HIDDEN IN THE HAYSTACK: SMALLER NEEDLES ARE MORE DIFFICULT FOR LLMS TO FIND

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) face significant challenges with needle-in-a-haystack tasks, where relevant information ("the needle") must be drawn from a large pool of irrelevant context ("the haystack"). Previous studies have highlighted positional bias and distractor quantity as critical factors affecting model performance, yet the influence of *gold context size*, the length of the answer-containing document, has received little attention. We present the first systematic study of gold context size in long-context question answering, spanning three diverse benchmarks (general knowledge, biomedical reasoning, and mathematical reasoning), eleven state-of-the-art LLMs (including recent reasoning models), and more than 150K controlled runs. Our experiments reveal that LLM performance drops sharply when the gold context is shorter, i.e., **smaller gold contexts consistently degrade model performance and amplify positional sensitivity**, posing a major challenge for agentic systems that must integrate scattered, fine-grained information of *varying lengths*. This effect persists under rigorous confounder analysis: even after controlling for gold document position, answer token repetition, gold-to-distractor ratio, distractor volume, and domain specificity, gold context size remains a decisive, independent predictor of success. Our work provides clear insights to guide the design of robust, context-aware LLM-driven systems.

1 INTRODUCTION

Large language models (LLMs) increasingly power applications that require reasoning over vast amounts of information, from synthesizing evidence across scientific literature (Gao et al., 2025; Baek et al., 2024; Sprueill et al., 2024; Bazgir et al., 2025; Wysocki et al., 2024; Cui et al., 2025; Wang et al., 2025), to navigating complex codebases (Liu et al., 2023; Zhang et al., 2023; Bogomolov et al., 2024), to maintaining coherence in multi-turn conversations. These applications share a common requirement: strong long-context understanding. This is particularly vital for *agentic systems*, in which autonomous agents must integrate heterogeneous information streams from specialized components to reason, plan, and act effectively.

A critical stage in such systems is *aggregation*, the synthesis of retrieved evidence into an accurate, actionable response. This stage determines what content to include, cite, or ignore, and has direct implications for safety, reliability, and factual correctness. Aggregation becomes especially challenging in *needle-in-a-haystack* scenarios, where relevant evidence (the 'gold context') is embedded within a large volume of topically related or superficially plausible but ultimately irrelevant or misleading, 'distractor context' (Tay et al., 2021; Shaham et al.). Successful aggregation requires precise identification and prioritization of minimal but essential content while discarding noisy signals.

Although LLMs now support context windows stretching into the millions of tokens, recent studies show that simply increasing input length does not ensure strong long-context reasoning. Prior work has explored *positional bias* (Wang et al., 2023; Liu et al.; Zheng et al., 2023), showing that early content is more likely to be attended to, and that distractors degrade performance. However, one key dimension remains underexplored: *how does the size of the gold context influence model performance?*

In this study, we present the first systematic analysis of gold context size as an independent variable in LLM long-context performance. We adapt three diverse benchmarks, CARDBiomedBench (Bianchi et al., 2025) (biomedical reasoning), NaturalQuestions (Kwiatkowski et al., 2019) (general

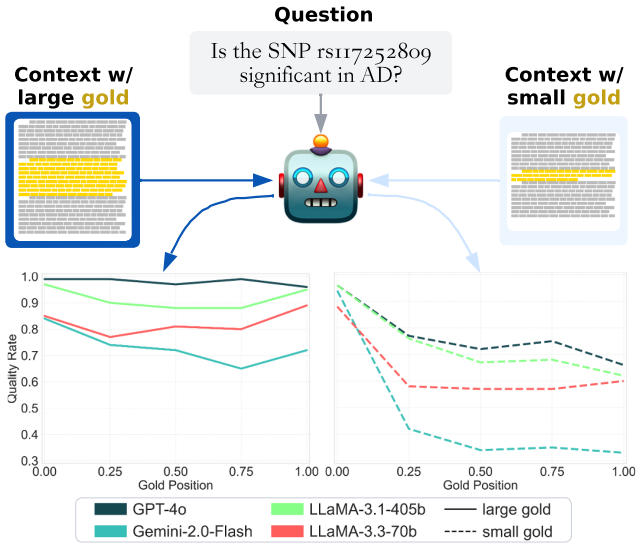


Figure 1: Changing both the size and position of gold context (relevant information) within a set of distractor context (irrelevant information), we observe that LLM Needle-In-A-Haystack Performance is *overall lower* and *more sensitive to position* when models are given short gold documents (dashed line) as opposed to long (solid line).

knowledge), and NuminaMath1.5 (LI et al., 2024) (mathematical reasoning), to vary both the size and position of the gold context while keeping the distractor content fixed (see Section 2). Our study spans over 150K controlled runs on eleven state-of-the-art LLMs, both general-purpose and reasoning-focused, ensuring that the observed effects hold across architectures and tasks. Smaller gold contexts lead to (1) *significantly worse performance* and (2) *higher positional sensitivity*.

These results complicate the common narrative that longer inputs degrade model performance. We find instead that **larger gold documents can improve accuracy**, while smaller golds are disproportionately vulnerable to positional bias. This reveals a brittleness not captured in prior work. Notably, models achieve near-perfect scores in no-distractor settings, confirming that these failures are due to aggregation breakdowns rather than task difficulty (see Section 3).

We conduct extensive confounder analyses to ensure that this effect is not an artifact. Controlling for gold document position, answer token repetition, gold-to-distractor ratio, distractor volume, and domain specificity, gold context size remains a decisive, independent predictor of performance. Smaller golds are more vulnerable to primacy bias, whereas larger golds show greater robustness to position and distractors.

These findings carry practical implications. In real-world deployments, context size, position, distractor size, and noise are rarely controllable. Aggregation failures due to overlooked gold context size can degrade trust, safety, and downstream task performance. Based on our empirical findings, designers of agentic systems should monitor length disparities across evidence documents—especially when shorter documents may carry critical information—to mitigate potential fragility.

In summary, our contributions are as follows:

- **Novel determinant of long-context performance:** To the best of our knowledge, we are the first to demonstrate that the *size* of gold contexts functions as a hidden variable in LLM long-context performance. Smaller golds degrade accuracy and amplify positional bias, underscoring a potential fragility in real-world applications.
- **Robust to confounders:** We identify and analyze five potential confounding variables, (1) gold document position, (2) answer token repetition, (3) gold-to-distractor ratio, (4) distractor volume, and (5) domain specificity, demonstrating that gold context size remains a decisive predictor of success despite these factors.
- **Large-scale experimentation:** We repeat our experiments and aggregate findings across eleven state-of-the-art LLMs, three diverse benchmarks, three sizes of gold, and six positions in the context window totaling over 150k controlled runs.

2 EXPERIMENTAL SETUP

We have designed our experiments to systematically evaluate how gold context size affects long-context LLM performance. This section outlines our design objectives, benchmark adaptations, baseline validations, primary evaluations simulating realistic aggregation, and potential confounding variables.

2.1 DESIDERATA

To systematically evaluate the impact of gold context size on long-context LLM performance, our guiding desiderata were (1) **Realism**, (2) **Gold Size Variability**, (3) **Distractors**, and (4) **Generality**.

Realism. In real-world agentic systems, aggregation involves synthesizing outputs from multiple specialized agents, each retrieving information from their domain of expertise. Usually, one agent returns the document containing the correct answer (the "gold" document), while others provide distractors, topically relevant but ultimately uninformative. We simulate this by inserting a gold document of varying size at different positions within a fixed-length sequence of distractors. Document order is randomized to reflect natural uncertainty in agent contributions and retrieval quality.

Gold Size Variability. We constructed three nested gold variants for each benchmark:

- **Small Gold**: Minimal span sufficient to answer the question.
- **Medium Gold**: Additional explanatory or supporting content.
- **Large Gold**: Complete reasoning process and/or extended relevant context.

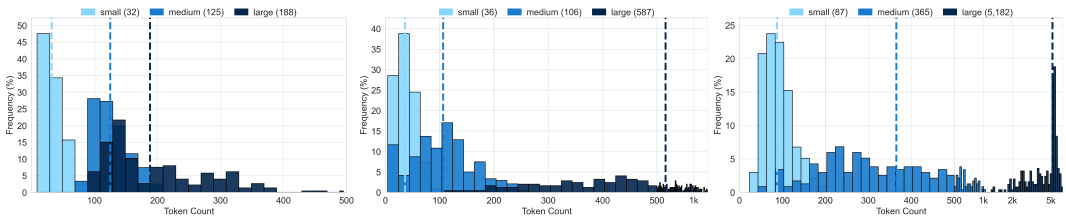
These were wrapped in pseudo-documents (titles, questions). Variants are hierarchically structured (small $\subset$ medium $\subset$ large) and validated for sufficiency. See Figure 8 for examples. Performance is high and uniform when observing only the gold of any size (Appendix B.1).

Distractors. To simulate realistic scenarios, we curate distractors topically relevant and lexically similar to the question but lacking the answer. Quantities per benchmark were calibrated to match token distributions observed in a real-world multi-agent retrieval system ($\sim 20k$ tokens).

Generality. We select three diverse benchmarks spanning biomedical, general knowledge, and mathematical reasoning, and evaluate performance across eleven leading LLMs of varying architecture and scale. This ensures that our findings generalize across domains and model classes.

2.2 TASK CONSTRUCTION: NEEDLES AND HAYSTACKS

We adapt three established question and answering benchmarks—CARDBiomedBench (biomedical reasoning), NaturalQuestions (general knowledge), and NuminaMath1.5 (mathematical reasoning)—to create controlled needle-in-a-haystack settings. Gold context sizes were varied, accompanied by distractors explicitly designed to be topically relevant yet answer-free. Figure 2 displays token count distributions for the varying sizes of gold.



(a) CARDBiomedBench (b) NaturalQuestions (c) NuminaMath1.5

Figure 2: Token count distributions for varying sizes of gold context on each benchmark. Median token count is in parenthesis in the legend. X-axis is scaled as linear (0-500) and logarithmic (500+).

CARDBiomedBench (CBB). CBB is a question-answering dataset focused on neurodegenerative diseases, designed to evaluate LLM performance on biomedical reasoning tasks involving genetic,

molecular, and clinical information. The BiomedSQL (for Alzheimer's & , CARD) variant augments each example with SQL queries and database rows to support structured reasoning experiments:

Full answer ⊂ SQL query + answer ⊂ SQL query + rows + answer

Distractors were drawn from documents retrieved by four independent domain-specific agents in a real-world system. These documents are semantically relevant but verifiably do not contain the answer, presenting realistic aggregation challenges.

NaturalQuestions (NQ). NQ is an open-domain QA benchmark of Google queries, with Wikipedia passages from the KILT corpus as evidence (Petroni et al., 2021):

Sentence with the answer ⊂ Paragraph with the sentence ⊂ Paragraph ± 4 paragraphs

Distractors were drawn from the HELMET (Yen et al., 2025) adaptation of NQ-KILT as 100-token segments. We exclude any documents labeled as evidence or containing the answer. This ensures distractors remain lexically and topically aligned with the question, but free of the answer.

NuminaMath1.5 (NM). NM is the largest open-source dataset for math reasoning, with problems ranging from high school to International Mathematical Olympiad (IMO)-level difficulty, originating from diverse sources like Chinese K-12 exams, AMC/AIME contests, and global math forums. We used the OpenR1Math (R1, 2024) variant, which includes model-generated solution traces from DeepSeekR1 (DeepSeek-AI et al., 2025) verified for correctness. We filter for examples with complete reasoning streams and define gold variants as:

Full answer ⊂ Textbook-style solution + answer ⊂ Full LLM-generated chain-of-thought + solution + answer

Distractors were reasoning traces to different questions. Due to length variability, we cap large gold contexts at the final 5k tokens, which include concluding reasoning and answers.

