed by embedding multiple interlinked fact units ("needles") throughout a long document. These facts are interlinked in such a way that the model must accurately track and retain all of them to complete the final blank in the question(See Fig. 12). Any missing or misinterpreted fact will lead to an incorrect output. Notably, despite the fact that these tasks are conceptually simple—requiring only repeated retrieval and reasoning rather than complex logic—the challenge arises from the need to maintain perfect recall and integration of all information over long contexts. Additionally, to comprehensively evaluate long-context reasoning, ATC introduces diversity along several axes:

- **Name Diversity:** Names are randomized for each critical information, ensuring that every question features unique and unpredictable entities. This prevents memorization and encourages genuine reasoning.
- **Relationship Diversity:** The task features a wide variety of relationship types, including but not limited to parent-child, ancestor-descendant (across multiple generations), and dual-role relationships (e.g., one individual may simultaneously be a parent and a lifelong mentor). This diversity requires models to handle a broad range of relational structures.
- **Task Diversity:** ATC includes multiple question types, which can be categorized as follows: (1) identifying the eldest ancestor; (2) tracing the n-th ancestor of a given individual; (3) tracing the n-th descendant of a given individual; (4) calculating the relationship distance between two individuals. This variety ensures that models are evaluated on a broad spectrum of reasoning skills.
- **Logical Complexity and Context Length Diversity:** The number of needles per question is varied from 2 to 512, increasing both logical complexity and context length, requiring the model to continuously integrate information from multiple sources in the context.
- **Language Diversity:** ATC supports bilingual evaluation in both English and Chinese, ensuring that models are tested on their ability to process and reason over information-dense contexts in multiple languages.

Evaluation Metrics. For the information-dense ATC task, we use an exact match (EM) metric based on the model's ability to output the correct answer in the required format `\boxed{...}`, allowing for precise and automated evaluation. To aggregate results across different levels of task difficulty (i.e., different numbers of embedded needles), we compute a weighted average of the exact match accuracy, where the weight for each subtask is proportional to the number of needles it contains. This is similar to the weighted average metric in classification evaluation (Pedregosa et al., 2011), where each class is weighted by its support (number of instances). For each configuration, we repeat the evaluation $R = 10$ times to ensure stability.

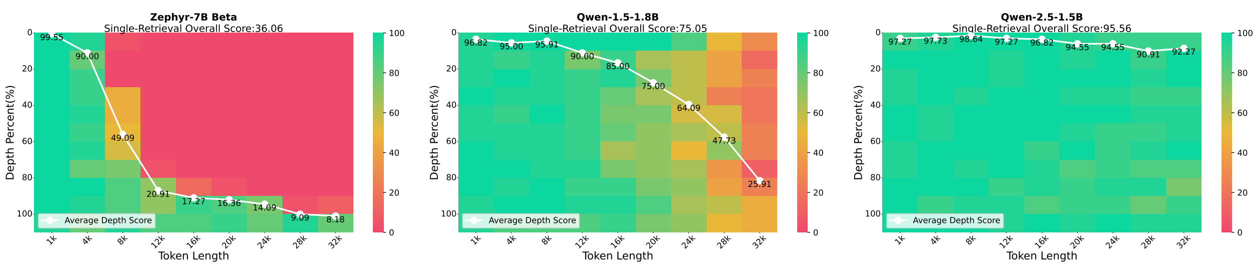
Formally, let P_{2^k} denote the exact-match accuracy (in percentage) of a model on the ATC subtask with 2^k needles, and let $\mathcal{N} = \{2^k \mid k = 1, 2, \ldots, 9\}$ be the set of needle counts. We report following two metrics:

$$\text{Weighted Average} = \frac{\sum_{N \in \mathcal{N}} (P_N \times N)}{\sum_{N \in \mathcal{N}} N} \qquad \text{ENL}_\tau = \max\{N \in \mathcal{N} \mid P_N \geq \tau\} \tag{4}$$

[1] `https://github.com/openai/tiktoken`

Table 1: **Main Results of *NeedleBench* 32K.** Overall denotes the mean score across all tasks. **Bold** denotes the best score among all models, and underline denotes the best score under the same model scale. The same notation applies in the following tables.

Model	Single-Retrieval			Multi-Retrieval			Multi-Reasoning			Overall
	Chinese	English	Overall	Chinese	English	Overall	Chinese	English	Overall	
Models with Fewer Than 7B Parameters										
Qwen-2.5-1.5B	96.67	94.44	95.56	95.39	97.29	96.34	0.00	15.63	7.82	66.57
Qwen-1.5-4B	95.66	99.60	97.63	95.76	97.01	96.38	2.68	7.05	4.86	66.29
ChatGLM3-6B-32K	93.64	98.89	96.26	90.83	94.38	92.61	0.18	9.07	4.62	64.50
Qwen-1.5-1.8B	78.99	71.11	75.05	54.26	52.93	53.60	0.00	0.00	0.00	42.88
Models with 7-20B Parameters										
Qwen-2.5-14B	99.19	98.79	98.99	**99.07**	99.23	99.15	29.65	17.90	23.78	73.97
Qwen-2.5-7B	**100.00**	99.80	99.90	97.70	99.31	98.51	12.65	18.64	15.64	71.35
Qwen-1.5-14B	99.60	99.49	99.55	92.57	99.15	95.86	0.58	10.20	5.39	66.93
Mistral-7B-Instruct-v0.2	92.73	96.36	94.55	87.23	96.97	92.10	11.57	14.27	12.92	66.52
Zephyr-7B-Beta	35.35	36.77	36.06	18.14	27.60	22.87	1.87	7.45	4.66	21.20
Models Larger Than 20B Parameters										
Qwen-2.5-72B	**100.00**	**100.00**	**100.00**	98.71	**99.96**	**99.33**	**39.80**	**52.07**	**45.93**	**81.76**
Qwen-2.5-32B	**100.00**	**100.00**	**100.00**	98.71	95.72	97.21	33.31	38.96	36.14	77.78
Qwen-1.5-32B	99.60	**100.00**	99.80	98.02	98.95	98.48	11.67	14.85	13.26	70.51
Mixtral-8x7B-Instruct-v0.1	95.76	99.60	97.68	94.63	99.43	97.03	5.93	15.88	10.91	68.54
Qwen-1.5-72B	97.37	89.60	93.48	93.49	92.24	92.87	9.75	7.35	8.55	64.97



(a) Zephyr-7B-Beta (b) Qwen-1.5-1.8B (c) Qwen-2.5-1.5B

Figure 2: Comparison of different model generations on the Single-Retrieval Task. Newer models such as Qwen-2.5 show clear improvements over earlier models like Zephyr-7B-Beta and Qwen-1.5.

where τ is a threshold (we use $\tau = 50\%$ and denote the metric as ENL-50). The ENL metric (Effective Needle Length) reflects the largest number of needles N for which the model's exact-match accuracy P_N remains at least τ. In other words, ENL-50 measures the model's effective reasoning depth: the maximum task difficulty (needle count) at which the model can still achieve at least 50% accuracy.

4 Experiments

We evaluate mainstream open-source LLMs on the information-sparse tasks in *NeedleBench* at two representative context lengths: 32K and 128K tokens. Each model is tested at the maximum context length it officially supports. To enable direct comparison across model generations, we also include Qwen-2.5 models in the 32K setting, even though they support longer contexts. All evaluated models are instruction-tuned rather than base models. Due to the high computational cost of long-context evaluation for closed-source API models, our evaluation of information-sparse tasks is limited to open-source models. Results are shown in Tabs. 1 and 2. The full list of evaluated models and their context window sizes is provided in Appendix A. For the information-dense ATC task, which is designed as a long-context evaluation with high information density and minimal irrelevant content, we expand our evaluation to include leading close-sourced API models such as GPT-4.1 (OpenAI, 2025a), OpenAI's o3-mini (OpenAI, 2025b), Claude-3.7-Sonnet-Thinking (Anthropic, 2024a), and DeepSeek R1 (DeepSeek-AI, 2025).

