Isolating Context Depth Effects on LLM Reasoning: Evidence of Chain-of-Thought Compression Decoupled from Accuracy

Anonymous ACL submission

Abstract

Standard benchmarks evaluate large language model (LLM) reasoning on isolated problems with minimal surrounding context, yet real-world deployments embed reasoning queries within long retrieval-augmented or conversational contexts. Prior work has shown that accuracy degrades with context length, but has not isolated context depth as an independent variable, nor examined its effect on reasoning *structure* rather than output correctness. We introduce a controlled experimental design that holds retrieval constant—always placing the target problem at the end of context—to isolate the effect of depth alone. Across 3,600 inference calls spanning four open-weight models, three depth conditions (1K, 8K, 20K tokens), and three reasoning domains (mathematics, logical reasoning, causal inference), we find that reasoning step count compresses significantly under depth pressure (Qwen3-8B: -16.3%, Llama-3.1-8B: -13.0%), while output accuracy remains stable across all four models and all three domains. We further document a stable architectural divide in self-correction behavior—reasoning-trained models self-correct in 78–97% of traces versus 9–23% for instruction-tuned models—that is invariant to context depth, indicating this property reflects training rather than deployment conditions. These findings suggest that standard benchmarks may systematically overestimate the reasoning quality of models in long-context deployments, and motivate context-aware evaluation protocols. Code and data are publicly available.[1]

1 Introduction

Reasoning benchmarks such as GSM8K (Cobbe et al., 2021), ARC (Clark et al., 2018), and BIG-Bench (Srivastava et al., 2023) present models with isolated problems embedded in minimal context. This design isolates reasoning ability cleanly, but it does not reflect the conditions under which LLMs increasingly operate: retrieval-augmented generation (RAG) pipelines prepend retrieved documents before the query, multi-turn conversational agents accumulate long interaction histories, and document analysis tasks embed questions within extensive source material.

A growing literature has examined whether this mismatch matters for output accuracy. Liu et al. (2024) showed that accuracy drops when relevant information appears in the middle of long contexts. Du et al. (2025) showed that context length alone—even with perfect retrieval positioning—degrades reasoning performance. These results establish that something in the context affects outputs. What remains unknown is what changes in the *reasoning process itself*.

The central gap. Prior long-context studies conflate two distinct variables: (1) *retrieval difficulty*—how hard it is to locate relevant information in context—and (2) *context depth pressure*—the effect of having a large volume of preceding text regardless of retrieval position. When the problem is buried mid-context, retrieval difficulty and depth pressure co-occur, making it impossible to attribute observed effects to either cause alone. To study depth pressure alone, one must hold retrieval constant.

This work. We introduce a controlled design that fixes the problem at the *end* of context across all conditions, eliminating the positional confound entirely. This allows us to ask: does increasing context depth alone affect reasoning structure, and if so, does structural change predict accuracy degradation?

We measure three structural properties of chain-of-thought responses—step count, self-correction rate, and reasoning domain sensitivity—across four models and three depth conditions. We report per-

[1] https://anonymous.4open.science/r/reasoning-position-experiment-9F26/

model results throughout to avoid pooling across architecturally distinct models.

Contributions.

- A controlled experimental design isolating context depth from retrieval position, fully replicable via a released codebase.
- Per-model evidence that reasoning step count compresses under depth pressure (Qwen3-8B: -16.3%; Llama-3.1-8B: -13.0%) while accuracy remains stable across all conditions.
- Documentation of a stable, training-determined architectural divide in self-correction behavior that is invariant to context depth.
- An empirical basis for reconsidering short-context benchmarks as proxies for long-context deployment quality.

2 Related Work

Long-context accuracy. Liu et al. (2024) established that LLMs perform worse when relevant content appears in the middle of long contexts, motivating positional controls in long-context evaluation. Du et al. (2025) extended this by showing that context length itself hurts reasoning even when retrieval is perfect, directly motivating our investigation of the underlying mechanism. Chroma Research (2025) document consistent accuracy degradation across frontier models as a function of context length, with steeper declines on complex tasks.