2.3 Baseline Experiments

We run three baseline conditions to validate that observed performance differences in main experiments result from changes to gold size, rather than underlying flaws in datasets or distractor construction. Baseline results across all benchmarks and models can be found in Appendix B.1:

- **Closed-book.** No context is provided, assessing whether models could answer from internal knowledge. This gauges possible benchmark saturation.
- **Gold-only.** Each gold context (sm, md, lg) is presented alone, without distractors. This confirms variants were sufficient to solve the task and that downstream performance drops are due to aggregation effects (e.g., distractor interference or gold placement).
- **Distractor-only.** Models are given only distractor documents. For CBB, we also test distractors from each agent separately to confirm they were individually non-informative. These checks ensure that distractors lack sufficient signal to answer correctly (Appendix B.1).

2.4 Main Experiments

We simulate realistic aggregation scenarios by embedding each gold context size at varying positions within a fixed sequence of distractors. This tests both gold size and positional sensitivity simultaneously. We evaluate eleven leading LLMs:

- **Closed-weight**: o3-mini (OpenAI, 2025), GPT-4o (OpenAI et al., 2024), GPT-4o-Mini (OpenAI, 2024), Gemini-2.5-Flash, Gemini-2.0-Flash, and Gemini-2.0-Flash-Lite (Mallick & Kilpatrick, 2025)
- **Open-weight**: DeepSeek-R1 (DeepSeek-AI et al., 2025), Phi-4-reasoning (Abdin et al., 2025), LLaMA-3.1-405B, LLaMA-3.3-70B, and LLaMA-3.1-8B (Dubey et al., 2024)

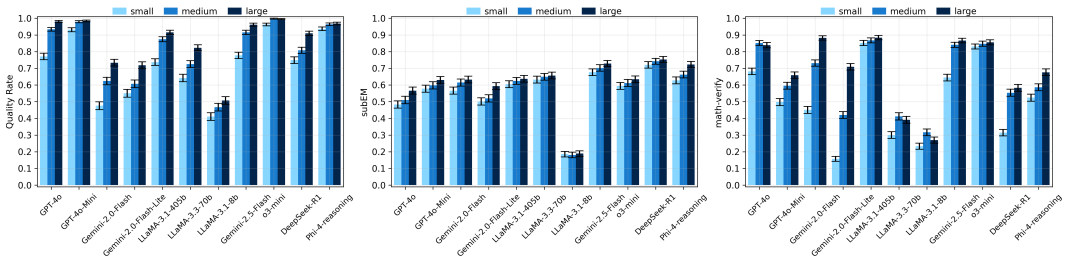
We evaluate each model on every size-position combination in a deterministic setting. Prompts were standardized within each benchmark. This enables rigorous, cross-model evaluation of gold context sensitivity and simulates the type of unpredictable document ordering in LLM systems.

3 Main Finding: Smaller Gold Contexts Lead to Lower Performance

Our experiments reveal that gold context size has a substantial and consistent effect on long-context performance, irrespective of confounding variables, across different benchmarks and models.

Increasing the size of the gold context significantly improves accuracy (Figure 3). On CBB, Gemini-2.0-Flash went from 48% with small to 62% with medium and 73% with large. GPT-4o performs similarly, rising from 77% (small) to 98% (large), while LLaMA-3.1-405B went from 74% to 92%.

Notably, performance with large gold contexts approaches the *gold-only baselines* (i.e., accuracy when the gold context is shown without any distractors) recorded at 96% for Gemini-2.0-Flash, and 100% for both GPT-4o and LLaMA-3.1-405B. This suggests that large gold contexts allow models to nearly recover ideal aggregation performance, while small golds fall significantly short.



(a) CARDBiomedBench (b) NaturalQuestions (c) NuminaMath1.5

Figure 3: Average performance across all gold positions for each benchmark and gold context size. Metrics are benchmark-specific (BioScore, subEM, math-verify). Higher is better. Error bars indicate 90% confidence intervals. Colors correspond to gold context sizes: small, medium, large. **Across all settings, performance improves monotonically with gold context size.**

4 Analysis of Confounding Factors

Our goal is to isolate the effect of *gold context size* as an independent factor in LLM performance. Therefore, one must rule out any potential confounding factors that may impact the outcome of the findings. Specifically, in this section we study and rule out the following confounds: gold document position in the context window (§4.1), the total number of repetitions of the answer in the context (§4.2), relative length of gold document to the total distractor evidence length (§4.3), total distractor length (§4.4) and domain specificity (§4.5). Some of these factors may conflate size effects, as variations in gold document size inherently shift their values, while others may introduce potential interactions that complicate attribution.

4.1 Gold Document Position

From the results in Figure 4, we observe that **smaller gold documents are hard to find regardless of their position**. Nevertheless, **certain positions amplify the bias against smaller gold documents**. Performance systematically declines when small gold contexts appear later in the input, while large gold contexts are more robust to position (Full results in Appendix B.5).

For instance, in CBB, Gemini-2.0-Flash achieves 94% accuracy when the small gold context is placed at the start of the context window, but only 33% when placed near the end, a 61-point drop. In contrast, the large gold context declines more gradually, from 84% to 65%, demonstrating greater positional resilience. This pattern held across all evaluated models and benchmarks.

Importantly, the positional effect is more pronounced in domain-specific tasks (CBB and NM) than in general knowledge (NQ), suggesting that task type and gold size compound aggregation difficulty.

Smaller Gold Contexts Exhibit Stronger Primacy Bias. We also observe a primacy bias across models: performance is consistently higher when the gold context appears early in the input window.

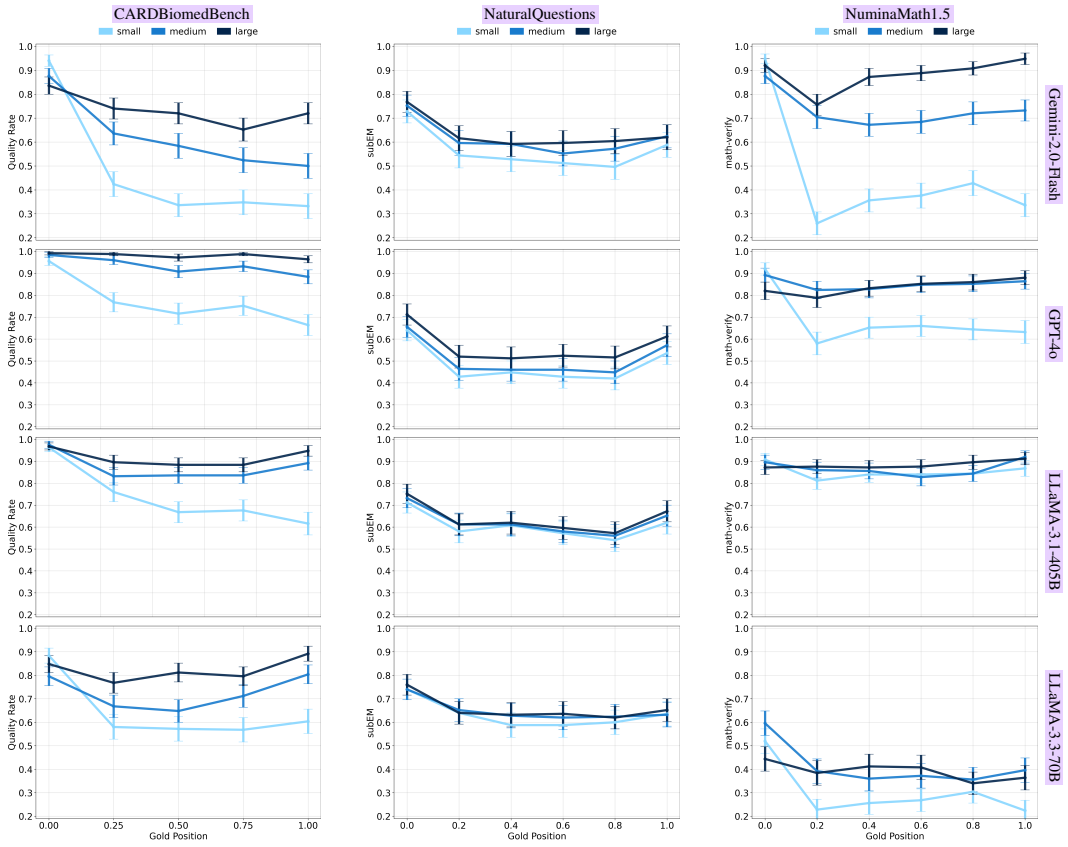


Figure 4: Model performance by gold context position (early to late in input), higher is better and error bars are 90% CIs. Each row is a model, columns are benchmarks. **Smaller gold contexts exhibit sharper performance degradation with later placement, especially in specialized domains (CBB, NM).** Larger contexts mitigate this sensitivity, highlighting the stabilizing effect of richer input.

This effect is especially pronounced for small gold contexts. In some cases, small gold contexts placed at the beginning of the input even outperformed medium or large contexts placed later, despite their reduced information content. This occurs often in the left and right columns of Figure 4, where the small gold line starts at the top at gold position 0.0 before crossing to the bottom.

This inversion highlights the sensitivity of model attention to positional cues when dealing with minimal evidence. While some bias exists for larger contexts, they are substantially more robust to position and do not exhibit the same sharp drop in middle and tail placements.

4.2 Answer Token Repetition

If larger gold documents contained the exact answer span more frequently, this would explain the phenomenon we have observed. The answer would be encoded multiple times and allow the model to attend to it more. Distributions plotted in the Appendix C.2 demonstrate various metrics to measure repetition among small, medium, and large gold documents. For example, given an answer a, a context c, we can compute answer token repetition as:

$$\text{AnsTokRepetition}(a, c) = \frac{1}{|T(a)|} \sum_{t \in T(c)} \mathbf{1}[t \in T(a)], \tag{1}$$

where $T(x)$ are the tokens of x, $|T(a)|$ is the number of unique tokens in a, and $\mathbf{1}[\cdot]$ is the indicator function. Figure 5, which shows binning task performance by similar answer token repetition, shows larger golds are often on top. **Despite the fact that repetition does occur, our claim of small gold documents being harder to find still holds**.

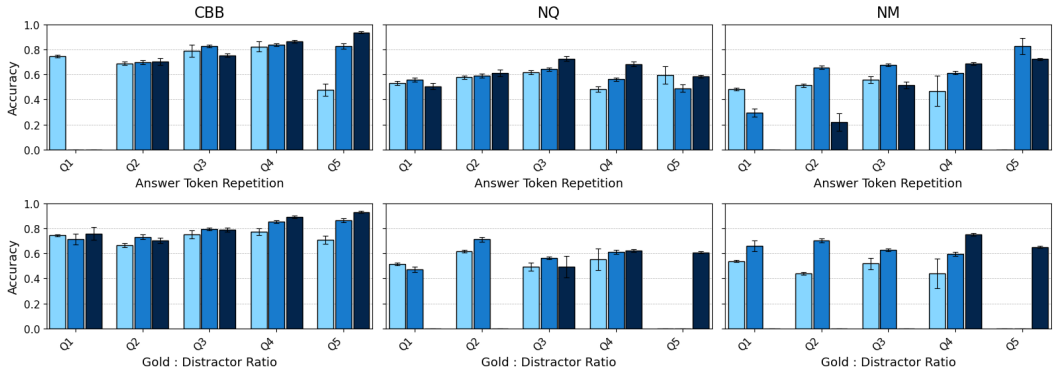


Figure 5: Performance when bucketing tasks into quintiles per confounder (answer token repetition and gold-to-distractor ratio). smaller golds typically yields lower accuracy compared to larger golds. Error bars are 95% confidence intervals, some are larger due to small sample size in that bin.