4.1 Performance of *NeedleBench* Information-Sparse Tasks

4.1.1 Impact of Model Architecture and Technical Advances on Retrieval Performance

In the 32K context setting, we observe a pronounced difference in retrieval performance not only between older and newer model generations, but also across different architectural and technical choices. Early-generation models such as Zephyr-7B-Beta (Tunstall et al., 2023b) and Qwen-1.5-1.8B (Bai et al., 2023a) achieve relatively low overall scores on the Single-Retrieval task, and often fail to achieve perfect recall—especially when the relevant information is located far from the end of the context, making it difficult for the model to maintain and utilize that information. See Fig. 2a and Fig. 2b for their performance.

This phenomenon can be partially attributed to the underlying architectural and technical designs of these models. Zephyr-7B-Beta, for example, employs sliding window attention (SWA) (Beltagy et al., 2020), which restricts attention to local windows and may hinder the propagation of information across distant tokens. Qwen-1.5 adopts a mixture of sliding window and full attention, which offers some improvement but still falls short of optimal long-range retrieval. In contrast, newer models such as Qwen-2.5 (Yang et al., 2024a) leverage advanced techniques for long-context extrapolation, including Dual Chunk Attention (DCA) (An et al., 2024) and YaRN (Peng et al., 2023), which enable effective modeling of both local and global dependencies. In practice, even at the small 1.5B scale, Qwen-2.5 achieves near-perfect or perfect scores on retrieval tasks (see Fig. 2c), indicating that these new architectural techniques can substantially enhance retrieval ability in long-context settings, and that accurate long-context retrieval is now commonly observed in modern LLMs.

It is also important to note that the use of local attention mechanisms such as SWA does not necessarily hinder strong long-context retrieval. For example, the recent Gemma-3 (Team et al., 2025) models employ a hybrid attention strategy, interleaving local and global attention layers at a 5:1 ratio (compared to the 1:1 ratio in Gemma-2). Despite having a higher proportion of local attention, Gemma-3 still achieves strong performance at very long context lengths, as demonstrated by the 128K context results (Tab. 2), suggesting carefully balancing the ratio of local and global attention layers is crucial for optimizing long-context performance.

4.1.2 Challenges in Multi-Needle Reasoning Compared to Retrieval Tasks

While retrieval tasks, such as Single-Needle Retrieval and Multi-Needle Retrieval, have become relatively straightforward for modern LLMs, the Multi-Needle Reasoning task presents a far greater challenge in Tab. 1. This task requires models to extract key information from multiple locations within a long context and integrate these interdependent pieces to perform complex reasoning. Only a few large-scale models—such as Qwen-2.5-72B and Qwen-2.5-32B—demonstrate significantly higher scores (with Qwen-2.5-72B achieving the best result at 45.93%, still below 50%), indicating their superior ability to handle such reasoning demands. In contrast, smaller models or earlier generations, particularly those below 20B parameters, struggle to perform effectively, often achieving near-zero scores. This highlights the inherent difficulty of reasoning tasks that demand multi-point integration over densely packed, interrelated information within long sequences.

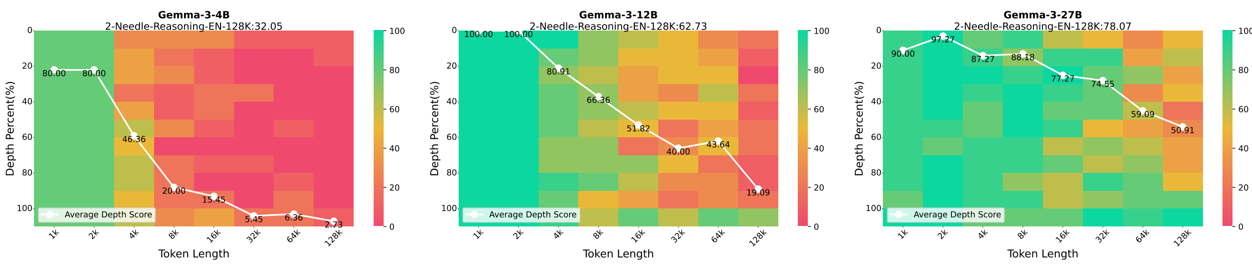
When scaling to 128K context length, we find that retrieval remains a largely solved problem for modern LLMs. Most models that support this length are relatively recent, and as such, exhibit strong performance on both Single and Multi-Needle Retrieval tasks, as shown in Tab. 2. Similar to the trends observed in the 32K setting (Tab. 1), Multi-Needle Reasoning continues to reveal substantial performance gaps across models. Notably, InternLM3-8B stands out among sub-10B models, achieving reasoning performance comparable to mid-sized models like Qwen-2.5-14B, though it still trails behind the strongest models in the 20B+ range.

4.1.3 Effect of Model Scale on Multi-Needle Reasoning Performance

To provide a clear visualization of the impact of model parameter size on Multi-Needle Reasoning, we present the performance of the Gemma-3 series (4B, 12B, and 27B) on the 2-Needle Reasoning task at 128K context length. As shown in Fig. 3, as the model size increases, the color in the figure gradually shifts to green, indicating higher scores and stronger reasoning ability for larger models in integrating multiple information points within long contexts.

Table 2: **Main Results of *NeedleBench* 128K.** Most models achieve excellent performance on retrieval tasks, indicating that they can easily handle long-context retrieval. However, there is still significant room for improvement on Multi-Needle Reasoning tasks, with even the best model scoring below 50.

Model	Single-Retrieval			Multi-Retrieval			Multi-Reasoning			Overall
	Chinese	English	Overall	Chinese	English	Overall	Chinese	English	Overall	
Models with Fewer Than 10B Parameters										
InternLM3-8B	99.09	99.66	99.38	96.00	98.91	97.45	23.44	35.85	29.64	75.49
LLaMA-3.1-8B	**100.00**	**100.00**	**100.00**	95.18	98.64	96.91	10.82	21.22	16.02	70.98
Qwen-2.5-7B	99.89	96.82	98.35	96.00	98.00	97.00	10.68	23.12	16.90	70.75
GLM-4-9B-Chat	98.98	88.41	93.69	97.32	**99.91**	98.61	4.40	10.06	7.23	66.51
Gemma-3-4B	95.23	89.89	92.56	83.00	86.77	84.89	15.28	16.34	15.81	64.42
InternLM2.5-7B-Chat-1M	99.43	99.66	99.55	90.95	98.55	94.75	14.57	11.88	13.22	69.17
Models with 10-20B Parameters										
Gemma-3-12B	92.61	99.55	96.08	91.77	94.86	93.32	33.72	39.32	36.52	75.31
Qwen-2.5-14B	99.89	95.91	97.90	98.09	97.73	97.91	29.20	22.95	26.08	73.96
Models Larger Than 20B Parameters										
Gemma-3-27B	96.70	98.98	97.84	94.18	96.36	95.27	**47.93**	48.15	**48.04**	80.38
Qwen-2.5-32B	99.43	99.77	99.60	98.91	99.68	**99.30**	32.19	39.55	35.87	78.25
LLaMA-3.1-70B	**100.00**	99.89	99.94	**99.00**	99.09	99.05	15.71	20.51	18.11	72.37
Qwen-2.5-72B	99.77	**100.00**	99.89	98.73	99.77	99.25	36.79	**51.05**	43.92	**81.02**



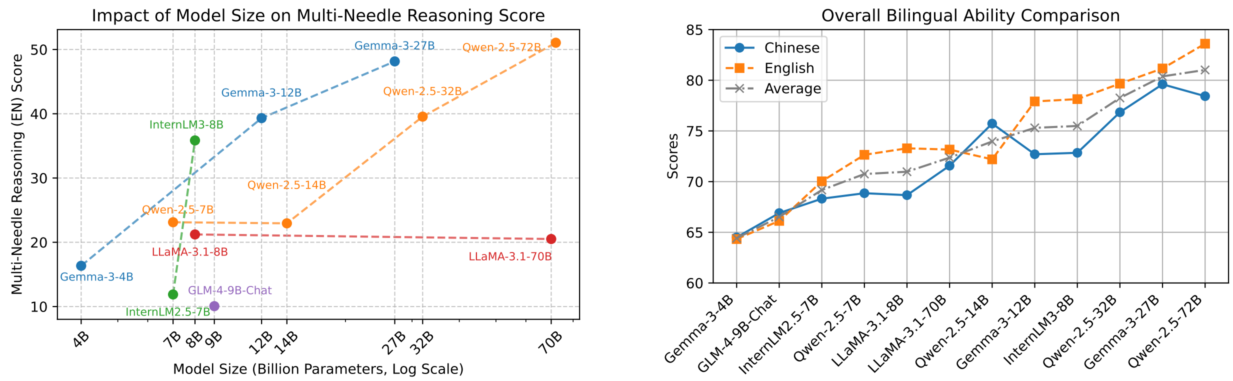
(a) Gemma-3-4B (b) Gemma-3-12B (c) Gemma-3-27B

Figure 3: Performance of Gemma-3 models with increasing parameter size on the 2-Needle Reasoning (EN, 128K) task. Larger models consistently achieve higher scores, highlighting the benefit of increased capacity for multi-point reasoning in long contexts.

