WHEN LONG CONTEXTS BREAK LOGIC: SEPARATING EVIDENCE USE AND DECISION BIAS IN INSTRUCTION-TUNED LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) increasingly operate over long contexts, yet their logical reasoning remains brittle when many irrelevant tokens intervene between premises and query. A recurring challenge is *diagnosis*: when an LLM answers incorrectly in a long context, is the failure due to (i) not using the relevant premises, (ii) failing to compose them into a valid inference, or (iii) a biased decision rule at the final Yes/No readout? We present a compact suite of probes that disentangle these failure modes using *matched-prior subtraction*—a distractor-conditioned control prompt that preserves formatting and length while removing the content of the evidence. Across three open instruction-tuned models (Qwen2.5, Llama-3.2, Gemma-2) we find that evidence influence on the final decision is near-zero in early layers and rises sharply only in late layers on a "needle-in-a-haystack" variant of LogicBench. For synthetic multi-premise rules (modus tollens, disjunctive syllogism, etc.), we show that many "oracle" failures under naive scoring are actually decision-level miscalibration: simple calibrated decision rules raise oracle accuracy to 0.83–0.93 on several rules. Finally, a *local calibratability* analysis reveals that the required decision correction depends systematically on evidence placement (front/middle/end/interleaved), indicating multiple long-context bias regimes rather than a single global calibration.

1 INTRODUCTION

Logical reasoning remains a key bottleneck for deploying LLMs in high-stakes settings (e.g., scientific hypothesis checking, legal reasoning, and medical triage), where conclusions must follow from stated premises and remain consistent across related queries. While recent benchmarks show strong short-context performance, long-context settings introduce a confound: the model must retrieve and use relevant premises amid many irrelevant tokens, and failures are often summarized as "lost-in-the-middle" effects (Liu et al., 2024) or long-context degradation on broad benchmarks (Bai et al., 2024; An et al., 2024). However, a wrong answer in a long context does not uniquely identify *where* the reasoning pipeline failed.

This paper studies long-context logical inference through three questions: (1) *Evidence use:* does the model's decision depend on the evidence document at all under heavy distractors? (2) *Composition:* if the relevant premises are isolated, can the model apply the intended inference rule? (3) *Decision bias:* even when evidence and composition are present, is the final Yes/No decision rule miscalibrated by distractors?

To address these questions, we adapt a control strategy used in contrastive/context-aware decoding (Li et al., 2023; Shi et al., 2024) into a *probe* rather than a decoding algorithm: we compare the model under a *full* prompt containing the evidence to a *matched prior* prompt that preserves the same distractors, formatting, and evidence length but replaces evidence *content* with a dummy span. Subtracting these two conditions estimates the *marginal evidence contribution* to the Yes/No decision in that context. We then combine this with a stage decomposition (Direct vs Extracted vs Oracle premises) and a local calibration analysis to separate failures that are plausibly fixable by decision debiasing from those that are not.

Contributions.

- We introduce a **matched-prior evidence probe** for long-context logical reasoning that isolates the marginal contribution of the evidence document while holding distractors, length, and schema fixed.
- We provide a **stage decomposition** (Direct / Extracted / Oracle) and show that several apparent "oracle" failures arise from **decision-level miscalibration**, not a lack of compositional competence.
- We propose a **local calibratability** metric (existence of a correct decision correction within a realistic neighborhood around a tuned calibration) and find **placement-dependent bias regimes** across reasoning rules.
- We report **cross-model mechanistic signatures** on a needle-in-haystack version of LogicBench (Parmar et al., 2024): evidence influence emerges sharply in late layers for Qwen2.5-3B-Instruct (Qwen Team, 2024), Llama-3.2-3B-Instruct (Meta AI, 2024), and Gemma-2-2B-it (Google DeepMind, 2024).

2 Experimental setup

2.1 Models

We evaluate three instruction-tuned decoder-only Transformers (Vaswani et al., 2017): Qwen2.5-3B-Instruct (Qwen Team, 2024), Llama-3.2-3B-Instruct (Meta AI, 2024), and Gemma-2-2B-it (Google DeepMind, 2024). All experiments are run with greedy decoding disabled; we score by log-likelihood (Section 2.4) rather than free-form generation.

2.2 Tasks

LogicBench BQA under distractors. We use LogicBench Boolean Question Answering (BQA) tasks (Parmar et al., 2024) and convert each instance into a long-context "needle-in-a-haystack" prompt (Kamradt, 2023): the target LogicBench context (rules + facts) appears inside one document among many distractor documents, and the question is asked at the end of the overall prompt.