Chain-of-thought length and quality. Jin et al. (2024) found that reasoning step length is the primary factor affecting CoT performance, with degradation at both extremes. Work on deliberate CoT compression—CoT-Valve (Ma et al., 2025) and step-entropy pruning (Chen et al., 2025)—shows that substantial fractions of reasoning steps can be removed without accuracy loss, suggesting redundancy in verbose chains. Our work is complementary: we document that *spontaneous* compression under context pressure already occurs at inference time, without any deliberate compression mechanism being applied.

Distinguishing our contribution. Existing work either measures accuracy without examining reasoning structure, or studies engineered compression rather than naturally occurring compression. We are the first, to our knowledge, to measure CoT structural properties (step count, self-correction rate) as a function of context depth in a design that controls for retrieval position.

3 Experimental Design

3.1 Research Questions

RQ1 Does reasoning step count decrease as context depth increases, holding retrieval position constant?

RQ2 Is any structural compression accompanied by accuracy degradation, or are the two measures decoupled?

RQ3 Does self-correction behavior change under context depth pressure, and does it differ by model training regime?

3.2 Models

We evaluate four open-weight models representing two architectural categories:
Instruction-tuned: Llama-3.1-8B-Instruct (Meta AI, 2024) and Mistral-7B-Instruct-v0.1 (Jiang et al., 2023), trained for general instruction following without explicit chain-of-thought reasoning.
Reasoning-trained: Qwen3-8B (Qwen Team, 2025) and DeepSeek-R1-Distill-Qwen-32B (Guo et al., 2025), trained with reinforcement learning to produce explicit extended reasoning traces.

All models were accessed via the OpenRouter API at temperature 0.0 to ensure deterministic outputs.

3.3 Problem Set

We use 300 problems across three domains, 100 each:
Mathematics: Multi-step arithmetic and algebra word problems from GSM8K (Cobbe et al., 2021), with verifiable numerical ground truth.
Logical reasoning: Multiple-choice science questions from ARC-Challenge (Clark et al., 2018), requiring deductive reasoning from given premises.
Causal inference: Causal reasoning tasks from BIG-Bench (Srivastava et al., 2023), requiring identification of cause-effect relationships.

3.4 Depth Conditions

We operationalise depth as total context token count, varied across three conditions:

- **Baseline (≈1K tokens):** Minimal prompt framing only.

Model	Base	Deep	Δ	%
Qwen3-8B	9.57	8.02	−1.56	−16.3
Llama-3.1-8B	4.99	4.33	−0.65	−13.0
DeepSeek-R1-32B	5.96	5.91	−0.05	−0.8
Mistral-7B	4.30	4.36	+0.06	+1.4

Table 1: Reasoning step count at baseline and deep conditions. Two models show consistent compression with non-overlapping confidence intervals; two show no meaningful change.

- **Medium (≈8K tokens):** Problem embedded in 8K tokens of Wikipedia filler.
- **Deep (≈20K tokens):** Problem embedded in 20K tokens of Wikipedia filler.

Retrieval control. The target problem is always positioned at the *end* of the context window across all conditions. By fixing retrieval to the trivially easy end-of-context position, we eliminate the positional confound and isolate the effect of context volume alone.

3.5 Metrics

Step count is computed by counting discrete reasoning steps, identified by numbered markers and logical connective phrases. For reasoning-trained models (Qwen3, DeepSeek-R1), extraction uses the model's explicit reasoning field rather than the final answer output.

Self-correction rate is a binary indicator of whether a response contains markers of self-revision (*wait*, *actually*, *let me reconsider*, etc.).

Accuracy is determined by ground-truth comparison with domain-appropriate parsing.

3.6 Statistical Analysis

We report per-model statistics with 95% bootstrap confidence intervals as primary results. We do not pool across models in significance tests, as the four models are architecturally heterogeneous and pooling would introduce pseudoreplication. Effect sizes are reported as percentage change from baseline. Total inference calls: $4 \times 300 \times 3 = 3{,}600$.