Figure 4a demonstrates a clear positive correlation between model parameter size and Multi-Needle Reasoning (EN) performance, with a marked improvement observed beyond the 10B–20B parameter range, highlighting model scale as a key factor for long-context reasoning. Within individual model series, such as Gemma and Qwen, reasoning scores consistently increase with scale, reflecting classic scaling laws (Kaplan et al., 2020); however, this trend does not universally apply, as the LLaMA-3.1 series shows minimal performance gains from 8B to 70B parameters. While most small models (<10B) achieve low scores, certain models like InternLM3-8B outperform some larger counterparts, suggesting that training strategies, architecture, or fine-tuning can significantly boost small model capabilities. Notably, models with similar parameter counts from different series can exhibit substantial performance differences—for example, Gemma-12B-IT significantly outperforms Qwen-2.5-14B—indicating that architecture, pretraining data, and instruction tuning play crucial roles. Finally, the LLaMA series displays a saturation effect, with scores plateauing around 20 despite increasing parameters, implying that scaling alone is insufficient and further improvements require additional optimization strategies.

4.1.4 Effect of Needle Count on Multi-Needle Reasoning Performance

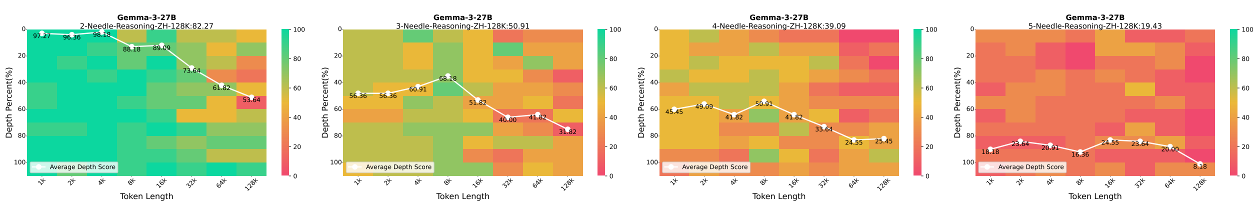
To further analyze how the number of reasoning points (needles) affects model performance, we visualize the results of Gemma-3-27B on the Multi-Needle Reasoning (ZH, 128K) task as the number of needles increases from 2 to 5, where "ZH" indicates that the task is conducted in Chinese. As shown in Fig. 5, the heatmaps become progressively redder, indicating a clear decline in performance as the reasoning complexity grows. This trend highlights the significant challenge that even strong models face when required to integrate a larger



(a) Model size vs. long-context reasoning ability.

(b) English vs. Chinese performance.

Figure 4: Impact of language and model size on NeedleBench 128K performance. Left: Increasing model parameter size generally leads to better long-context reasoning. Right: Most models exhibit a clear performance gap between English and Chinese, with English scores typically higher.



(a) 2-Needle

(b) 3-Needle

(c) 4-Needle

(d) 5-Needle

Figure 5: Performance of Gemma-3-27B on Multi-Needle Reasoning as the number of needles increases. The color shift towards red in the heatmaps indicates a clear decline in performance as the task complexity grows, highlighting the increasing challenge of integrating more information points within long contexts.

number of interdependent information points within long contexts. While retrieval over long context windows is becoming a solved problem for modern LLMs, integrating multiple critical details to form a coherent answer remains a significant challenge for all models. More advanced and larger-scale models generally perform better on such tasks, but even the strongest models are still far from perfect in this regard.

4.1.5 Impact of Language: Which Model Performs Better under the Bilingual Scenario?

To directly address which model performs best in the bilingual scenario, we analyze the overall NeedleBench 128K performance for both English and Chinese, as shown in Fig. 4b. **Qwen2.5-72B** achieves the highest scores in both languages, indicating the strongest bilingual performance. Across most evaluated models, there is generally a performance gap between English and Chinese, with English results typically being higher. Performance in English tends to be higher than in Chinese for most models, though this is not universal; for example, Qwen2.5-14B and GLM4-9B achieve slightly better results in Chinese. Such differences may be related to factors including pretraining data distribution, tokenization strategies, or language-specific modeling approaches. These observations suggest that further research is needed to improve cross-lingual generalization and to develop more robust multilingual long-context models.

4.2 *NeedleBench* Information-Dense Task

We present the results of the ATC task in Tab. 3, which evaluates model performance under different context lengths determined by the number of embedded factual units ('needles'). Here, "reasoning models" in Tab. 3 refer to models that explicitly output a "think" step or intermediate reasoning process before giving the

Table 3: ATC task results for all evaluated models. Models with stronger reasoning abilities generally achieve higher scores, with DeepSeek-R1 achieving the best overall performance (total score 44.01). In contrast, models with fewer parameters often achieve only single-digit scores.

Model	Needle Count									Evaluation Metric		
	2	4	8	16	32	64	128	256	512			
	Context Length (tokens)									Weighted Score		ENL
	0.4K	0.5K	0.6K	0.7K	1.0K	1.5K	2.7K	5K	9.6K	$\leq$ 2K	All	ENL-50
Closed-Source and Reasoning Models												
Claude-3.7-Sonnet-Thinking	**100.0**	**100.0**	92.5	92.5	67.5	40.0	15.0	7.5	2.5	59.84	12.39	32
DeepSeek-R1	**100.0**	**100.0**	87.5	95.0	**97.5**	**90.0**	70.0	65.0	**15.0**	**92.86**	**44.01**	**256**
GPT-4o	**100.0**	72.5	82.5	42.5	17.5	0.0	0.0	0.0	0.0	18.97	2.34	8
GPT-4.1	**100.0**	95.0	87.5	82.5	75.0	62.5	37.5	2.5	0.0	71.43	14.13	64
o3-mini	97.5	**100.0**	**97.5**	92.5	82.5	30.0	0.0	0.0	0.0	58.85	7.26	32
QwQ-32B	**100.0**	97.5	92.5	65.0	32.5	12.5	0.0	0.0	0.0	33.41	4.12	16
OREAL-32B	92.5	55.0	45.0	20.0	22.5	7.5	5.0	0.0	0.0	18.13	2.86	4
DeepSeek-R1-Qwen-32B	95.0	75.0	47.5	35.0	12.5	5.0	0.0	0.0	0.0	17.06	2.10	4
DeepSeek-R1-Qwen-14B	**100.0**	62.5	37.5	22.5	5.0	0.0	0.0	0.0	0.0	10.08	1.24	4
DeepSeek-R1-Qwen-7B	77.5	25.0	2.5	0.0	0.0	0.0	0.0	0.0	0.0	2.18	0.27	2
Models with 20B or More Parameters												
Qwen1.5-72B	70.0	25.0	10.0	0.0	0.0	0.0	0.0	0.0	0.0	2.54	0.31	2
Qwen2.5-72B	92.5	62.5	45.0	10.0	0.0	0.0	2.5	0.0	0.0	7.58	1.25	4
Qwen1.5-32B	75.0	20.0	7.5	7.5	5.0	0.0	0.0	0.0	0.0	4.52	0.56	2
Qwen2.5-32B	97.5	62.5	27.5	17.5	5.0	2.5	0.0	0.0	0.0	10.04	1.24	4
Gemma-3-27B	82.5	70.0	67.5	47.5	30.0	2.5	5.0	0.0	0.0	22.74	3.43	8
Models with 7-20B Parameters												
Qwen1.5-14B	52.5	17.5	10.0	2.5	2.5	0.0	0.0	0.0	0.0	2.98	0.37	2
Qwen2.5-14B	72.5	47.5	25.0	7.5	5.0	0.0	0.0	0.0	0.0	6.47	0.80	2
Gemma-3-12B	72.5	55.0	45.0	17.5	10.0	2.5	0.0	0.0	0.0	11.79	1.45	4
Mixtral-8x7B	50.0	12.5	15.0	7.5	0.0	0.0	0.0	0.0	0.0	3.10	0.38	2
GLM-4-9B	52.5	35.0	10.0	7.5	2.5	0.0	0.0	0.0	0.0	4.17	0.51	2
InternLM3-8B	55.0	32.5	27.5	20.0	2.5	0.0	0.0	5.0	0.0	6.83	2.09	2
Models with Fewer Than 7B Parameters												
Qwen1.5-1.8B	5.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.08	0.01	0
Qwen1.5-4B	35.0	10.0	7.5	0.0	0.0	0.0	0.0	0.0	0.0	1.35	0.17	0
Qwen2.5-1.5B	37.5	25.0	2.5	0.0	0.0	0.0	0.0	0.0	0.0	1.55	0.19	0
Qwen2.5-7B	75.0	37.5	7.5	5.0	5.0	0.0	0.0	0.0	0.0	4.76	0.59	2
Mistral-7B	40.0	17.5	2.5	2.5	0.0	0.0	0.0	0.0	0.0	1.67	0.21	0
Gemma-3-4B	57.5	27.5	15.0	17.5	2.5	0.0	0.0	0.0	0.0	5.60	0.69	2
ChatGLM3-6B-32K	27.5	7.5	10.0	5.0	2.5	0.0	2.5	0.0	2.5	2.58	1.88	0
InternLM2.5-7B-1M	60.0	27.5	15.0	2.5	5.0	0.0	0.0	0.0	0.0	4.37	0.54	2