Synthetic multi-premise rules. To control placements and labels, we also generate templated instances for classical inference rules (modus tollens, disjunctive syllogism, hypothetical syllogism, and a mixed "syllogism_all" set). Each instance contains $k \in \{2, 3\}$ relevant premises plus optional irrelevant premises. Labels are balanced by construction unless noted.

2.3 Long-context construction

We format the user message as $n_{\text{docs}} = 20$ documents separated by headers. Nineteen documents are distractors sampled from WikiText (Merity et al., 2016) ("fluent noise") or other noise sources (Appendix). One document is the *evidence document* containing the logic rules and facts. If $\texttt{pos} = 9$, the evidence is placed in the middle:

$$\underbrace{D_0 \mid D_1 \mid \cdots \mid D_8}_{\text{distractors}} \mid \underbrace{D_9 \text{ (evidence)}}_{\text{needle}} \mid \underbrace{D_{10} \mid \cdots \mid D_{19}}_{\text{distractors}} \mid \text{[QUESTION]}.$$

We keep the question at the end to ensure it is always the most recent span. A full prompt template is provided in Appendix A.

2.4 Answer scoring

Direct "first-token" Yes/No comparisons can be brittle (e.g., whitespace/newline tokens or "Answer:" prefixes can absorb probability mass). We therefore score with sequence likelihood: given a fixed answer prefix ("Answer:"), we compare the conditional log-likelihood of the completions " Yes" vs " No" (including leading whitespace) and predict the higher-likelihood label. We additionally report calibrated decision rules in Section 3.3.

3 PROBES: EVIDENCE CONTRIBUTION, STAGES, AND CALIBRATION

3.1 MATCHED-PRIOR SUBTRACTION

Let x be the full prompt (distractors + evidence + question) and x^- the *matched prior* prompt where the evidence document is replaced by a dummy span of identical token length and identical formatting (header, position, separators). For a Yes/No task we define the logit-difference score $s(x) = \log p(\text{“Yes”} \mid x) - \log p(\text{“No”} \mid x)$ under the scoring protocol of Section 2.4. The marginal evidence contribution is

$$\Delta(x) \;=\; s(x) - s(x^-). \tag{1}$$

We also compute a layerwise version Δ_ℓ by applying a logit lens to the residual stream at layer ℓ (Appendix). We visualize *signal recovery* as Δ_ℓ/Δ_L where L is the final layer.

3.2 STAGE DECOMPOSITION

We evaluate each instance under three stages: **Direct** (full long context), **Extracted** (a short prompt containing only premises selected by a lightweight extractor), and **Oracle** (a short prompt containing the ground-truth relevant premises only). This separates failures due to selecting the wrong premises from failures to compose selected premises into the correct conclusion.

3.3 DECISION CALIBRATION AND LOCAL CALIBRATABILITY

Matched-prior subtraction yields a natural one-dimensional decision statistic $\Delta(x)$. We consider calibrated rules of the form $\hat{y} = \mathbb{1}[\alpha\Delta(x) > t]$, including (i) **threshold-only** ($\alpha = 1$) and (ii) **affine** (α, t) tuned on a held-out calibration set. To avoid overfitting and to quantify when a single calibration is insufficient, we compute a *local calibratability* metric: for each example, we test whether *any* α in a realistic neighborhood around a tuned global value (here, $\alpha \in [0.75, 0.95]$ around $\alpha_{\text{global}} = 0.85$) yields the correct label. If no α in this band succeeds, we call the example **nonlocal** (not fixable by plausible calibration). We further track whether success requires α above or below the band, defining a direction index.