4.3 Gold-to-Distractor Ratio

Given a fixed amount of distractor documents, varying the size of gold documents will also vary the proportion of gold tokens to distractor tokens within a context window. Given a gold document g and a set of distractor documents D, we can compute:

$$\text{Gold-to-Distractor Ratio}(g, D) = \frac{T(g)}{\sum_{d \in D} T(d)}, \tag{2}$$

Where $T(x)$ is the token count of passage x. This raises the question of if the positive effect of larger gold documents is due to the size itself, or the increased ratio? By grouping the tasks into similar ranges of gold-to-distractor ratio, we can see if size still has an effect when the ratio is held constant. Figure 5 shows just this, and larger golds consistently outperform smaller ones within bins. **Even after controlling for gold-to-distractor ratio, gold context size remains a strong indicator of performance.**

4.4 Distractor Volume

To evaluate the robustness of the gold context size effect under varying degrees of context noise, we systematically increased the number of distractor documents. We leveraged our adaptation of NuminaMath1.5 to run experiments with 5, 10, and 15 distractors, approximately 25k, 50k, and 75k distractor tokens, respectively.

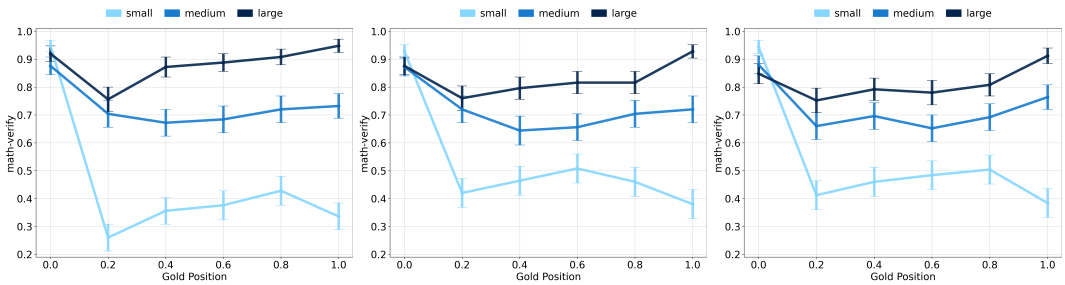
Figure 6 shows that performance is strongly influenced by gold context size, regardless of distractor volume. This reinforces that size remains a dominant variable, even when noise levels change.

4.5 Domain Specificity of Tasks

The effects of gold context size are notably amplified in domain-specific tasks compared to general knowledge. Figure 7 quantifies this by measuring the range in model performance across different gold context positions. For each model and gold size, we compute the performance range as the difference between maximum and minimum scores across all positions:

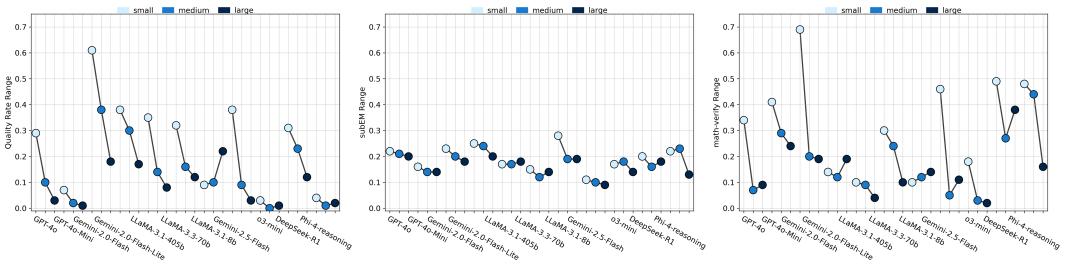
$$\text{Range} = \max_{i \in \{1,\ldots,n\}} \text{perf}(\text{position}_i) - \min_{i \in \{1,\ldots,n\}} \text{perf}(\text{position}_i) \tag{3}$$

For example, on NuminaMath1.5, Gemini-2.0-Flash showed a performance range of 72% for small gold contexts, compared to only 20% for large gold. A similar pattern held in CARDBiomedBench. In contrast, NaturalQuestions exhibited smaller variation across all sizes, likely due to easier questions and higher closed-book baseline scores. This suggests that general knowledge tasks may be inherently more resilient to gold context variability.



(a) 5 Distractors (b) 10 Distractors (c) 15 Distractors

Figure 6: Gemini-2.0-Flash performance on NuminaMath1.5 as the number of distractor documents increases (error bars are 90% CIs). Despite growing distractor noise (up to ∼75k tokens), the performance gap between small and large gold contexts persists. **This confirms that gold context size remains a key factor in long-context reasoning under high-noise conditions.**



(a) CARDBiomedBench (b) NaturalQuestions (c) NuminaMath1.5

Figure 7: Positional sensitivity by benchmark. For each model and gold context size, we compute the range (Equation 3) of performance across positions. **Smaller gold contexts exhibit much higher sensitivity (larger ranges), especially in domain-specific tasks (CBB, NM). Larger gold contexts yield more stable performance across positions.**

Summary. These additional analyses confirm that the observed effects are not artifacts of a single benchmark or setup. Small gold contexts not only reduce performance, but also magnify positional bias. These effects are more severe in noisy environments and domain-specific tasks.

5 Related Work

We review related work in the context of long-context reasoning, focusing on three themes: positional biases in LLMs, long-context evaluation frameworks, and mitigation strategies.

Positional biases in LLMs. Position bias, the tendency of LLMs to over- or under-attend to different parts of the input, has emerged as a fundamental challenge. Prior work has identified several variants: *primacy bias*, where early content is favored (Wang et al., 2023); *recency bias*, where later content dominates (Zheng et al., 2023); and *U-shaped bias*, where mid-context is under-attended (Liu et al.). These effects persist across model architectures, alignment strategies (Liu et al.), extended context lengths (Lee et al., 2024), and, to some extent, in internal representations (Lu et al., 2024). Our work contributes to this literature by introducing a new dimension: we show that *the size of the gold context modulates the strength of positional bias*. Specifically, smaller gold contexts are significantly more vulnerable to primacy effects, while larger contexts confer greater robustness to variation.

The closest work to our setup is (Levy et al.) who study needle-in-a-haystack performance under variable input lengths. While both works investigate positional dynamics in noisy settings, our approach holds the distractor context fixed and instead varies the gold context size, allowing us to isolate the effects of gold signal sparsity. Another related work is by (Dai et al., 2024), who examine how in-context factors, including "data size", affect NIAH performance in synthetic key-value retrieval tasks. Their needle length analysis examines different data subsets, each defined by a specific answer span length. As a result, they effectively evaluate different sets of tasks for each

target length. Whereas we hold the questions and their answers constant and systematically vary the size of the surrounding gold (relevant) context in open-ended QA.

Frameworks for long-context evaluation. Evaluation strategies for long-context reasoning have evolved from synthetic toy tasks to richer, more realistic setups. Long-Range Arena (Tay et al., 2021) introduced standardized tasks for comparing various transformer variants. Recent benchmarks explore broader benchmarking variations (Guan et al., 2022; Hudson & Al Moubayed; An et al.; Bai et al.; Li et al., a; Gao et al.; Li et al., b; Modarressi et al., 2025; Ling et al., 2025; Jacovi et al.; Zhang et al.; Ye et al., 2024; Yen et al., 2024), such as document synthesis (Shaham et al.; 2023), document-level retrieval (Yen et al., 2025), citation verification (Zhang et al., 2024a), and biomedical reasoning (Adams et al., 2024; Cui et al., 2025). Most of these setups use the "needle-in-a-haystack" formulation (Kamradt, 2023; Hsieh et al., 2024a) where a small relevant span must be retrieved from a large set of distractors. Some efforts push beyond this setup, incorporating aggregation, multi-hop inference (Zhuang et al.; Katsis et al.), or mixed-modality inputs (Wu et al.). Our work builds on this direction by adapting natural, domain-specific datasets to simulate realistic multi-agent aggregation within a "needle-in-a-haystack" framework due to its practical relevance.

Mitigation strategies for position bias. Several mitigation approaches have been proposed to reduce position sensitivity in LLMs. These include compressing or abstracting context (Jiang et al., 2024), distilling long-context information into weights (Cao et al., 2025), reweighting attention via calibration (Hsieh et al., 2024b), modifying positional encoding schemes (Zhang et al., 2024b; Zheng et al., 2024), and fine-tuning on debiased distributions (Xiong et al., 2024). While some methods mitigate positional biases, many introduce side effects (Zhao et al., 2024), leaving long-context generalization an ongoing challenge. Our contribution is diagnostic rather than corrective. We uncover a novel interaction between input structure (gold context size) and positional bias severity, showing that simply increasing the amount of gold evidence can systemically impact position bias. Whether existing mitigation strategies can address this effect remains an open question.

6 Discussion, Limitations, and Conclusion

Why does gold context size strongly affect aggregation accuracy? Our findings reveal two interconnected factors: First, we hypothesize that larger gold contexts attract attention by offering a *higher density* of semantically relevant tokens, making them more prominent within distracting content. This richer semantic environment helps models retrieve relevant signals and reduces positional sensitivity. The effect is especially pronounced in domain-specific tasks, where coherent reasoning chains in larger contexts help models follow structured logic needed for accurate answers.

Practical implications of our findings. While prior work has studied factors like positional bias and distractor count, we highlight an overlooked and less controllable factor: gold context size. Therefore, practitioners should recognize that aggregation quality is sensitive to context length variations, even when retrieval mechanisms functions as expected. Practitioners can address this by strategically balancing retrieved document sizes and accounting for potential biases against shorter contexts.

Limitations of our study. First, we did not explicitly control the proportion of gold context within the total context window. Instead, we fixed distractor lengths to better reflect real-world conditions, resulting in varying gold-to-distractor ratios. This may confound whether performance differences stem from gold context size alone or its relative share. Second, while our benchmarks and distractors were curated for realism and domain diversity, only the CBB dataset used a real-world retriever; NQ and NM relied on synthetic setups. Future work should address these.

Conclusion. Our study reveals a fundamental yet previously overlooked limitation in LLM aggregation capabilities: **the size of relevant information critically influences aggregation effectiveness in long-context tasks**. Through systematic evaluation, we demonstrated that smaller gold contexts degrade model performance substantially and exacerbate positional sensitivity, especially in domain-specific tasks. This discovery underscores a crucial vulnerability in real-world agentic deployments, where relevant evidence often appears unpredictably scattered amidst extensive distractors. As language models become central to applications requiring precise and trustworthy reasoning-from scientific discovery to personalized assistants-our findings highlight the urgent need to rethink aggregation strategies. Future LLM-driven systems must explicitly address context-size variability to ensure reliability, safety, and user trust in the face of complex, noisy real-world information streams.

7 Reproducibility Statement

In the Supplementary Material, we have included a folder with our self-contained code and instructions to reproduce all of the experimental results from this paper. Upon publication, we plan to push the code to a public GitHub repository, along with the data used to run the experiments, in order to support reproducible research.

References

Marah Abdin, Sahaj Agarwal, Ahmed Awadallah, Vidhisha Balachandran, Harkirat Behl, Lingjiao Chen, Gustavo de Rosa, Suriya Gunasekar, Mojan Javaheripi, Neel Joshi, Piero Kauffmann, Yash Lara, Caio César Teodoro Mendes, Arindam Mitra, Besmira Nushi, Dimitris Papailiopoulos, Olli Saarikivi, Shital Shah, Vaishnavi Shrivastava, Vibhav Vineet, Yue Wu, Safoora Yousefi, and Guoqing Zheng. Phi-4-reasoning technical report, 2025. URL https://arxiv.org/abs/2504.21318.