final answer. As the needle count increases, the input context becomes longer and the question increasingly requires the model to perform continuous retrieval and reasoning.

Across all models, we observe a clear downward trend in performance as the number of needles increases in Fig. 6. Most models fail entirely beyond the 64-needle level, indicating that longer, information-dense contexts remain a major challenge. Additionally, **model size** plays a significant role: larger models tend to achieve higher scores. For example, within the Gemma family, performance steadily improves from 5.60 (4B) to 22.74 (27B), demonstrating the benefit of increased capacity in tackling information-dense tasks. When focusing on the average performance under short contexts ($\leq$2K tokens), the Gemma-3 series shows strong results across all model scales. In fact, Gemma-3 models of sizes 4B, 12B, and 27B each achieve the best score in their respective size categories, highlighting the series' robustness in low-to-medium complexity settings.

Can State-of-the-Art Reasoning Models Generalize to Long-Chain Reasoning? We use the ENL-50 metric to quantify the effective reasoning depth of each model—that is, the maximum number of compositional steps a model can reliably handle. As shown in Tab. 3, small models ($\leq$7B) can only generalize to 2-step reasoning, medium models (7–20B) up to 4 steps, and large models (>20B) up to 8 steps, with only the strongest models such as GPT-4.1 (ENL-50 = 64) and DeepSeek R1 (ENL-50 = 256) demonstrating the ability to generalize to much longer reasoning chains. Notably, the DeepSeek-R1-Qwen distillation models,

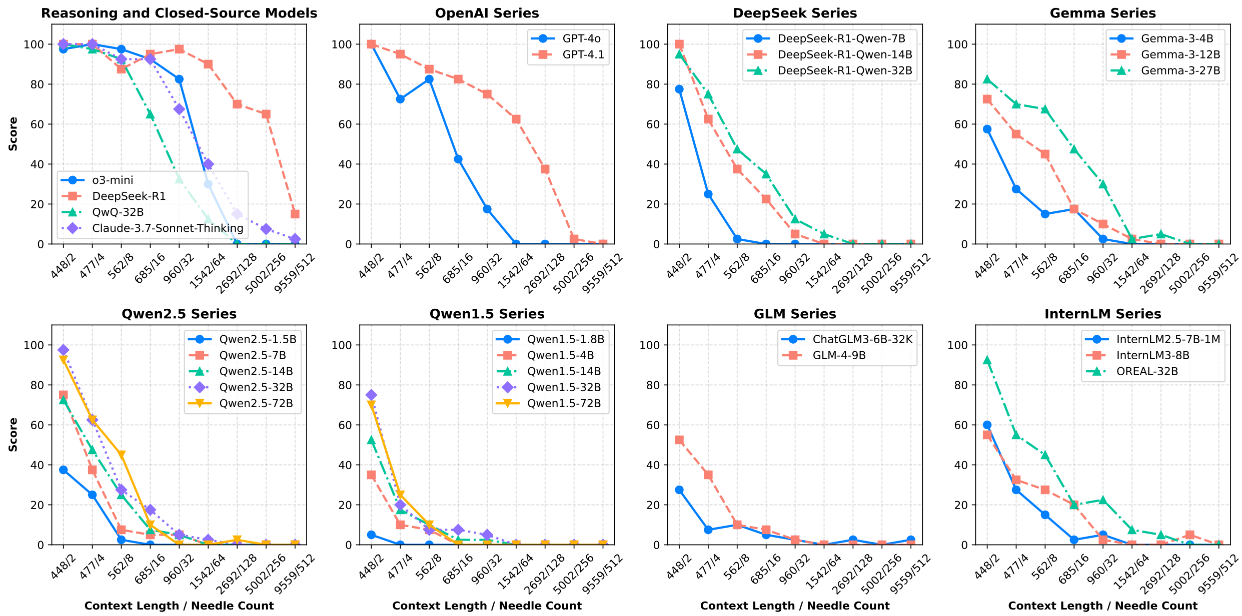


Figure 6: Performance decline trend of various models on ATC. OpenAI models (GPT-4 series) are shown separately for clarity. Notably, DeepSeek-R1 and GPT-4.1 exhibit a much slower decline in performance as the number of needles increases, demonstrating strong resistance to performance degradation on information-dense tasks. In contrast, most other models experience a rapid drop to near-zero scores as the length of the information-dense context increases.

which are trained by distilling DeepSeek R1's reasoning data, still perform quite poorly on our synthetic tasks: for example, DeepSeek-R1-Qwen-32B achieves an ENL-50 of only 4, and the 14B and 7B variants also fail to generalize beyond 2–4 steps. This indicates that, although these models may have memorized reasoning patterns from their teacher, they struggle to transfer such reasoning to broader, more diverse compositional tasks, highlighting the challenge of achieving true generalizable reasoning beyond rote pattern replication.

The "Under-Thinking" Bottleneck in Information-Dense Long-Context Tasks. To investigate why state-of-the-art (SOTA) models like o3-mini and DeepSeek R1 frequently fail at information-dense tasks, we manually annotated approximately 10% of the observed errors from the ATC task, focusing specifically on cases where the question type is "identifying the eldest ancestor." The main error types identified in our analysis are summarized in Table 4. For each error category, we provide representative prompt and response examples in Appendices C.1 to C.5 to illustrate how these errors occur in practice.

Our analysis reveals that the most prevalent failure mode among strong models is what we term *under-thinking*: models prematurely conclude that no further inference can be made, even when clear clues remain in the context. While the term "under-thinking" has previously been used to describe models abandoning a reasoning path too early in favor of another (often in math problem solving) (Wang et al., 2025), here we use it to refer to a distinct phenomenon: the inability to sustain inference by fully leveraging all relevant information, resulting in reasoning that **halts midway under the mistaken belief that nothing more can be inferred.** This under-thinking bottleneck, together with other characteristic errors, highlights the persistent challenges faced by LLMs in reliably handling information-dense, long-context tasks.

Beyond under-thinking errors, our analysis reveals several other characteristic failure modes in current LLMs on the ATC task. Instruction following errors are predominantly observed in smaller-parameter models, which often fail to comply with required output formats. Partial understanding errors occur when a model fixates on part of a statement—such as interpreting "Dan Newton is more than just a mother; Dan Newton is a lifelong mentor of Andrew Williams" as indicating only a mentorship—while overlooking the explicit familial role, in this case, "mother," thus breaking the full kinship chain. Repetitive output errors are frequently seen

Table 4: Analysis of Common Error Types in ATC Task: Under-thinking is the most prevalent error, especially in strong models like DeepSeek R1 and o3-mini. Partial understanding errors indicate models only grasp part of key information. Smaller models frequently exhibit instruction following and repetitive output errors.