4 Results

4.1 Step Count Compresses Under Depth Pressure

Figure 1 shows mean step count per model across depth conditions. Per-model results are in Table 1.

Two models show consistent step-count compression from baseline to deep (Qwen3-8B and Llama-3.1-8B), with confidence intervals that do not overlap across conditions. The compression pattern does not align with model architecture category: one reasoning-trained model (Qwen3) compresses, while another (DeepSeek) does not; the same split occurs among instruction-tuned models.

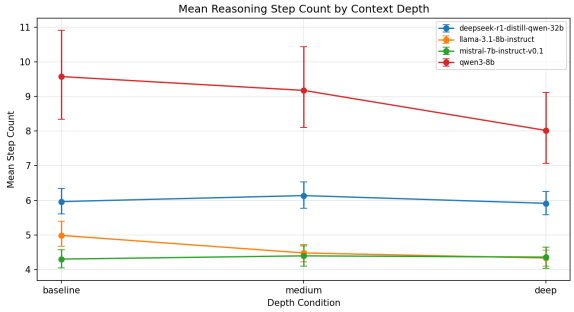

Figure 1: Mean reasoning step count by context depth condition. Error bars show 95% bootstrap confidence intervals. Qwen3-8B and Llama-3.1-8B show meaningful compression; Mistral and DeepSeek remain approximately flat.

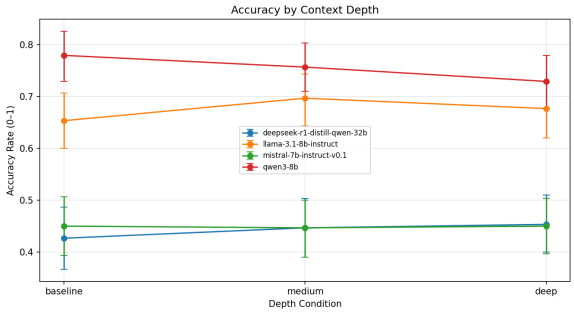

Figure 2: Accuracy by context depth condition. Across all four models, confidence intervals overlap substantially across depth conditions, indicating no meaningful degradation attributable to context depth under our controlled design.

4.2 Accuracy Is Stable Across Depth Conditions

Figure 2 and Table 2 show accuracy across depth conditions. No model shows a statistically significant accuracy change. Qwen3-8B shows the largest step compression (−16.3%) and the largest (but still non-significant) accuracy change (−0.050); confidence intervals overlap across all depth conditions.

4.3 Structural Compression Is Decoupled from Accuracy

Figure 3 directly visualises the structural-accuracy relationship. All four models fall well below the

Model	Base	Deep	Δ
Qwen3-8B	0.779	0.729	−0.050
DeepSeek-R1-32B	0.427	0.453	+0.027
Llama-3.1-8B	0.653	0.677	+0.023
Mistral-7B	0.450	0.450	0.000

Table 2: Accuracy at baseline and deep conditions. No model shows a statistically significant accuracy change across depth conditions.

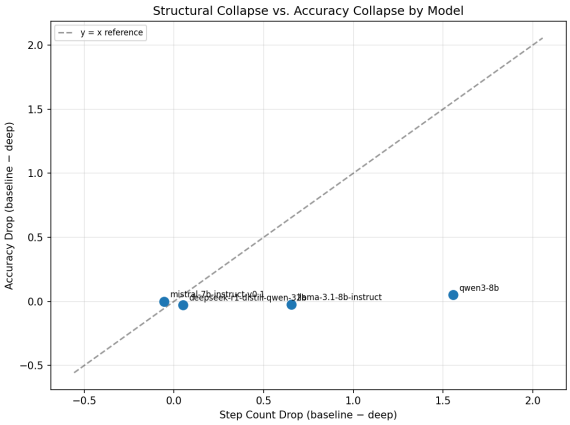

Figure 3: Step count drop vs. accuracy drop (baseline − deep) per model. All models cluster near zero on the accuracy axis regardless of step count change. The dashed line shows where the two metrics would change equally.