Lisa Adams, Felix Busch, Tianyu Han, Jean-Baptiste Excoffier, Matthieu Ortala, Alexander Löser, Hugo JWL. Aerts, Jakob Nikolas Kather, Daniel Truhn, and Keno Bressem. Longhealth: A question answering benchmark with long clinical documents, 2024. URL https://arxiv.org/abs/2401.14490.

Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. L-eval: Instituting standardized evaluation for long context language models. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.776/.

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. Researchagent: Iterative research idea generation over scientific literature with large language models. *arXiv preprint arXiv:2404.07738*, 2024.

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. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. URL https://aclanthology.org/2024.acl-long.172/.

Adib Bazgir, Rama chandra Praneeth Madugula, and Yuwen Zhang. Agentichypothesis: A survey on hypothesis generation using LLM systems. In *Towards Agentic AI for Science: Hypothesis Generation, Comprehension, Quantification, and Validation*, 2025. URL https://openreview.net/forum?id=UeeyfR4CUg.

Owen Bianchi, Maya Willey, Chelsea X Avarado, Benjamin Danek, Marzieh Khani, Nicole Kuznetsov, Anant Dadu, Syed Shah, Mathew J Koretsky, Mary B Makarious, Cory Weller, Kristin S Levine, Sungwon Kim, Paige Jarreau, Dan Vitale, Elise Marsan, Hirotaka Iwaki, Hampton Leonard, Sara Bandres-Ciga, Andrew B Singleton, Mike A. Nalls, Shekoufeh Mokhtari, Daniel Khashabi, and Faraz Faghri. Cardbiomedbench: A benchmark for evaluating large language model performance in biomedical research. *biorxiv preprint 2025.01.15.63327*, abs/2025.01.15.63327, 2025. URL https://pubmed.ncbi.nlm.nih.gov/39868292/.

Egor Bogomolov, Aleksandra Eliseeva, Timur Galimzyanov, Evgeniy Glukhov, Anton Shapkin, Maria Tigina, Yaroslav Golubev, Alexander Kovrigin, Arie van Deursen, Maliheh Izadi, et al. Long code arena: a set of benchmarks for long-context code models. *arXiv preprint arXiv:2406.11612*, 2024.

Bowen Cao, Deng Cai, and Wai Lam. Infiniteicl: Breaking the limit of context window size via long short-term memory transformation. *arXiv preprint arXiv:2504.01707*, 2025.

Hao Cui, Zahra Shamsi, Gowoon Cheon, Xuejian Ma, Shutong Li, Maria Tikhanovskaya, Peter Christian Norgaard, Nayantara Mudur, Martyna Beata Plomecka, Paul Raccuglia, Yasaman Bahri, Victor V. Albert, Pranesh Srinivasan, Haining Pan, Philippe Faist, Brian A Rohr, Michael J. Statt,

Dan Morris, Drew Purves, Elise Kleeman, Ruth Alcantara, Matthew Abraham, Muqthar Mohammad, Ean Phing VanLee, Chenfei Jiang, Elizabeth Dorfman, Eun-Ah Kim, Michael Brenner, Sameera S Ponda, and Subhashini Venugopalan. CURIE: Evaluating LLMs on multitask scientific long-context understanding and reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=jw2fC6REUB.

Hui Dai, Dan Pechi, Xinyi Yang, Garvit Banga, and Raghav Mantri. Deniahl: In-context features influence llm needle-in-a-haystack abilities, 2024. URL https://arxiv.org/abs/2411.19360.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri,

Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan

Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Hugging Face. Math-verify: A robust mathematical expression evaluation system, 2025. URL https://github.com/huggingface/Math-Verify.

Center for Alzheimer's and Related Dementias (CARD). Biomedsql. https://huggingface.co/datasets/NIH-CARD/BiomedSQL, 2025. Accessed: 2025-05-14.

Muhan Gao, Jash Shah, Weiqi Wang, and Daniel Khashabi. Science hierarchography: Hierarchical abstractions of scientific literature. *arXiv preprint arXiv:2504.13834*, 2025. URL https://arxiv.org/abs/2504.13834.

Yunfan Gao, Yun Xiong, Wenlong Wu, Zijing Huang, Bohan Li, and Haofen Wang. U-niah: Unified rag and llm evaluation for long context needle-in-a-haystack. URL https://arxiv.org/abs/2503.00353.

Jian Guan, Zhuoer Feng, Yamei Chen, Ruilin He, Xiaoxi Mao, Changjie Fan, and Minlie Huang. LOT: A story-centric benchmark for evaluating Chinese long text understanding and generation. *Transactions of the Association for Computational Linguistics*, 2022.

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

Cheng-Yu Hsieh, Yung-Sung Chuang, Chun-Liang Li, Zifeng Wang, Long Le, Abhishek Kumar, James Glass, Alexander Ratner, Chen-Yu Lee, Ranjay Krishna, et al. Found in the middle: Calibrating positional attention bias improves long context utilization. In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 14982–14995, 2024b.

George Hudson and Noura Al Moubayed. MuLD: The multitask long document benchmark. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. URL https://aclanthology.org/2022.lrec-1.392/.

Alon Jacovi, Andrew Wang, Chris Alberti, Connie Tao, Jon Lipovetz, Kate Olszewska, Lukas Haas, Michelle Liu, Nate Keating, Adam Bloniarz, Carl Saroufim, Corey Fry, Dror Marcus, Doron Kukliansky, Gaurav Singh Tomar, James Swirhun, Jinwei Xing, Lily Wang, Madhu Gurumurthy, Michael Aaron, Moran Ambar, Rachana Fellinger, Rui Wang, Zizhao Zhang, Sasha Goldshtein, and Dipanjan Das. The facts grounding leaderboard: Benchmarking llms' ability to ground responses to long-form input. URL https://arxiv.org/abs/2501.03200.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. Longllmlingua: Accelerating and enhancing llms in long context scenarios via prompt compression. In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 1658–1677, 2024.

Greg Kamradt. Needle in a haystack - pressure testing llms. *GitHub*, 2023. URL https://github.com/gkamradt/LLMTest_NeedleInAHaystack.

Yannis Katsis, Sara Rosenthal, Kshitij Fadnis, Chulaka Gunasekara, Young-Suk Lee, Lucian Popa, Vraj Shah, Huaiyu Zhu, Danish Contractor, and Marina Danilevsky. Mtrag: A multi-turn conversational benchmark for evaluating retrieval-augmented generation systems. URL https://arxiv.org/abs/2501.03468.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

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.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. Same task, more tokens: the impact of input length on the reasoning performance of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. URL https://aclanthology.org/2024.acl-long.818/.

Jia LI, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang, Kashif Rasul, Longhui Yu, Albert Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann Fleureau, Guillaume Lample, and Stanislas Polu. Numinamath. [https://huggingface.co/AI-MO/NuminaMath-1.5](https://github.com/project-numina/aimo-progress-prize/blob/main/report/numina_dataset.pdf), 2024.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. LooGLE: Can long-context language models understand long contexts? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, a. URL https://aclanthology.org/2024.acl-long.859/.

Kuan Li, Liwen Zhang, Yong Jiang, Pengjun Xie, Fei Huang, Shuai Wang, and Minhao Cheng. Lara: Benchmarking retrieval-augmented generation and long-context llms – no silver bullet for lc or rag routing, b. URL https://arxiv.org/abs/2502.09977.

Zhan Ling, Kang Liu, Kai Yan, Yifan Yang, Weijian Lin, Ting-Han Fan, Lingfeng Shen, Zhengyin Du, and Jiecao Chen. Longreason: A synthetic long-context reasoning benchmark via context expansion, 2025. URL https://arxiv.org/abs/2501.15089.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*. URL https://aclanthology.org/2024.tacl-1.9/.

Tianyang Liu, Canwen Xu, and Julian McAuley. Repobench: Benchmarking repository-level code auto-completion systems. *arXiv preprint arXiv:2306.03091*, 2023.

Taiming Lu, Muhan Gao, Kuai Yu, Adam Byerly, and Daniel Khashabi. Insights into llm long-context failures: When transformers know but don't tell. *arXiv preprint arXiv:2406.14673*, 2024.

Sean MacAvaney, Andrew Yates, Sergey Feldman, Doug Downey, Arman Cohan, and Nazli Goharian. Simplified data wrangling with ir_datasets. In *SIGIR*, 2021.

Shrestha Basu Mallick and Logan Kilpatrick. Gemini 2.0: Flash, flash-lite and pro. https://developers.google.com/updates/gemini-2-0-flash-flash-lite-pro, February 2025. Google for Developers Blog, February 5, 2025.

Ali Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Trung Bui, Ryan A. Rossi, Seunghyun Yoon, and Hinrich Schütze. Nolima: Long-context evaluation beyond literal matching, 2025. URL https://arxiv.org/abs/2502.05167.

National Center for Biotechnology Information (NCBI). Ncbi [internet]. https://www.ncbi.nlm.nih.gov/, 1988. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988]– [cited 2025 May 20].

NIH Biowulf. Nih high-performance computing (hpc) biowulf cluster. https://hpc.nih.gov, 2024. Accessed May 2025.

OpenAI. tiktoken: A fast bpe tokenizer for use with openai's models. https://github.com/openai/tiktoken, 2023. Accessed: 2025-05-15.

OpenAI. Gpt-4o mini: advancing cost-efficient intelligence. https://openai.com/research/gpt-4o-mini-advancing-cost-efficient-intelligence, July 2024. https://openai.com/research/gpt-4o-mini-advancing-cost-efficient-intelligence.

OpenAI. Introducing openai o3 and o4-mini. https://openai.com/index/introducing-o3-and-o4-mini/, April 2025. Accessed: September 25, 2025.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick

Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. Kilt: a benchmark for knowledge intensive language tasks, 2021. URL https://arxiv.org/abs/2009.02252.

Open R1. Openr1-math-220k. https://huggingface.co/datasets/open-r1/OpenR1-Math-220k, 2024. Accessed: 2025-05-15.

Uri Shaham, Elad Segal, Maor Ivgi, Avia Efrat, Ori Yoran, Adi Haviv, Ankit Gupta, Wenhan Xiong, Mor Geva, Jonathan Berant, and Omer Levy. SCROLLS: Standardized CompaRison over long language sequences. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. URL https://aclanthology.org/2022.emnlp-main.823/.

Uri Shaham, Maor Ivgi, Avia Efrat, Jonathan Berant, and Omer Levy. Zeroscrolls: A zero-shot benchmark for long text understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 7977–7989, 2023.

Henry W Sprueill, Carl Edwards, Khushbu Agarwal, Mariefel V Olarte, Udishnu Sanyal, Conrad Johnston, Hongbin Liu, Heng Ji, and Sutanay Choudhury. Chemreasoner: Heuristic search over a large language model's knowledge space using quantum-chemical feedback. *arXiv preprint arXiv:2402.10980*, 2024.

Peter D. Stenson, Matthew Mort, Edward V. Ball, Molly Chapman, Katy Evans, Luisa Azevedo, Matthew Hayden, Sally Heywood, David S. Millar, Andrew D. Phillips, and David N. Cooper. The human gene mutation database (hgmd®): optimizing its use in a clinical diagnostic or research setting. *Human Genetics*, 139(10):1197–1207, 2020. doi: 10.1007/s00439-020-02199-3. URL https://doi.org/10.1007/s00439-020-02199-3.

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. Long range arena : A benchmark for efficient transformers. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=qVyeW-grC2k.