Error Type (%)	**Explanation**	**Representative Models (non-exhaustive)**
Under-thinking (60.2%)	The model prematurely halts reasoning, asserting that no further inference can be made despite the presence of remaining clues.	Reasoning models (e.g., DeepSeek R1, GPT-4.1, o3-mini, Claude-3-7-Sonnet, InternLM3-8B-Instruct, Qwen2.5-1.5B-Instruct)
Partial Understanding Error (12.9%)	Model only identifies part of the relationships mentioned.	Gemma3 27B, InternLM3-8B-Instruct, Qwen2.5-7B-Instruct
Instruction Following Error (7.5%)	Model fails to follow the required output format.	Qwen1.5-1.8B-Chat, Qwen2.5-1.5B-Instruct
Repetitive Output (7.5%)	Model either repeats reasoning steps or outputs meaningless text.	Deepseek-R1-Distill-Qwen-7B, Qwen2.5-1.5B-Instruct
Hallucination Error (7.5%)	Model introduces information that is not present in the original text.	DeepseekR1, Qwen1.5-1.8B-Chat, Deepseek-R1-Distill-Qwen-7B,
Other Errors (4.5%)	Miscellaneous errors not covered by the above categories.	Various models

in models such as Deepseek-R1-Distill-Qwen-7B, suggesting a potential drawback of directly distilling large model outputs into smaller models via supervised fine-tuning. In addition, we observe other errors, such as misunderstanding the question intent, misinterpreting examples in the prompt as information to be used in reasoning, or making logical mistakes in the reasoning process. These issues are more prevalent in models with generally weaker overall performance. Collectively, these diverse error types highlight the persistent challenges faced by LLMs in reliably handling information-dense long-context tasks.

5 Conclusion and Future Work

In this research, we conduct a comprehensive evaluation of large language models (LLMs) on retrieval and reasoning tasks in long-context scenarios. Our results reveal that even state-of-the-art LLMs—including Claude 3.7 Sonnet-Thinking, o3-mini, and DeepSeek R1—exhibit notable shortcomings, especially on the Ancestral Trace Challenge, which is designed to test information-dense, multi-step retrieval and reasoning across extended texts. While recent long-context models have made progress in information retrieval, we find that they still struggle considerably when required to perform sustained, multi-step retrieval and reasoning over contexts where critical information is densely interwoven throughout the input.

Our research highlights the importance of targeted assessments in pinpointing and addressing critical gaps in LLMs' abilities to manage information-dense scenarios. The "under-thinking" phenomenon identified in our study further underscores the need to improve models' reasoning strategies—especially their tendency to prematurely conclude tasks even when additional evidence is available. Future work can include exploring reinforcement learning to help models improve their reasoning and mitigate under-thinking, as well as expanding NeedleBench to cover more diverse and realistic information-dense scenarios, since NeedleBench is a synthetic benchmark and may not fully reflect the complexity of real-world tasks.

References

Chenxin An, Fei Huang, Jun Zhang, Shansan Gong, Xipeng Qiu, Chang Zhou, and Lingpeng Kong. Training-free long-context scaling of large language models, 2024. URL https://arxiv.org/abs/2402.17463.

Anthropic. The claude 3 model family: Opus, sonnet, haiku — model card. https://docs.anthropic.com/en/docs/resources/claude-3-model-card, 2024a.

Anthropic. Introducing the next generation of claude. https://www.anthropic.com/news/claude-3-family, 2024b. Accessed: 2024-03-27.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report, 2023a. URL https://arxiv.org/abs/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. Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*, 2023b.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. *arXiv preprint arXiv:2412.15204*, 2024.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer, 2020. URL https://arxiv.org/abs/2004.05150.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, et al. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*, 2024.

LMDeploy Contributors. Lmdeploy: A toolkit for compressing, deploying, and serving llm. https://github.com/InternLM/lmdeploy, 2023.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Di Zhang. Aime_1983_2024 (revision 6283828), 2025. URL https://huggingface.co/datasets/di-zhang-fdu/AIME_1983_2024.

Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. https://storage.googleapis.com/deepmind-media/gemini/gemini_v1_5_report.pdf, 2024. Accessed: 2024-03-17.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. Ruler: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*, 2024.

Greg Kamradt. LLMs Need Needle In A Haystack Test-Pressure Testing LLMs. https://github.com/gkamradt/LLMTest_NeedleInAHaystack, 2023.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020. URL https://arxiv.org/abs/2001.08361.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Jinhyuk Lee, Anthony Chen, Zhuyun Dai, Dheeru Dua, Devendra Singh Sachan, Michael Boratko, Yi Luan, Sébastien MR Arnold, Vincent Perot, Siddharth Dalmia, et al. Can long-context language models subsume retrieval, rag, sql, and more? *arXiv preprint arXiv:2406.13121*, 2024.

Chengqi Lyu, Songyang Gao, Yuzhe Gu, Wenwei Zhang, Jianfei Gao, Kuikun Liu, Ziyi Wang, Shuaibin Li, Qian Zhao, Haian Huang, et al. Exploring the limit of outcome reward for learning mathematical reasoning, 2025.

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

OpenAI. Gpt-4 and gpt-4 turbo - documentation. https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo, 2023. Accessed: 2024-01-23.

OpenAI. Introducing gpt-4.1 in the api. https://openai.com/index/gpt-4-1/, 2025a. Accessed 2025-04-19.

OpenAI. Openai o3-mini. https://openai.com/index/openai-o3-mini/, 2025b. Accessed 2025-04-19.

Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. scikit-learn: Machine learning in python, 2011. URL https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models, 2023. URL https://arxiv.org/abs/2309.00071.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL `https://arxiv.org/abs/2503.19786`.

Qwen Team. Qwq-32b: Embracing the power of reinforcement learning, March 2025. URL `https://qwenlm.github.io/blog/qwq-32b/`.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. Zephyr: Direct distillation of lm alignment, 2023a.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. Zephyr: Direct distillation of lm alignment, 2023b. URL `https://arxiv.org/abs/2310.16944`.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Large search model: Redefining search stack in the era of llms. In *ACM SIGIR Forum*, volume 57, pp. 1–16. ACM New York, NY, USA, 2024.

Yue Wang, Qiuzhi Liu, Jiahao Xu, Tian Liang, Xingyu Chen, Zhiwei He, Linfeng Song, Dian Yu, Juntao Li, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. Thoughts are all over the place: On the underthinking of o1-like llms, 2025. URL `https://arxiv.org/abs/2501.18585`.

Tianwen Wei, Liang Zhao, Lichang Zhang, Bo Zhu, Lijie Wang, Haihua Yang, Biye Li, Cheng Cheng, Weiwei Lü, Rui Hu, Chenxia Li, Liu Yang, Xilin Luo, Xuejie Wu, Lunan Liu, Wenjun Cheng, Peng Cheng, Jianhao Zhang, Xiaoyu Zhang, Lei Lin, Xiaokun Wang, Yutuan Ma, Chuanhai Dong, Yanqi Sun, Yifu Chen, Yongyi Peng, Xiaojuan Liang, Shuicheng Yan, Han Fang, and Yahui Zhou. Skywork: A more open bilingual foundation model. *arXiv preprint arXiv:2310.19341*, 2023.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024a. URL `https://arxiv.org/abs/2407.10671`.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024b.

An Yang, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoyan Huang, Jiandong Jiang, Jianhong Tu, Jianwei Zhang, Jingren Zhou, et al. Qwen2. 5-1m technical report. *arXiv preprint arXiv:2501.15383*, 2025.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. Big bird: Transformers for longer sequences, 2021. URL `https://arxiv.org/abs/2007.14062`.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Qihao Wu, Junhao Chen, Zihang Xu, Zhenning Dai, Xu Han, Shuo Wang, Zhiyuan Liu, and Maosong Sun. Infinitebench: 128k long-context benchmark for language models. *arXiv preprint*, 2023.

A Evaluated Models

The following Tab. 5 presents a list of models evaluated in this study, along with their maximum context lengths.