$y = x$ reference line, clustered near zero on the accuracy axis. The model with the largest step compression (Qwen3-8B, $x \approx 1.56$) shows only 0.050 accuracy change. This decoupling pattern holds regardless of model architecture.

4.4 Self-Correction Rate: A Stable Architectural Divide

Figure 4 reveals a large and stable gap in self-correction rates. Reasoning-trained models self-correct in nearly every response regardless of depth; instruction-tuned models rarely self-correct at any depth. The within-model change across depth conditions is negligible for all four models, suggesting that self-correction is a property established during training rather than a response to contextual pressure.

4.5 Domain-Level Heterogeneity

Figure 5 shows accuracy across the model × domain × depth grid. Mathematical reasoning is the most stable domain: Llama maintains 0.85–0.87 and Qwen3 maintains 0.94–0.96 across all depth conditions. Causal reasoning is hardest

Figure 4: Self-correction rate by context depth. Reasoning-trained models (Qwen3-8B: 95–97%; DeepSeek-R1: 78–82%) self-correct far more frequently than instruction-tuned models (Llama: 20–23%; Mistral: 9–11%). Neither group changes meaningfully across depth conditions.

and most variable, with DeepSeek ranging 0.16–0.23 and Mistral 0.26–0.28. Several domain-depth cells show improving accuracy at greater depth (e.g., Llama causal: 0.44 → 0.53), suggesting non-monotonic effects that warrant further investigation.

5 Discussion

5.1 Why Compression Without Accuracy Loss?

Two interpretations are consistent with our data. First, baseline reasoning chains may contain redundant scaffolding—verbalised steps that are not causally necessary for reaching the correct answer. Under this view, context depth triggers implicit efficiency pressure that removes overhead without affecting load-bearing reasoning. This is consistent with deliberate compression work: Chen et al. (2025) found that up to 80% of low-entropy steps can be pruned without accuracy loss.

Second, the problems in our benchmark may be insufficiently demanding to reveal the cost of compression. On problems where every reasoning step is genuinely necessary for correctness, the same compression pressure might produce accuracy degradation. This motivates extending our design to harder benchmarks such as MATH (Hendrycks et al., 2021) and BIG-Bench Hard (Suzgun et al., 2023).

5.2 Self-Correction as an Architectural Property

The stability of the self-correction divide across depth conditions has a practical implication for model selection. Practitioners deploying models

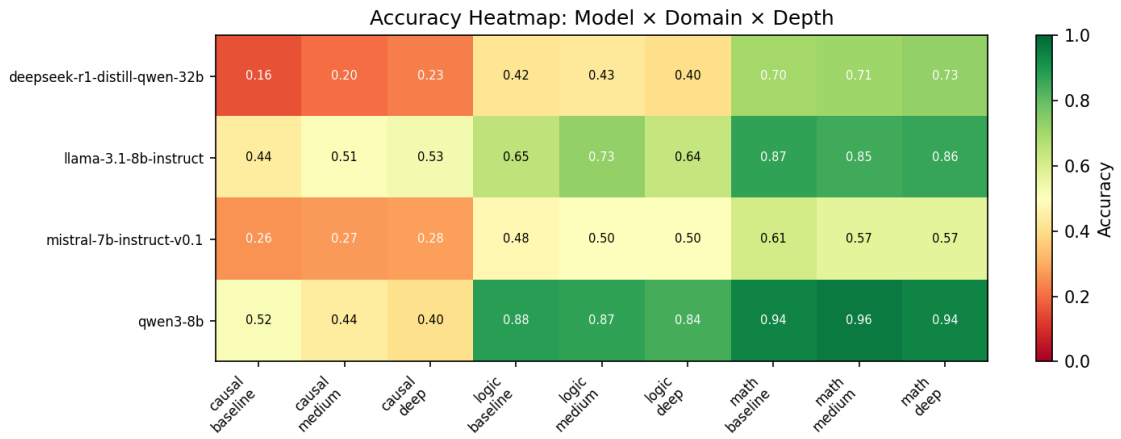

Figure 5: Accuracy heatmap by model, domain, and depth. Mathematical reasoning is most stable across depths; causal reasoning shows the most cross-model variance and depth sensitivity.