Weiqi Wang, Jiefu Ou, Yangqiu Song, Benjamin Van Durme, and Daniel Khashabi. Can llms generate tabular summaries of science papers? rethinking the evaluation protocol. *arXiv preprint arXiv:2504.10284*, 2025. URL https://arxiv.org/abs/2504.10284.

Yiwei Wang, Yujun Cai, Muhao Chen, Yuxuan Liang, and Bryan Hooi. Primacy effect of chatgpt. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 108–115, 2023.

Chih-Hsuan Wei, Alexis Allot, Po-Ting Lai, Robert Leaman, Shubo Tian, Ling Luo, Qiao Jin, Zhizheng Wang, Qingyu Chen, and Zhiyong Lu. Pubtator 3.0: an ai-powered literature resource for

unlocking biomedical knowledge. *Nucleic Acids Research*, 52(W1):W540–W546, 04 2024. ISSN 0305-1048. doi: 10.1093/nar/gkae235. URL https://doi.org/10.1093/nar/gkae235.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. Longmemeval: Benchmarking chat assistants on long-term interactive memory. In *The Thirteenth International Conference on Learning Representations*. URL https://openreview.net/forum?id=pZiyCaVuti.

Oskar Wysocki, Magdalena.wysocka@cruk.manchester.ac.uk Magdalena.wysocka@cruk.manchester.ac.uk, Danilo Carvalho, Alex Bogatu, Danilo.miranda@idiap.ch Danilo.miranda@idiap.ch, Maxime.delmas@idiap.ch Maxime.delmas@idiap.ch, Harriet.unsworth@cruk.manchester.ac.uk Harriet.unsworth@cruk.manchester.ac.uk, and Andre Freitas. An LLM-based knowledge synthesis and scientific reasoning framework for biomedical discovery. In Yixin Cao, Yang Feng, and Deyi Xiong (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-demos.34/.

Zheyang Xiong, Vasilis Papageorgiou, Kangwook Lee, and Dimitris Papailiopoulos. From artificial needles to real haystacks: Improving retrieval capabilities in llms by finetuning on synthetic data. *arXiv preprint arXiv:2406.19292*, 2024.

Xiao Ye, Andrew Wang, Jacob Choi, Yining Lu, Shreya Sharma, Lingfeng Shen, Vijay Tiyyala, Nicholas Andrews, and Daniel Khashabi. AnaloBench: benchmarking the identification of abstract and long-context analogies. In *Conference on Empirical Methods in Natural Language Processing* (EMNLP), 2024. URL https://arxiv.org/abs/2402.12370.

Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izasak, Moshe Wasserblat, and Danqi Chen. Helmet: How to evaluate long-context language models effectively and thoroughly. *arXiv preprint arXiv:2410.02694*, 2024.

Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. Helmet: How to evaluate long-context language models effectively and thoroughly, 2025. URL https://arxiv.org/abs/2410.02694.

Fengji Zhang, Bei Chen, Yue Zhang, Jacky Keung, Jin Liu, Daoguang Zan, Yi Mao, Jian-Guang Lou, and Weizhu Chen. Repocoder: Repository-level code completion through iterative retrieval and generation. *arXiv preprint arXiv:2303.12570*, 2023.

Jiajie Zhang, Yushi Bai, Xin Lv, Wanjun Gu, Danqing Liu, Minhao Zou, Shulin Cao, Lei Hou, Yuxiao Dong, Ling Feng, et al. Longcite: Enabling llms to generate fine-grained citations in long-context qa. *arXiv preprint arXiv:2409.02897*, 2024a.

Yuxiang Zhang, Jing Chen, Junjie Wang, Yaxin Liu, Cheng Yang, Chufan Shi, Xinyu Zhu, Zihao Lin, Hanwen Wan, Yujiu Yang, Tetsuya Sakai, Tian Feng, and Hayato Yamana. ToolBeHonest: A multi-level hallucination diagnostic benchmark for tool-augmented large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. URL https://aclanthology.org/2024.emnlp-main.637/.

Zhenyu Zhang, Runjin Chen, Shiwei Liu, Zhewei Yao, Olatunji Ruwase, Beidi Chen, Xiaoxia Wu, and Zhangyang Wang. Found in the middle: How language models use long contexts better via plug-and-play positional encoding. *arXiv preprint arXiv:2403.04797*, 2024b.

Xinyu Zhao, Fangcong Yin, and Greg Durrett. Understanding synthetic context extension via retrieval heads. *arXiv preprint arXiv:2410.22316*, 2024.

Chuanyang Zheng, Yihang Gao, Han Shi, Minbin Huang, Jingyao Li, Jing Xiong, Xiaozhe Ren, Michael Ng, Xin Jiang, Zhenguo Li, et al. Dape: Data-adaptive positional encoding for length extrapolation. *Advances in Neural Information Processing Systems*, 37:26659–26700, 2024.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*, 2023.

Tianyi Zhuang, Chuqiao Kuang, Xiaoguang Li, Yihua Teng, Jihao Wu, Yasheng Wang, and Lifeng Shang. Docpuzzle: A process-aware benchmark for evaluating realistic long-context reasoning capabilities. URL https://arxiv.org/abs/2502.17807.

A Appendix

A Extended Experimental Details

We provide extended experimental details on benchmark construction, model configuration, and evaluation methodology to support the reproducibility and interpretability of our results.

A.1 Original Benchmarks

We describe the sources, licenses, and preprocessing procedures for each of the three adapted benchmarks used in our experiments. All experiments were run on a sampled subset of 250 examples per benchmark. See the code repository for exact methodology.

CARDBiomedBench

- **Source:** CARDBiomedBench on Hugging Face and its BiomedSQL variant on Hugging Face. Distractor documents were retrieved using a multi-agent retrieval system at NIH, which retrieves content from: (1) Google search over NIH domains, (2) PubTator3.0 (Wei et al., 2024), (3) the Human Gene Mutation Database (HGMD) (Stenson et al., 2020), and (4) NCBI gene and variant pages (National Center for Biotechnology Information (NCBI), 1988).
- **License:** Apache 2.0 for benchmark code and data. Some distractor sources (e.g., HGMD) are not redistributable but are publicly accessible on their respective platforms.
- **Preprocessing:** None, the distractor and gold documents are as-is from the retriever.

NaturalQuestions

- **Source:** NQ with evidence spans aligned to Knowledge Intensive Language Tasks (KILT) on Hugging Face. Gold documents were loaded using the Ai2 ir_datasets python package (MacAvaney et al., 2021) and distractors were sourced from HELMET on Hugging Face.
- **License:** Creative Commons Share-Alike 3.0 (NQ), MIT (KILT & HELMET), and Apache 2.0 (ir_datasets).
- **Preprocessing:** We filtered for validation examples that had matching HELMET distractors. Examples with missing KILT provenance, absent or unresolvable answer spans, or malformed metadata were excluded. Gold and distractor documents included the title of the article *'Title: {title} Document: {gold_document}'* to give them context.

NuminaMath1.5

- **Source:** NuminaMath1.5 (NM) and its OpenR1Math (OR1M) variant on Hugging Face, that contains DeepSeekR1 reasoning chains.
- **License:** Apache 2.0. (NM and OR1M).
- **Preprocessing:** Filtered to retain only examples with 'complete' and 'verified' fields for question, final answer, structured solution, and long-form generation. DeepSeekR1 generations were truncated to the final 5,000 tokens using GPT-4o tiktoken (OpenAI, 2023) tokenization to normalize document length across tasks. Distractors sampling was among the other questions and excluded duplicates. Sizes of gold and distractors were strung into a pseudo-document by including *'The answer to {question} is {gold_document}'* to give them context.

	CARDBiomedBench (CBB)	NaturalQuestions (NQ)	NuminaMath1.5 (NM)
Question	What is the genomic location of rs12255438 in the GRCh38/hg38 build of the human genome and what gene is it located on or near?	Who is playing the halftime show at super bowl 2016?	A ship traveling along a river has covered 24 km upstream and 28 km downstream... Determine the speed of the ship in still water and the speed of the river.
Small Gold	The SNP rs12255438 is located on or closest to the gene CTNNA3 on chromosome 10 at base pair position 66465707 in the GRCh38/hg38 build of the human genome.	Super Bowl 50 halftime show It was headlined by the British rock group Coldplay with special guest performers Beyoncé and Bruno Mars, who previously had headlined the Super Bowl XLVII and Super Bowl XLVIII halftime shows, respectively.	The answer to the question "A ship traveling along a river has covered 24 km ..." is: v_{R}=4\mathrm{~}/\mathrm{}, v_{B}=10\mathrm{~}/\mathrm{}
Medium Gold	`SELECT 'AlzheimerDisease_GeneData' AS source_table, UUID, SNP, chr_38, bp_38, nearestGene ... WHERE SNP = 'rs12255438' LIMIT 100` The SNP rs12255438 is located on or closest to the gene CTNNA3 on chromosome 10 at base pair position 66465707 in the GRCh38/hg38 build of the human genome.	Super Bowl 50 halftime show The Super Bowl 50 Halftime Show took place on February 7, 2016, at Levi's Stadium in Santa Clara, California as part of Super Bowl 50. It was headlined by the British rock group Coldplay with special guest performers Beyoncé and Bruno Mars, who previously had headlined the Super Bowl XLVII and Super Bowl XLVIII halftime shows, respectively.	Let t be the time required for the boat to travel $24 \mathrm{~}km}$ upstream and $28 \mathrm{~}km}$ downstream, v_{R} the speed of the river, and v_{B} the speed of the boat. When the boat is traveling upstream, its speed is $v_{B}-v_{R}$, and when it is traveling downstream, its speed is $v_{B}+v_{R}$. ... The speed of the river is $v_{R}=4 \mathrm{~}km} / \mathrm{h}$, and the speed of the boat is $v_{B}=10 \mathrm{~}km} / \mathrm{h}$.

Large Gold	`SELECT 'AlzheimerDisease_GeneData' AS source_table, UUID, SNP, chr_38, bp_38, nearestGene ... WHERE SNP = 'rs12255438' LIMIT 100`
	`[{'SNP': 'rs12255438', 'chr_38': 10, 'bp_38': 66465707, 'nearestGene': 'CTNNA3'}, ... {'SNP': 'rs12255438', 'chr_38': 10, 'bp_38': 66465707, 'nearestGene': 'CTNNA3'}]`
	The SNP rs12255438 is located on or closest to the gene CTNNA3 on chromosome 10 at base pair position 66465707 in the GRCh38/hg38 build of the human genome.
		Super Bowl 50 halftime show
		The Super Bowl 50 Halftime Show took place on February 7, 2016, at Levi's Stadium in Santa Clara, California as part of Super Bowl 50. It was headlined by the British rock group Coldplay with special guest performers Beyoncé and Bruno Mars, who previously had headlined the Super Bowl XLVII and Super Bowl XLVIII halftime shows, respectively.
		...
		At that time, Mars and Beyoncé were both doing a diet and stressing out. One day before the performance they were "watching playback backstage", while Beyonce ate a bag of Cheetos.
		... (+5 more paragraphs)
			<think>
			Okay, so I need to find the speed of the ship in still water and the speed of the river. Let me start by recalling that when a ship is moving upstream, its effective speed is the speed of the ship minus the speed of the river.
			...
			Wait, actually, the problem states: "For this journey, it took half an hour less than for traveling 30 km upstream ... Hmm, let me parse that again...
			...
			the final answer is v_{R}=4\mathrm{~}/\mathrm{},v_{B}=10\mathrm{~}/\mathrm{}

Figure 8: Gold context construction across benchmarks. The "small" gold context is minimally sufficient to answer the question; "medium" and "large" add further relevant information. In CARDBiomedBench (left), this includes SQL and result rows; in NQ (center), adjacent Wikipedia paragraphs; in NM (right), full solution traces and DeepSeekR1 reasoning chain.