Table 5: Evaluated Models. We used LMDeploy (Contributors, 2023) and vLLM (Kwon et al., 2023) to accelerate the inference process. Unless otherwise specified, we use greedy decoding with temperature set to 0 for all model outputs.

Series	**Models**	**Context Window**
Qwen	Qwen-1.5-1.8B, Qwen-1.5-4B, Qwen-1.5-14B, Qwen-1.5-32B, Qwen-1.5-72B Qwen-2.5-1.5B, Qwen-2.5-7B, Qwen-2.5-14B, Qwen-2.5-32B, Qwen-2.5-72B, QwQ-32B	32K-128K
Zhipu AI	ChatGLM3-6B-32K, GLM-4-9B-Chat, GLM-4-9B-Chat-1M	32K, 128K, 1M
InternLM	InternLM3-8B, InternLM2.5-7B-Chat-1M, OREAL-32B	32K–1M
LLaMA	LLaMA-3.1-8B, LLaMA-3.1-70B	128K
Mistral	Mistral-7B-Instruct-v0.2, Mixtral-8x7B-Instruct-v0.1	32K
Zephyr	Zephyr-7B-Beta	32K
Gemma	Gemma-3-4B-IT, Gemma-3-12B-IT, Gemma-3-27B-IT	128K
DeepSeek	DeepSeek-R1	128K
OpenAI	GPT-4o, GPT-4.1, o3-mini	128K–1M
Claude	Claude-3.7-Sonnet-Thinking	200K

B *NeedleBench* Prompt Examples

This section presents representative prompt examples for each major task in *NeedleBench*.

B.1 Single-Needle Retrieval

Figure 7, Fig. 8, and Fig. 9 show prompt examples for the Single-Needle Retrieval task, where the target information (needle) is placed at different positions within the context (beginning, middle, and end, respectively).

Single-Needle Retrieval (Needle First - Demonstration with English Version)

Prompt:

This is a test of long-text capability. You need to first read the long document below, and then answer the final question based on the information in the document.
The content of the long document is as follows

Hidden on Emerald Island is the legendary Stardust Shard.
—*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*—
—*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*—

Based on the information in the document, now please answer: What legendary item is hidden on Emerald Island? Please answer in the format 'The legendary item hidden on the Emerald Island is ________.'

Figure 7: An example prompt of Single-Needle Retrieval showcasing key information with the single needle positioned at the very beginning. In actual tests, the needle is placed at various depths within extended texts to evaluate performance under different conditions.

Single-Needle Retrieval (Needle Middle - Demonstration with English Version)

Prompt:
This is a test of long-text capability. You need to first read the long document below, and then answer the final question based on the information in the document.
The content of the long document is as follows

—*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*—
Hidden on Emerald Island is the legendary Stardust Shard.
—*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*—

Based on the information in the document, now please answer: What legendary item is hidden on Emerald Island? Please answer in the format 'The legendary item hidden on the Emerald Island is ________.'

Figure 8: An example prompt of Single-Needle Retrieval showcasing key information with the single needle positioned at the middle

B.2 Multi-Needle Retrieval

Figure 10 provides a prompt example for the Multi-Needle Retrieval task, which requires the model to extract multiple target items from a long context.

Single-Needle Retrieval (Needle Last - Demonstration with English Version)

Prompt:
This is a test of long-text capability. You need to first read the long document below, and then answer the final question based on the information in the document.
The content of the long document is as follows

—*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*—
—*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*—
Hidden on Emerald Island is the legendary Stardust Shard.

Based on the information in the document, now please answer: What legendary item is hidden on Emerald Island? Please answer in the format 'The legendary item hidden on the Emerald Island is _________.'

Figure 9: An example prompt of Single-Needle Retrieval showcasing key information with the single needle positioned at last

Multi-Needle Retrieval (Demonstration with five Needles English Version Prompt)

Prompt:
This is a test of long-text capability. You need to first read the long document below, and then answer the final questions one by one based on the information in the document.
The content of the long document is as follows

—*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*—
Hidden on Forgotten Island is the legendary Stardust Shard.
—*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*—
Hidden on Mythical Island is the legendary Time-Space Key.
—*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*—
The ruler of the Alpha Bot star system is Cosmic Ruler Starshine.
—*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*—
Hidden on Storm Island is the legendary Goodness Heart.
—*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*—
The ruler of the Orion star system is Guardian of Time Lightspeed.
—*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*—

Based on the information in the document, now please answer: What legendary item is hidden on Forgotten Island? What legendary item is hidden on Mythical Island? Who is the ruler of the Alpha Bot star system? What legendary item is hidden on Storm Island? Who is the ruler of the Orion star system? Please answer in the format of "The legendary item hidden on the Forgotten Island is _________, The legendary item hidden on the Mythical Island is _________, The ruler of the Alpha Bot star system is _________, The legendary item hidden on the Storm Island is _________, The ruler of the Orion star system is _________."

Figure 10: An example prompt of Multi-Needle Retrieval with the new question template (5 needles)

Multi-Needle Reasoning (Demonstration with Three Needles English Version)

Prompt:
This is a test of long-text capability. You need to first read the long document below, and then answer the final question based on the information in the document.
The content of the long document is as follows
—*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*—
Jasmine Lane is not only James Hill's father but also James Hill's role model.
—*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*—
Janet Guzman is not only Carolyn Hicks's maternal grandmother but also Carolyn Hicks's role model.
—*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*—
James Hill, as Janet Guzman's paternal grandfather, has a significant impact on Janet Guzman's upbringing.
—*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*—
Based on the information in the document, now please answer: Given the context described above, who is the eldest relative that 'Carolyn Hicks' can trace back to in the context?

For example:
Example 1: If James Hill's father is Jasmine Lane, and no further information about familial relationships is provided in the text, then the oldest relative James Hill can trace back to in the provided text is \boxed{Jasmine Lane}.
Example 2: If Andrew Williams's grandmother is Dan Newton, and Dan Newton's father is James Hill, and no further information about familial relationships is provided in the text, then the oldest relative Andrew Williams can trace back to in the provided text is \boxed{James Hill}.
Example 3: If Jeff White's father is Kevin Le, Dan Newton's grandmother is Jeff White, and Jeff White's father is Kevin Le, and Shelley Mills is Dan Newton's great-granddaughter, and no further information about familial relationships is provided in the text, then the oldest relative Shelley Mills can trace back to in the provided text is \boxed{Kevin Le}.

Notes:
1. You do not need to worry about the gender consistency of names in this test. For example, a name that is typically considered feminine can still be the father of another person. Our primary focus is on who is older.
2. Ignore surname inheritance issues. For instance, Andrew Williams could still be the biological father of Christopher Baker. We only care about who is older and do not need to consider whether a child should inherit the father's or mother's surname.
3. At the end of your response, remember to put your final answer within \boxed{}. For example: "So the oldest relative 'Carolyn Hicks' can trace back to in the provided text is \boxed{(your answer here)}."

Figure 11: An example prompt of Multi-Needle Reasoning

B.3 Multi-Needle Reasoning

Figure 11 presents a prompt example for the Multi-Needle Reasoning task, where the model must perform reasoning over several interrelated pieces of information distributed throughout the context.

B.4 Ancestral Trace Challenge

Figure 12 shows a prompt example for the Ancestral Trace Challenge (ATC), an information-dense task that requires multi-step logical reasoning to trace relationships and infer the correct answer.

Ancestral Trace Challenge (3-needle version example with DeepSeek R1 response)

Prompt:
Here is a test for multi-step reasoning ability called the Ancestral Trace Challenge. In this test, we will simulate different people's familial relationships, and your task is to continuously reason through them until you identify the eldest ancestor.
Now, the scrambled family relationships are provided below:
Wyatt James is the child of Maria Watson. Emily Barry, as Maria Watson's paternal grandfather, has a significant impact on Maria Watson's upbringing. Joseph Taylor is not only Emily Barry's maternal grandfather but also Emily Barry's role model.

Given the scrambled family relationships described above, who is the eldest relative that 'Wyatt James' can trace back to in the context?