in long-context settings where reasoning reliability is critical cannot expect instruction-tuned models to self-correct more under pressure. The capacity for self-correction is instilled during reasoning-focused training and is not elicited by task difficulty or context depth.

5.3 Implications for Evaluation

Our controlled design reveals a phenomenon invisible to standard evaluation: models reason structurally differently when embedded in long context than when evaluated in isolation. Standard benchmarks, by presenting problems in minimal context, may systematically overestimate the reasoning quality that deployed systems exhibit. Evaluation protocols including long-context conditions alongside short-context benchmarks would provide a more complete picture of deployed model quality.

6 Conclusion

We introduced a controlled experimental design for isolating context depth effects on LLM reasoning, and used it to document spontaneous chain-of-thought compression under depth pressure that is decoupled from accuracy loss. Two of four tested models compress their reasoning step count by 13–16% at 20K-token context depth without any corresponding accuracy degradation. A further finding—that self-correction rates form a stable, depth-invariant architectural divide between reasoning-trained and instruction-tuned models—has direct implications for model selection in reliability-critical deployments. The controlled design, code, and full dataset are released to support replication and extension.

Limitations

Model scale. All models are in the 7–32B parameter range. The compression effect may differ at frontier scale, where models may have greater capacity to maintain reasoning depth under pressure.

Context length ceiling. Our deep condition reaches 20K tokens. Modern deployments routinely use 128K+ token contexts. Whether the compression and decoupling effects hold or intensify at longer contexts is an open question.

Step-count metric. Our step-count measure relies on surface markers and may conflate reasoning compression with formatting variation across models. A token-length-based or entropy-based measure would provide a complementary, more model-agnostic signal.

Filler text type. We used semantically neutral Wikipedia passages as filler. Adversarial or domain-matched filler might produce different compression patterns.

Problem difficulty. Our benchmark tasks have well-defined ground truth and moderate difficulty. The decoupling result may not generalise to open-ended or highly complex reasoning tasks.

Ethics Statement

This work uses publicly available benchmark datasets (GSM8K, ARC-Challenge, BIG-Bench) and open-weight models accessed via a commercial API. No human subjects were involved. The released code and data contain no personally identifiable information.

References

Fangwei Chen and 1 others. 2025. Making slow thinking faster: Compressing LLM chain-of-thought via step entropy. *arXiv preprint arXiv:2508.03346*.

Chroma Research. 2025. Context rot: How increasing input tokens impacts LLM performance. Technical report, Chroma.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Yufeng Du, Minyang Tian, Srikanth Ronanki, Subendhu Rongali, Sravan Bodapati, Aram Galstyan, and Azton Wells. 2025. Context length alone hurts LLM performance despite perfect retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2025*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

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

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Mingyu Jin, Qinkai Yu, Jingyuan Huang, Qingcheng Zeng, Zhenting Wang, Wenyue Hua, Haiyan Zhao, Kai Mei, Yanda Meng, Kaize Ding, Fan Yang, Mengnan Du, and Yongfeng Zhang. 2024. The impact of reasoning step length on large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*.

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

Xinyin Ma, Guangnian Wan, Runpeng Yu, Gongfan Fang, and Xinchao Wang. 2025. CoT-Valve: Length-compressible chain-of-thought tuning. In *Proceedings of the Association for Computational Linguistics: ACL 2025*.

Meta AI. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Qwen Team. 2025. Qwen3 technical report. *arXiv preprint*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adria Garriga-Alonso, and 1 others. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, and Jason Wei. 2023. Challenging BIG-Bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*.