A.2 Task Creation

A.3 LLM Configuration

We evaluated seven LLMs, each configured via provider-specific APIs. All evaluations were conducted as deterministically as possible.

API Providers. We used the following service providers for model access:

- **GPT models** (o3-mini, GPT-4o, GPT-4o-mini) were accessed via the `Azure OpenAI` service.
- **Gemini models** (Gemini-2.5-Flash, Gemini-2.0-Flash, Gemini-2.0-Flash-Lite) were accessed via the `Google AI GenAI SDK`, using the official `genai` Python client.
- **DeepSeek-R1 and Phi-4-reasoning** were accessed via the `Azure AI Inference` service.
- **LLaMA models >= 70b params** (Meta-LLaMA-3.1-405B-Instruct, LLaMA-3.3-70B-Instruct) were accessed via the `Azure AI Inference` service.
- **LLaMA model < 70B parameters** (LLaMA-3.1-8B-Instruct) was evaluated locally using the `meta-llama/Llama-3.1-8B-Instruct` checkpoint, loaded via Hugging Face `transformers`. All local evaluations were conducted on the NIH High-Performance Computing (HPC) Biowulf cluster (NIH Biowulf, 2024), leveraging GPU nodes for inference.

Prompting and Evaluation Configuration. Prompts were benchmark-specific and standardized across model types. All non-reasoning models were queried with `max_tokens=256` and

`temperature=0.0`. Provider-specific configurations (e.g., safety settings for Google GenAI, and device mapping for HuggingFace) were handled automatically during model initialization. See the code and YAML config files for full details. Reasoning models were queried with `max_tokens=2048` and their default generation params, to allow for reasoning. Reasoning models were additionally given instructions to encourage grounding their answer in the retrieved documents, to prevent relying on internal knowledge.

Grading LLMs. For CARDBiomedBench, an additional grading LLM was used to assess answer correctness via BioScore using GPT-4o, as done by the authors. It was instantiated using the same infrastructure and configurations as the primary LLMs, with `max_tokens=10`.

A.4 METRICS

We used evaluation metrics that align with the original datasets' scoring protocols:

Quality Rate. We evaluate responses to the CBB tasks following their proposed **BioScore** framework, an LLM-as-a-judge metric implemented with GPT-4o. Each response is scored on a 3-point scale according to the BioScore prompt 4, and a score ≥ 2 is considered factually correct. The **Quality Rate** is computed as the proportion of responses meeting this threshold.

Formally, given a reference set $Resp$ of expert-annotated responses and a corresponding set $\hat{Resp}$ of model-generated responses for n questions:

$$\text{Quality Rate} = \frac{1}{N}\sum_{n=1}^{N} Correct(r_n, \hat{r}_n) \tag{4}$$

$$\text{where} \quad Correct(r_n, \hat{r}_n) = \begin{cases} 1, & \text{if BioScore}(r_n, \hat{r}_n) \geq 2 \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad r_n \in Resp, \hat{r}_n \in \hat{Resp} \tag{5}$$

SubEM. For NQ we utilized substring exact match, which assigns a score of 1.0 if any normalized ground truth string is a subspan of the model's response (after normalization), and 0 otherwise. This is a correctness signal used by previous work on this data.

math-verify. Evaluated with math-verify (Face, 2025), a symbolic equivalence checker that parses LaTeX boxed answers and verifies correctness through structured math expression comparison. Parsing and verification are done using an extraction and comparison pipeline derived from the Math-Verify toolkit.

Error Bars. All plots showing aggregate scores (e.g., Figure 3) report 90% confidence intervals (CIs) estimated via non-parametric bootstrapping over tasks. Given N scores, we resample with replacement 1,000 times and compute the middle 90% interval from the resulting bootstrap distribution.

A.5 PROMPTS.

We show prompts used to collect results from the models and the BioScore grading prompt. There is a unique prompt for each benchmark, which is used on every model. {Variables} are in curly braces which are formatted with task data (question and documents). We encourage models to ground their answers in the context and abstain if unable to answer.

```
You are a highly knowledgeable and experienced expert in the healthcare and biomedical field,
possessing extensive medical knowledge and practical expertise. Create an answer to the question
using only the provided documents (some of which might be irrelevant). If you cannot answer the
question based on the documents, explicitly state that you do not know.
Question: {question}
Documents: {documents}
```

Prompt 1: The CARDBiomedBench prompt is adapted from the original paper's experimental methods and includes mention of biomedical expertise.

```
Create an answer to the question using only the provided documents (some of which might be
irrelevant). If you cannot answer the question based on the documents, explicitly state that
you do not know.
Question: {question}
Documents: {documents}
```

Prompt 2: The NaturalQuestions prompt is adapted from previous work's experimental methods (Liu et al.; Yen et al., 2025).

```
Create an ANSWER to the QUESTION using only the provided DOCUMENTS (some of which might be
irrelevant). Write nothing but your final answer in LaTeX within \\boxed{}. If you do not
know the answer to a question, explicitly state so in \\boxed{I don't know}.
QUESTION: {question}
DOCUMENTS: {documents}
QUESTION: {question}
ANSWER:
```

Prompt 3: The NuminaMath1.5 prompt uniquely repeats the question and has formatting guidelines, to comply with the math-verify metric. Without repeating the question models exhibited extremely poor performance in every configuration.

```
You are a highly knowledgeable and experienced expert in the healthcare and biomedical field, possessing
extensive medical knowledge and practical expertise.
### Scoring Instructions for Evaluating Analyst Responses

**Objective:** Evaluate an analyst's response against a gold standard.

**Scoring Criteria:**
    - **Exact Match:** 3 points for an exact or equally accurate response.
    - **Close Match:** 2 points for a very close response with minor inaccuracies.
    - **Partial Match:** 1 point for a partially accurate response with significant omissions.
    - **Irrelevant Information (Harmless):** Deduct 0.5 points for harmless irrelevant information.
    - **Irrelevant Information (Distracting):** Deduct 1 point for distracting irrelevant information.
    - **No Match:** 0 points for no match.
    - **Not Knowing Response:** -1 point for stating lack of knowledge or abstaining. An example
    of this scenario is when Analyst Response says \'There are various studies, resources or
    databases on this topic that you can check ... but I do not have enough information on this topic.

**Scoring Process:**
    1. **Maximum Score:** 3 points per question.
    2. **Calculate Score:** Apply criteria to evaluate the response.

**Question:** {question}
**Golden Answer:** {gold_ans}
**Analyst Response:** {pred_ans}

## Your grading
Using the scoring instructions above, grade the Analyst Response return only the numeric score
on a scale from 0.0-3.0. If the response is stating lack of knowledge or abstaining, give it
-1.0.
```

Prompt 4: BioScore grading prompt for LLM-as-a-judge on CBB tasks, awarding points for correct information and deducting points for incorrect information. It differentiates an abstention (-1) from an incorrect answer (0 or 1).

B Extended Results

We provide extended baselines for all benchmarks in Figure 9 and main results for CBB in Figure 10, for NQ in Figure 11, and for NM in Figure 12. Additionally, we provide positional curves for all models in Figure 14.

B.1 BASELINES

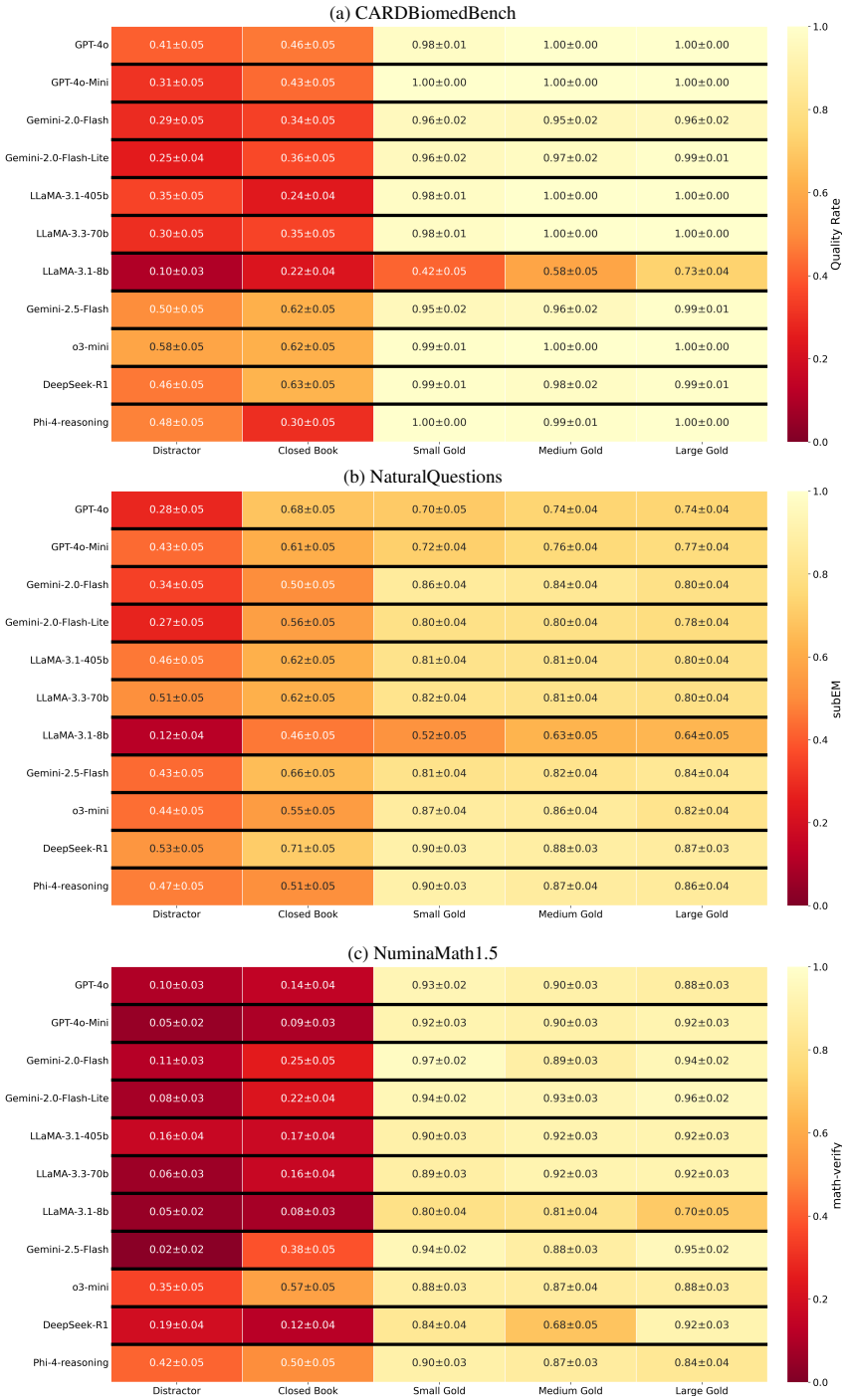


Figure 9: Baseline performance when viewing distractors only, closed book (no documents), and varying sizes of gold. This confirms both (1) models perform poorly without the gold documents and (2) performance is near perfect when viewing any size of gold document.