4 RESULTS

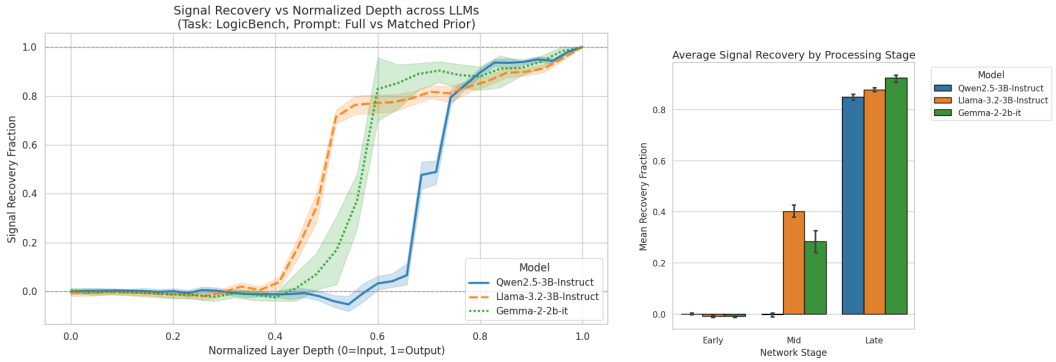
4.1 EVIDENCE INFLUENCE EMERGES LATE IN THE NETWORK ACROSS MODEL FAMILIES

Figure 1 reports layerwise signal recovery on LogicBench-BQA under fluent distractors (WikiText), with the evidence document fixed in the middle (pos=9 of 20 documents). Across all three models, the marginal evidence contribution is near zero for a large fraction of early layers, then rises sharply only in late layers. This pattern is robust across logic rule categories (Appendix) and suggests that, under heavy distractors, these instruction-tuned LLMs incorporate evidence into the final Yes/No decision primarily in late processing stages.

4.2 STAGE DECOMPOSITION SHOWS DECISION BIAS CAN MASK ORACLE COMPETENCE

Figure 2 summarizes a stage decomposition on synthetic multi-premise rules for Qwen2.5-3B-Instruct under distractors. Under naive (uncalibrated) decision rules, even the *Oracle* setting (only gold premises, no distractors) can appear near chance—a failure mode we observed repeatedly when using brittle first-token scoring. After switching to likelihood-based scoring and applying calibrated decision rules, oracle accuracy rises substantially (e.g., to ≈ 0.83 for hypothetical syllogism and ≈ 0.90 for the mixed syllogism set). For modus tollens, affine calibration yields oracle accuracy near ≈ 0.93, while the Extracted stage remains lower, indicating that premise selection (not only composition) limits performance.

This result has two implications for long-context logical reasoning: (i) long-context failures can reflect a *decision-level bias* induced by distractors, not just inability to derive the conclusion, and (ii) stage-decomposed evaluation can distinguish premise selection errors from compositional inference errors.



(a) Layerwise signal recovery Δ_ℓ/Δ_L (mean $\pm$ bootstrapped CI).

(b) Average recovery by coarse network stage (early/mid/late).

Figure 1: LogicBench needle-in-a-haystack (20 docs, fluent distractors): evidence influence is largely absent in early layers and emerges late across three model families.

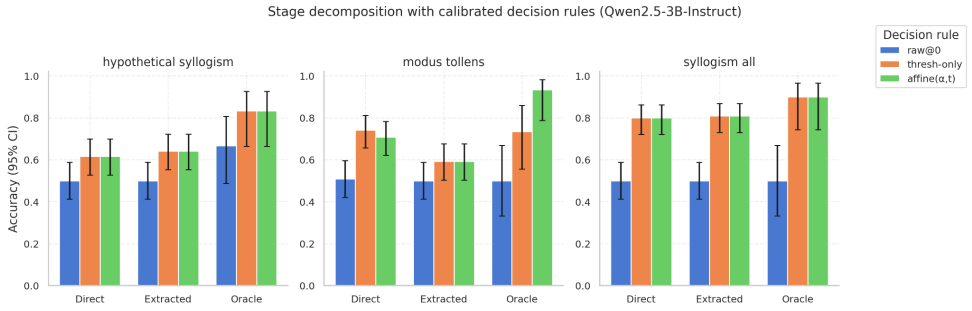


Figure 2: Stage decomposition with calibrated decision rules (Qwen2.5-3B-Instruct). "Direct" uses the full long context; "Extracted" uses premises selected by an extractor; "Oracle" uses ground-truth premises only. Calibrated decision rules substantially increase oracle accuracy on several rules, showing that some apparent oracle failures are decision-level effects.

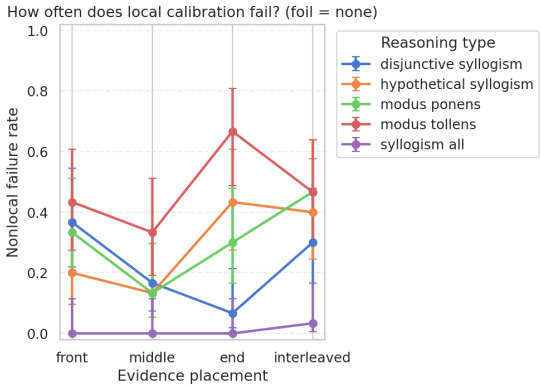
4.3 Local calibratability reveals placement-dependent bias regimes

If a single global calibration explained long-context failures, most errors would become correct under a narrow neighborhood of that calibration. Figure 3 shows this is not the case: the *nonlocal* rate (no $\alpha \in [0.75, 0.95]$ yields correctness) varies by inference rule and evidence placement (front/middle/end/interleaved). Moreover, among nonlocal cases, the *direction* of calibration need depends on placement: middle placements more often require *stronger* corrections (higher α), while front/end placements frequently require *weaker* corrections (lower α). This indicates multiple long-context bias regimes rather than a single global threshold shift.