For example:
Example 1: If James Hill's father is Jasmine Lane, and no further information about familial relationships is provided in the text, then the oldest relative James Hill can trace back to in the provided text is \boxed{Jasmine Lane}.
Example 2: If Andrew Williams's grandmother is Dan Newton, and Dan Newton's father is James Hill, and no further information about familial relationships is provided in the text, then the oldest relative Andrew Williams can trace back to in the provided text is \boxed{James Hill}.
Example 3: If Jeff White's father is Kevin Le, Dan Newton's grandmother is Jeff White, and Jeff White's father is Kevin Le, and Shelley Mills is Dan Newton's great-granddaughter, and no further information about familial relationships is provided in the text, then the oldest relative Shelley Mills can trace back to in the provided text is \boxed{Kevin Le}.

Notes:
1. You do not need to worry about the gender consistency of names in this test. For example, a name that is typically considered feminine can still be the father of another person. Our primary focus is on who is older.
2. Ignore surname inheritance issues. For instance, Andrew Williams could still be the biological father of Christopher Baker. We only care about who is older and do not need to consider whether a child should inherit the father's or mother's surname.
3. At the end of your response, remember to put your final answer within \boxed{}. For example: "So the oldest relative xxx(the person you need to find the eldest relative for) can trace back to in the provided text is \boxed{}."

DeepSeek R1 Response (Correct Example):
Wyatt James is the child of Maria Watson, making Maria one generation above Wyatt. Maria's paternal grandfather is Emily Barry, placing Emily two generations above Maria and three generations above Wyatt. Emily Barry's maternal grandfather is Joseph Taylor, which adds two more generations above Emily (Emily's mother and then Joseph). Therefore, Joseph Taylor is five generations above Wyatt. Among the mentioned relatives, Joseph Taylor is the eldest ancestor.

So the oldest relative Wyatt James can trace back to in the provided text is \boxed{Joseph Taylor}.

Figure 12: Example prompt and DeepSeek R1 response in the Ancestral Trace Challenge (ATC) with multi-step reasoning

C Error Analysis Examples

In this section, we present representative error cases observed in the Ancestral Trace Challenge (ATC) task. Each error type is illustrated with a concrete example and referenced in the main text.

C.1 Under-thinking Error

See Fig. 13 for an example of an Under-thinking Error. This error occurs when the model prematurely halts its reasoning process, incorrectly concluding that no further inference can be made, even though additional relevant clues remain in the context. As a result, the model fails to identify the true eldest ancestor.

C.2 Instruction Following Error

See Fig. 14 for an example of an Instruction Following Error. In this case, the model's reasoning may be correct, but it fails to adhere to the required output format specified in the task instructions (e.g., omitting the \boxed{} format), resulting in an incomplete or invalid response.

C.3 Partial Understanding Error

See Fig. 15 for an example of a Partial Understanding Error. This error type is characterized by the model identifying only a subset of the relationships or information present in the context, leading to an incomplete or partially correct answer.

C.4 Repetitive Output Error

See Fig. 16 for an example of a Repetitive Output Error. Here, the model either enters a repetitive reasoning loop—reiterating the same inference steps or chains without reaching a conclusion—or repeatedly outputs meaningless or irrelevant content.

C.5 Hallucination Error

See Fig. 17 for an example of a Hallucination Error. This error occurs when the model introduces information or relationships that are not present in the original context, resulting in fabricated or unsupported answers.

Ancestral Trace Challenge (Under-thinking Error Example by GPT-4.1)

Prompt:
The initial question prompt is omitted here for brevity; see Fig. 14 for identical content.
Now, the scrambled family relationships are provided below:
James Tate's father is Natasha Weeks. Gary Anderson is not only Casey Johnson's father but also Casey Johnson's role model. For Cesar Pacheco, Dana Martinez is not just a mom, but also a friend. Taylor Collins is the child of Douglas Harris. Casey Johnson is not only Natasha Weeks's maternal grandfather but also Natasha Weeks's role model. Daniel Morton is not only Jessica Stewart's mom but also Jessica Stewart's role model. For John Cole MD, Sandra Cameron is more than just a father; Sandra Cameron is a lifelong mentor of John Cole MD. Tami Woodward, as Sandra Cameron's grandmother, has a significant impact on Sandra Cameron's upbringing. Mary Smith is William Castaneda's dad. Stephen Hunter's maternal grandmother is James Tate. For Kevin Le, Nicole Downs is not just a paternal grandmother, but also a friend. Dennis Middleton is Michael Martinez's mother. Kelly Meyer's mom is Eileen Green. For Sheri Garrett, Holly Jones is more than just a paternal grandmother; Holly Jones is a lifelong mentor of Sheri Garrett. For Christy Dixon, Cesar Pacheco is not just a maternal grandmother, but also a friend. Gary Anderson's father is John Cole MD. Sheri Garrett is Kaitlyn Chandler's maternal grandmother. Jonathan Payne is the child of Julie Ross. Jessica Martin is the child of Shannon Garner. Julie Ross is the child of Taylor Collins. For Samantha Sheppard, Kaitlyn Chandler is more than just a mother; Kaitlyn Chandler is a lifelong mentor of Samantha Sheppard. For Nicole Downs, Jessica Stewart is not just a mother, but also a friend. Kelly Meyer, as Mary Smith's father, has a significant impact on Mary Smith's upbringing. For Shannon Garner, Stephen Hunter is not just a mother, but also a friend. Christy Dixon is Sean Mccoy's grandfather. Jonathan Payne is not only Daniel Morton's mother but also Daniel Morton's role model. Holly Jones's paternal grandfather is Michael Martinez. Douglas Harris is the child of William Castaneda. **Jessica Martin is Eileen Green's paternal grandmother.** Samantha Sheppard, as Dana Martinez's mom, has a significant impact on Dana Martinez's upbringing. Tami Woodward's maternal grandmother is Sean Mccoy.

Given the scrambled family relationships described above, who is the eldest relative that 'Kevin Le' can trace back to in the context?

GPT-4.1 Response (Under-thinking Error):
Let's break down the relationships relevant to **Kevin Le** by tracing up along every provided parental and grandparental link. We are looking for the furthest ancestor **Kevin Le** can trace from the data.
Step 1–11 are all correctly reasoned and omitted here due to space constraints.

Step 12: Does Eileen Green have an ancestor?
Look for Eileen Green as a child in any other statement. Let's check all the family chain we have so far.

No further information about Eileen Green's parentage appears in the provided text.

The model's response summary here is omitted to highlight the subsequent error in its answer.

Final answer:
So, the eldest relative 'Kevin Le' can trace back to in the provided text is \boxed{Eileen Green}.

Figure 13: An example of an Under-thinking Error in the Ancestral Trace Challenge (ATC) task. The GPT-4.1 model prematurely concludes the reasoning process and fails to identify the true eldest ancestor, despite the presence of additional relevant information in the context.

Ancestral Trace Challenge (Instruction Following Error Example by Qwen1.5-1.8B-Chat)

Prompt:
Here is a test for multi-step reasoning ability called the Ancestral Trace Challenge. In this test, we will simulate different people's familial relationships, and your task is to continuously reason through them until you identify the eldest ancestor.
For example:
Example 1: If James Hill's father is Jasmine Lane, and no further information about familial relationships is provided in the text, then the oldest relative James Hill can trace back to in the provided text is \boxed{Jasmine Lane}.
Example 2: If Andrew Williams's grandmother is Dan Newton, and Dan Newton's father is James Hill, and no further information about familial relationships is provided in the text, then the oldest relative Andrew Williams can trace back to in the provided text is \boxed{James Hill}.
Example 3: If Jeff White's father is Kevin Le, Dan Newton's grandmother is Jeff White, and Jeff White's father is Kevin Le, and Shelley Mills is Dan Newton's great-granddaughter, and no further information about familial relationships is provided in the text, then the oldest relative Shelley Mills can trace back to in the provided text is \boxed{Kevin Le}.

Notes:
1. You do not need to worry about the gender consistency of names in this test. For example, a name that is typically considered feminine can still be the father of another person. Our primary focus is on who is older.
2. Ignore surname inheritance issues. For instance, Andrew Williams could still be the biological father of Christopher Baker. We only care about who is older and do not need to consider whether a child should inherit the father's or mother's surname.
3. At the end of your response, remember to put your final answer within \boxed{}. For example: "So the oldest relative 'Andrew Williams' can trace back to in the provided text is \boxed{somebody (your answer here)}."