B.2 CARDBIOMEDBENCH

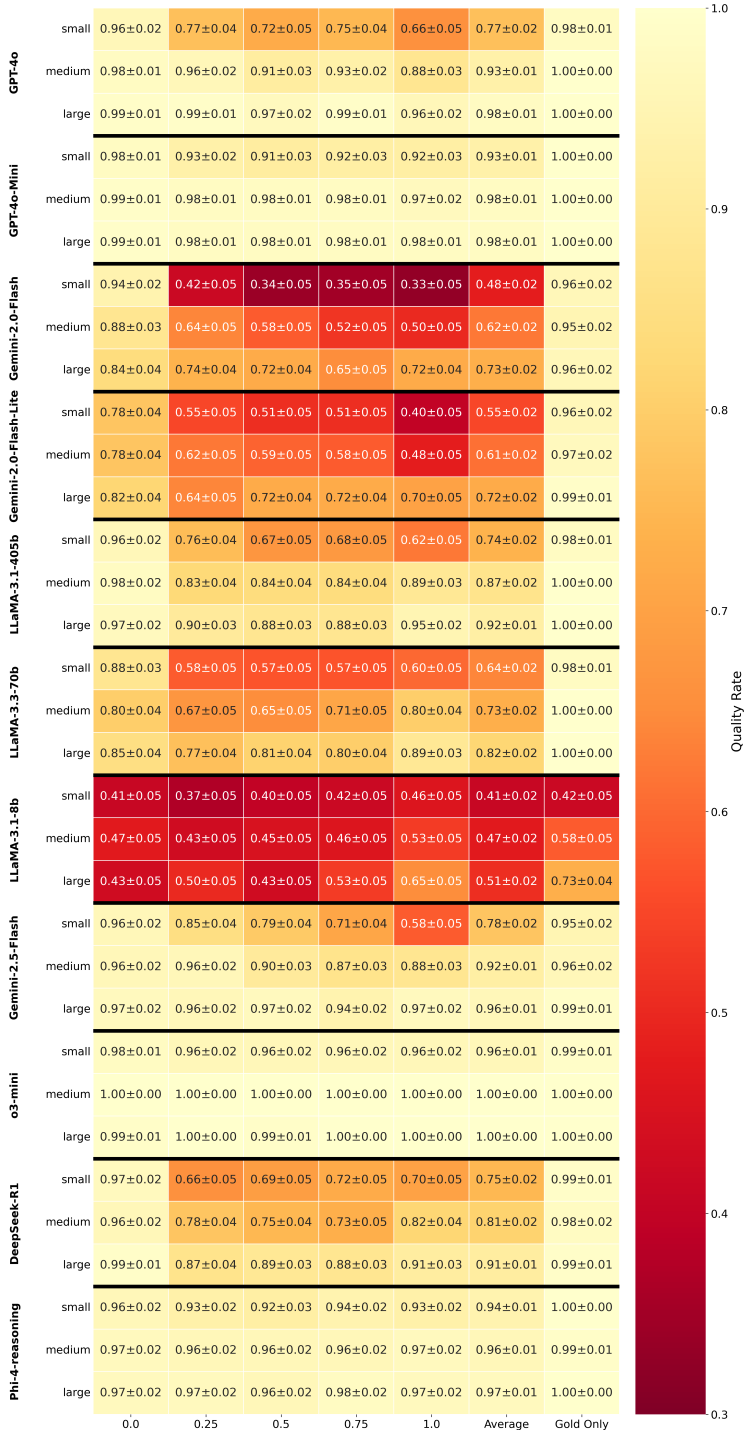


Figure 10: CARDBiomedBench performance for each model and size of gold for varying positions in the context window (0.0, 0.25, 0.5, 0.75, 1.0), the average across all positions, and baseline performance when seeing gold only. Higher scores (light yellow) is more desirable than low scores (dark red), 90% CI are reported.

B.3 NATURALQUESTIONS

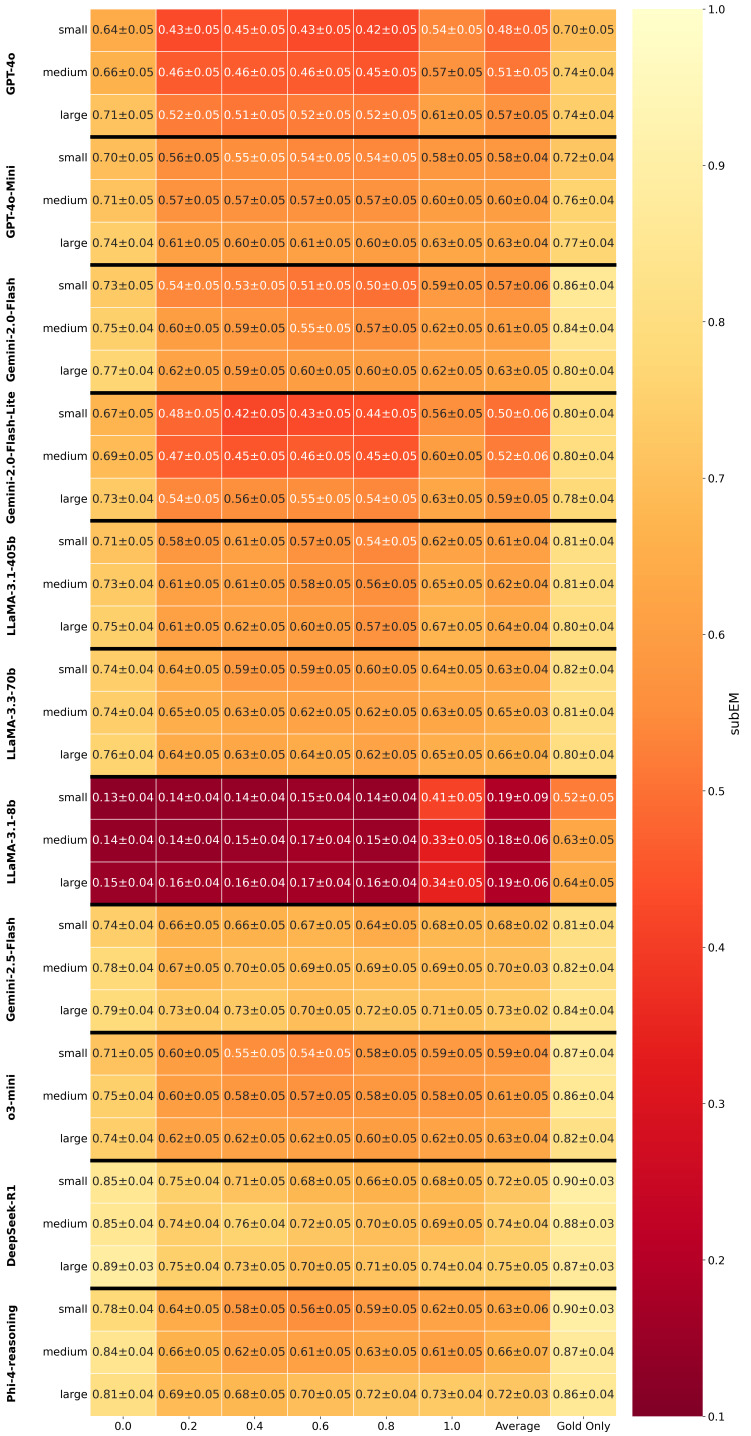


Figure 11: NaturalQuestions performance for each model and size of gold for varying positions in the context window (0.0, 0.2, 0.4, 0.6, 0.8, 1.0), the average across all positions, and baseline performance when seeing gold only. Higher scores (light yellow) are more desirable than low scores (dark red), 90% CI are reported.

B.4 NUMINAMATH1.5

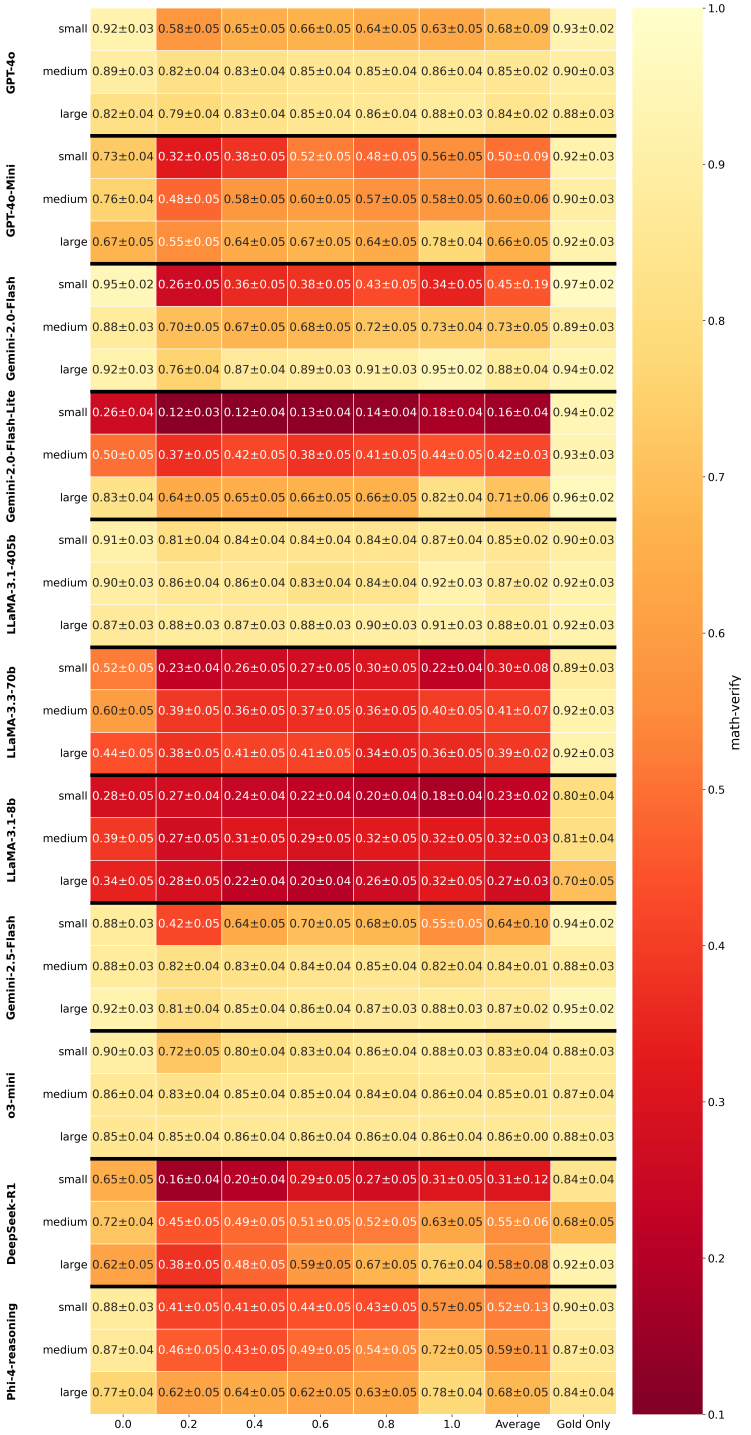


Figure 12: NuminaMath1.5 performance for each model and size of gold for varying positions in the context window (0.0, 0.2, 0.4, 0.6, 0.8, 1.0), the average across all positions, and baseline performance when seeing gold only. Higher scores (light yellow) are more desirable than low scores (dark red), 90% CI are reported.