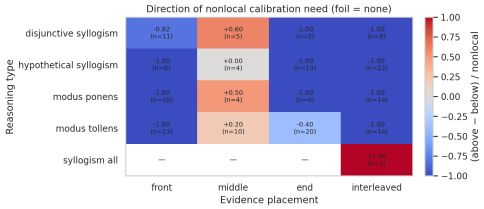
5 Discussion

Our results suggest that long-context logical reasoning failures are heterogeneous: **evidence influence** can be delayed to late layers (Figure 1), **premise selection** and **composition** can be separated by stage decomposition (Figure 2), and **decision bias** induced by distractors is *placement-dependent* and not globally correctable (Figure 3). This has practical implications for benchmark design and evaluation: reporting only end-task accuracy in long contexts can conflate evidence use, inference, and decision calibration.

More broadly, matched-prior subtraction provides a simple diagnostic tool for analyzing whether an LLM's decision depends on the intended evidence under heavy distractors—complementing



(a) Nonlocal (not locally calibratable) rate by placement.



(b) Direction index (above−below) among nonlocal cases; n shown per cell.

Figure 3: Local calibratability analysis (foil-free synthetic rules). Long-context decision bias is not explained by a single global calibration: the required correction varies systematically with evidence placement and rule type.

long-context benchmarks (Bai et al., 2024; An et al., 2024) and "lost-in-the-middle" analyses (Liu et al., 2024).

6 Limitations and Future Work

We focus on Yes/No logical inference and three small open models; larger models and multi-token answers may exhibit different regimes. Our extractor is intentionally lightweight; improving premise selection may shift the balance between selection and composition failures. Finally, while our layerwise evidence probe is suggestive, connecting these signatures to specific attention heads or circuits would require deeper mechanistic analysis.

References

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. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14388–14411, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.776. URL https://aclanthology.org/2024.acl-long.776/.

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 Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3119–3137, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.172. URL https://aclanthology.org/2024.acl-long.172/.

Google DeepMind. Gemma 2 2B-IT model card. https://huggingface.co/google/gemma-2-2b-it, 2024. Accessed: 2026-02-23.

Greg Kamradt. LLMTest_NeedleInAHaystack: Pressure testing long-context retrieval with a needle-in-a-haystack task. GitHub repository, 2023. URL https://github.com/gkamradt/LLMTest_NeedleInAHaystack. Accessed 2026-02-23.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 12286–12312, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.687. URL https://aclanthology.org/2023.acl-long.687/.

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*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models, 2016. Introduces the WikiText language modeling datasets.

Meta AI. Llama 3.2 3B Instruct model card. https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct, 2024. Accessed: 2026-02-23.

Mihir Parmar, Nisarg Patel, Neeraj Varshney, Mutsumi Nakamura, Man Luo, Santosh Mashetty, Arindam Mitra, and Chitta Baral. Logicbench: Towards systematic evaluation of logical reasoning ability of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13679–13707, Bangkok, Thailand, 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.739/.

Qwen Team. Qwen2.5-3B-Instruct model card. Hugging Face model card, 2024. URL https://huggingface.co/Qwen/Qwen2.5-3B-Instruct. Accessed 2026-02-23.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. Trusting your evidence: Hallucinate less with context-aware decoding. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 783–791, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-short.69. URL https://aclanthology.org/2024.naacl-short.69/.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

A Prompt templates and additional details

Needle-in-a-haystack template.

```
[SYSTEM] You are a careful logician. Answer exactly "Yes" or "No".

[USER]
[DOC 0] <distractor>
...
[DOC 8] <distractor>
[DOC 9] Context:
  Rules: <LogicBench or synthetic rules>
  Facts: <LogicBench or synthetic facts>
[DOC 10] <distractor>
...
[DOC 19] <distractor>

[QUESTION] <Yes/No query>
Answer:
```