Now, the scrambled family relationships are provided below:
For Andrew Williams, Dan Newton is more than just a mother; Dan Newton is a lifelong mentor of Andrew Williams.

Given the scrambled family relationships described above, who is the eldest relative that 'Andrew Williams' can trace back to in the context?

Qwen1.5-1.8B-Chat Response (Instruction Following Error):
Based on the scrambled family relationships provided, the eldest relative that 'Andrew Williams' can trace back to is Dan Newton. This is because Dan Newton is mentioned as Andrew's grandmother, and no further information about familial relationships is provided in the text. Therefore, we can infer that Dan Newton is the oldest relative and the one who has been a mentor to Andrew throughout his life.

Figure 14: An example of an Instruction Following Error in the ATC task. The Qwen1.5-1.8B-Chat model gives the correct reasoning but fails to output the answer in the required \boxed{} format.

Ancestral Trace Challenge (Partial Understanding Error Example by InternLM3-8B)

Prompt:
The initial question prompt is omitted here for brevity; see Fig. 14 for identical content.
Now, the scrambled family relationships are provided below:
For Andrew Williams, Dan Newton is more than just a mother; Dan Newton is a lifelong mentor of Andrew Williams. Andrew Williams, as Brooke Sanders's mother, has a significant impact on Brooke Sanders's upbringing. Brooke Sanders is not only Leah Williams's maternal grandfather but also Leah Williams's role model.

Given the scrambled family relationships described above, who is the eldest relative that 'Leah Williams' can trace back to in the context?

InternLM3-8B Response (Partial Understanding Error):
To determine the eldest relative that Leah Williams can trace back to, let's break down the given family relationships step by step:
1. **Andrew Williams and Dan Newton:**
- **Dan Newton is a lifelong mentor of Andrew Williams. This does not imply a familial relationship, so we can ignore this for our purpose of tracing ancestry.**

2. **Andrew Williams and Brooke Sanders:**
- Andrew Williams is Brooke Sanders's mother. This means Andrew Williams is the mother of Brooke Sanders.

3. **Brooke Sanders and Leah Williams:**
- Brooke Sanders is Leah Williams's maternal grandfather. This means Brooke Sanders is the father of Leah Williams's mother.

From these relationships, we can trace the lineage as follows:
- Leah Williams's mother is Andrew Williams.
- Andrew Williams's father is Brooke Sanders.

Therefore, the eldest relative that Leah Williams can trace back to in the provided text is \boxed{Brooke Sanders}.

Figure 15: An example of a Partial Understanding Error in the ATC task. The model locates the key information but only interprets part of it—for example, it notices "Dan Newton is more than just a mother; Dan Newton is a lifelong mentor of Andrew Williams" but only treats Dan Newton as a mentor, missing the familial relationship.

Ancestral Trace Challenge (Repetitive Output Error Example by Deepseek-R1-Distill-Qwen-7B)

Prompt:
The initial question prompt is omitted here for brevity; see Fig. 14 for identical content.
Now, the scrambled family relationships are provided below:
Keith Jordan is the child of Anna Ford. Anna Ford's grandfather is Brian Barron. Kayla Lucas is Brian Barron's mom.

Given the scrambled family relationships described above, who is the eldest relative that 'Keith Jordan' can trace back to in the context?

Deepseek-R1-Distill-Qwen-7B Response (Repetitive Output Error):
Okay, so I need to figure out who is the eldest ancestor that Keith Jordan can trace back to based on the given family relationships. Let me break down the information step by step.

First, the information provided is:
1. Keith Jordan is the child of Anna Ford.
2. Anna Ford's grandfather is Brian Barron.
3. Kayla Lucas is Brian Barron's mom.

I need to determine the eldest relative Keith can trace back to. Let's start by mapping out the family tree based on the given information.

Starting with Keith Jordan, he has a parent, Anna Ford. So, Anna is Keith's mother. Now, Anna's grandfather is Brian Barron. That means Brian is two generations above Anna. Since Anna is Keith's mother, Brian would be the great-grandfather of Keith. Because Brian is the grandfather of Anna, who is Keith's mother, Brian is the great-grandfather of Keith.

Additionally, Kayla Lucas is Brian Barron's mom. So, Kayla is Brian's daughter. Since Brian is the great-grandfather of Keith, Kayla would be the great-grandmother of Keith. **But wait,** Kayla is Brian's mom, so she is three generations above Keith. That makes her the great-grandmother, not the grandmother.

Wait, let me clarify. If Brian is the grandfather of Anna, then Anna is Brian's granddaughter. Since Anna is Keith's mother, Keith is Brian's great-grandson. Therefore, Brian is the great-grandfather of Keith. Kayla is Brian's mother, so she is three generations above Keith, making her the great-grandmother.

But the question is asking for the eldest relative that Keith can trace back to. So, starting from Keith, the immediate parent is Anna, then Anna's parent is Brian, and Brian's parent is Kayla. So, the chain is Keith → Anna → Brian → Kayla.

Therefore, the eldest relative Keith can trace back to is Kayla Lucas, who is his great-grandmother.

[The model continues to repeat the same reasoning and chain multiple times, restating the relationships and the conclusion in nearly identical wording, e.g., "Wait, but in the examples, when someone's grandmother is mentioned, the eldest is the father of that grandmother..." and so on, for several more paragraphs.]

Figure 16: An example of a Repetitive Output Error in the ATC task. The Deepseek-R1-Distill-Qwen-7B model repeatedly tries to apply the example pattern of tracing to a grandparent's parent, but since no further ancestor is given for Kayla Lucas, it gets stuck in a loop, restating the same reasoning chain over and over.

Ancestral Trace Challenge (Hallucination Error Example by Deepseek-R1-Distill-Qwen-7B)

Prompt:

The initial question prompt is omitted here for brevity; see Fig. 14 for identical content.

Now, the scrambled family relationships are provided below:

For Nancy Gilbert, Kathy Marshall is not just a maternal grandmother, but also a friend. For Carol Barron, George Estes is not just a mom, but also a friend. Carol Barron is not only Kathy Marshall's maternal grandfather but also Kathy Marshall's role model.

Given the scrambled family relationships described above, who is the eldest relative that 'Nancy Gilbert' can trace back to in the context?

Deepseek-R1-Distill-Qwen-7B Response (Hallucination Error):

Alright, let's try to figure out who the eldest ancestor Nancy Gilbert can trace back to. I'm a bit confused, but I'll take it step by step.

First, I need to understand the given information. The problem provides some scrambled family relationships:

1. For Nancy Gilbert, Kathy Marshall is not just a maternal grandmother but also a friend.
2. For Carol Barron, George Estes is not just a mom but also a friend.
3. Carol Barron is not only Kathy Marshall's maternal grandfather but also her role model.

The model's reasoning process here is omitted to highlight the subsequent error in its answer.

Additionally, Carol is the mother of George Estes, who is a friend but not necessarily related to Nancy. So, the key here is that Carol is the maternal grandfather of Kathy, who is the maternal grandmother of Nancy. Therefore, Carol is the great-grandfather of Nancy's mother, making him the great-great-grandfather of Nancy.

But wait, the question is about the eldest ancestor Nancy can trace back to. So, starting from Nancy, her maternal grandmother is Kathy, whose parent is her maternal great-grandmother, and her grandfather is Carol. So, Carol is the great-grandfather of Nancy's mother, which makes him the great-great-grandfather of Nancy. Therefore, Carol is the eldest ancestor Nancy can trace back to.

Wait, but let me double-check.

Nancy's maternal grandmother is Kathy. Kathy's parent is Nancy's maternal great-grandmother. Kathy's father is Carol, making Carol the maternal grandfather of Nancy's mother. Therefore, Carol is the great-grandfather of Nancy's mother, which makes him the great-great-grandfather of Nancy. So, Carol is the eldest ancestor Nancy can trace back to.

I think that's it. So, the answer should be Carol Barron.

So the eldest relative Nancy Gilbert can trace back to in the provided text is \boxed{Carol Barron}.

Figure 17: An example of a Hallucination Error in the ATC task. The Deepseek-R1-Distill-Qwen-7B model hallucinates the relationship in "For Carol Barron, George Estes is not just a mom, but also a friend," mistakenly treating Carol as the mother of George Estes and thus inferring the wrong ancestor.