B.5 PERFORMANCE BY POSITION

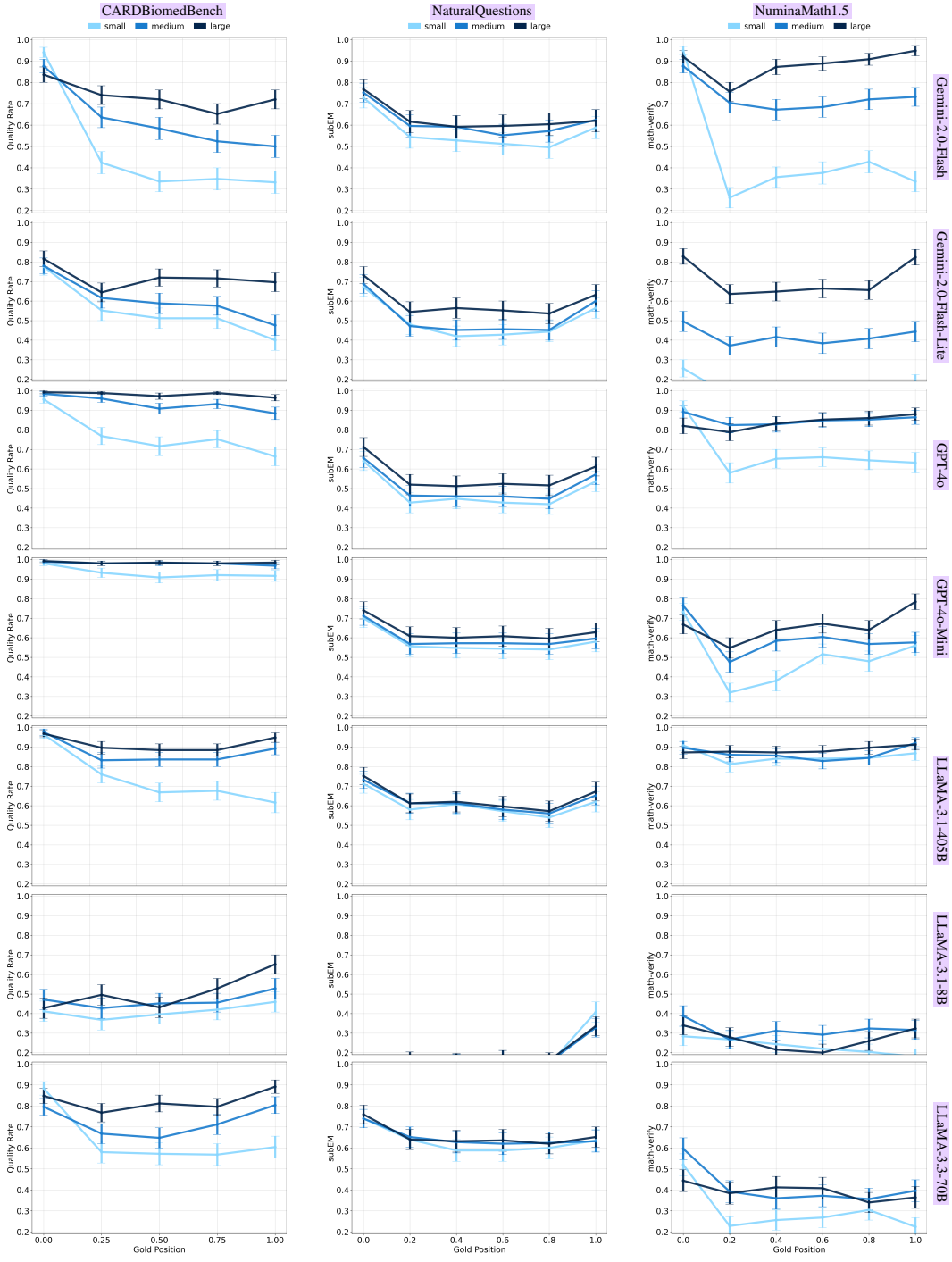


Figure 13: Model performance by gold context position (early to late in input), higher is better and error bars are 90% CIs. Each row is a model, columns are benchmarks. **Smaller gold contexts exhibit sharper performance degradation with later placement, especially in specialized domains (CBB, NM).** Larger contexts mitigate this sensitivity, highlighting the stabilizing effect of richer input. All non-reasoning models, including the ones in Figure 4, are here for comparison.

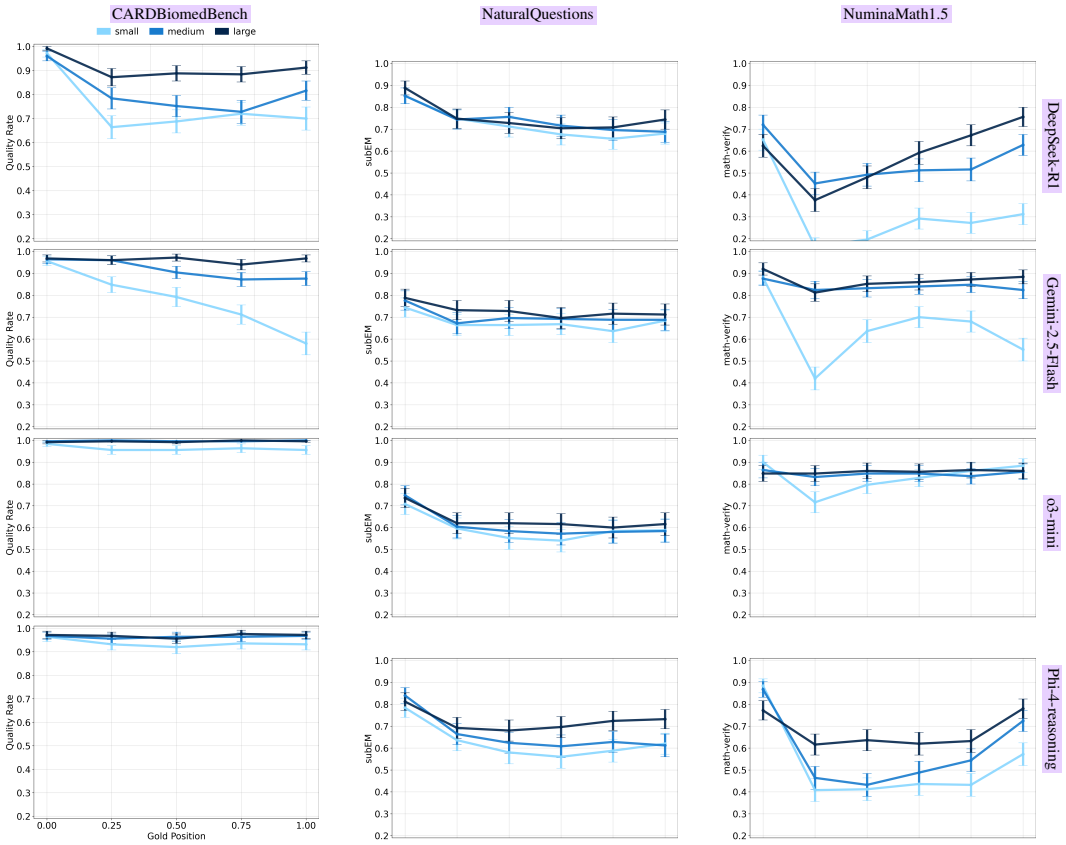


Figure 14: Reasoning model performance by gold context position (early to late in input), higher is better and error bars are 90% CIs. Each row is a model, columns are benchmarks. **Smaller gold contexts exhibit sharper performance degradation with later placement, especially in specialized domains (CBB, NM).** Larger contexts mitigate this sensitivity, highlighting the stabilizing effect of richer input.

C CONFOUNDER ANALYSIS

We provide details, formulas, distributions, and performance across benchmarks when considering the potential confounding variables.

C.1 MEASURING GOLD CONTEXT RATIOS AND ANSWER OVERLAP

We define several metrics to quantify the repetition of the answer across gold passages, as well as the gold-to-distractor ratio.

Gold-to-Distractor Ratios. To measure the ratio of gold to distractor tokens, we define $T(x)$ as the token count of passage x:

$$\text{Gold-to-Distractor Ratio}(g, D) = \frac{T(g)}{\sum_{d \in D} T(d)}, \tag{6}$$

Exact Mentions. We count exact string occurrences of the answer in the context, case-insensitive and word-bounded:

$$\text{ExactMentions}(a, c) = \sum_{a_i \in \mathcal{A}} \#\{\text{occurrences of } a_i \text{ in } c\}, \tag{7}$$

where $\mathcal{A}$ is the set of provided answer strings and c is the context.

Answer Token Hits. At the token level, we measure how many context tokens match any token from the answer:

$$\text{AnsTokHits}(a, c) = \sum_{t \in T(c)} \mathbf{1}[t \in T(a)], \tag{8}$$

where $T(x)$ is the tokenized version of x. This counts duplicates, i.e., repeated matches.

Answer Token Repetition. To normalize for answer length, we define redundancy as raw answer-token hits per unique answer token:

$$\text{AnsTokRepetition}(a, c) = \frac{\text{AnsTokHits}(a, c)}{|T(a)|}, \tag{9}$$

where $|T(a)|$ is the number of unique tokens in the answer. This measures the degree of repetition relative to the answer's own size.

C.2 Confounder Distributions

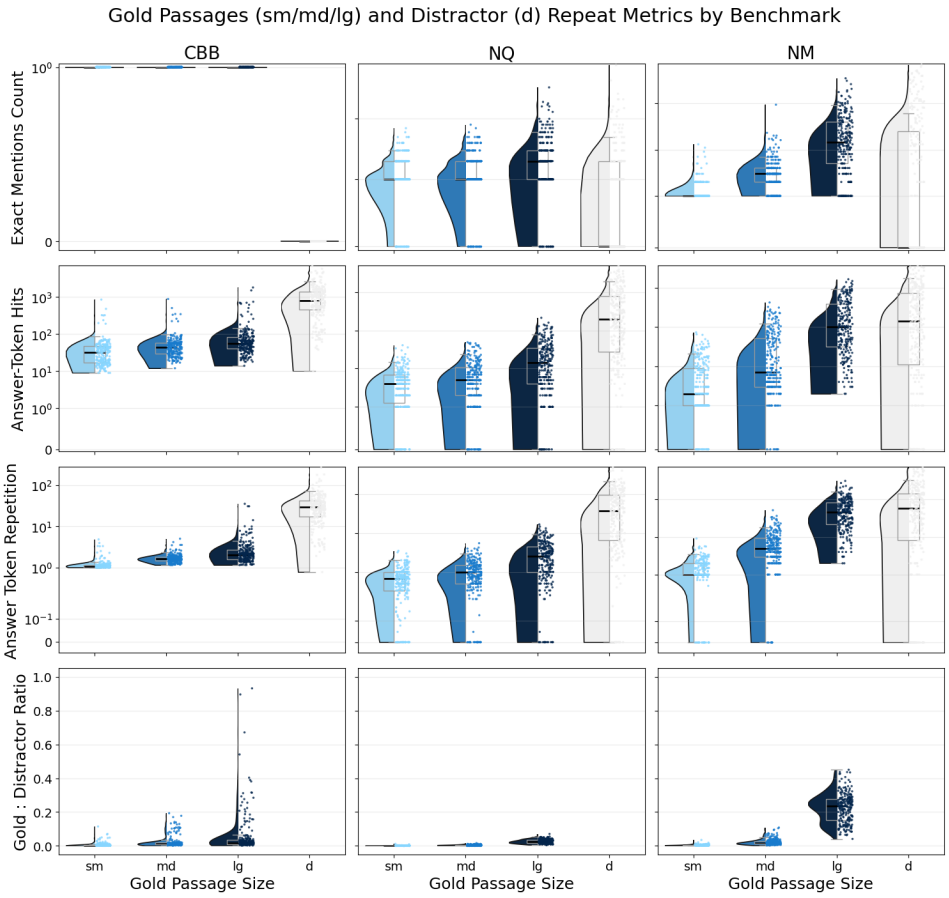


Figure 15: Raincloud plots across all benchmarks of Exact Mentions Counts, Answer-Token Hits, Redundancy, and Gold : Distractor Ratio across all sizes of gold and distractor documents for reference.

D LLM Declaration

LLMs were used to assist in editing and revising some of the language used throughout the manuscript. Additionally, LLMs were used to edit code to create some of the figures that appear in the manuscript. The authors take full responsibility for the work in its entirety.