Mixing Matters? Evidence and Its Limits for Position Bias Across Sequence Mixers

Anonymous Author(s)
Affiliation
Address
email

Abstract

Language models often use relevant evidence less effectively when it appears in the middle of a long context. We ask whether this position bias changes with the mechanism used to mix information across tokens. For each multi-document question, we move the answer-bearing document through ten positions while holding the question and distractors fixed. In a shared-corpus comparison near 2.8B parameters, Pythia shows a clear beginning advantage that two Mamba variants lack, while all three models show an end advantage. This near-2.8B family comparison confounds sequence mixing with depth, RoPE or other positional encoding, recurrent-state size, and checkpoint family, and each architecture is represented by a single checkpoint rather than a sample of checkpoints. A tighter comparison between pure and hybrid Mamba-2 models points in the same direction, but its paired effect remains statistically uncertain. Across additional exploratory studies, the family gap appears only once the compared models can answer the task, a corpus swap changes overall accuracy without a detectable change in curve shape, and Pythia reproduces the beginning advantage on synthetic retrieval. Prompt design is associated with materially different per-condition edge estimates: Pythia primacy is $+0.070$ under the baseline template and $+0.005$ under an instructional template, while Mamba primacy is 0.000 under the baseline layout and $+0.103$ under a bookended layout. These exploratory variant cells were not tested with a paired between-condition contrast, so their differences are descriptive. Attention-sink measurements and position probes motivate hypotheses about evidence use, but do not establish a causal mechanism. All ten-document QA results use an exploratory split, and the held-out confirmatory split remains unexamined.

1 Introduction

Retrieval-augmented systems often present a model with several documents and expect it to use whichever one contains the answer. Yet evidence is not equally usable everywhere in the prompt: accuracy commonly falls when the relevant document moves from an edge toward the middle (Liu et al., 2024). This failure is important even when retrieval itself succeeds, because a downstream model can still neglect the retrieved evidence. Long-context benchmarks expose related weaknesses across question answering, summarization, and retrieval (Bai et al., 2024; Shaham et al., 2023; Li et al., 2024).

The usual lost-in-the-middle curve has mostly been studied in Transformers. Dense attention can revisit any earlier token directly, whereas state-space models such as Mamba compress the prefix into a recurrent state (Gu and Dao, 2024; Dao and Gu, 2024). This architectural difference could change how evidence position affects an answer. It is difficult to test cleanly, however, because model families also differ in training data, tokenizer, depth, scale, positional encoding, and capability.

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

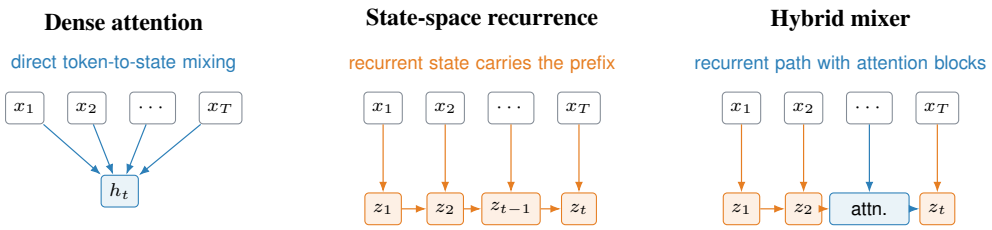


Figure 1: Schematic sequence mixers. Dense attention permits direct token interactions, state-space models carry prefix information through a recurrent state, and hybrids combine both. The drawing is conceptual. The depicted 1-in-4 attention block is illustrative; the hybrid checkpoint in Section 4 replaces about 7% of blocks with attention.

We study the question with a paired intervention on evidence position. For every question, the same answer document and distractors form one bundle, and only the answer document's position changes. The main comparison finds a beginning advantage for Pythia that is absent in two similarly sized Mamba models trained on the same corpus, while the end advantage is shared. This near-2.8B family comparison confounds sequence mixing with depth, RoPE or other positional encoding, recurrent-state size, and checkpoint family, and each architecture is represented by a single checkpoint rather than a sample of checkpoints. A more tightly matched pure-versus-hybrid Mamba-2 comparison is directionally compatible but does not detect a reliable paired checkpoint difference. Scale, corpus, task, production-model, prompt, sink, and probe studies then define where the observation does and does not generalize.

The paper makes three contributions. First, it provides an auditable position-intervention harness with anchors, preserved outputs, paired inference, and an untouched confirmatory split. Second, it presents an exploratory family-by-position comparison together with a tighter but statistically uncertain pure-versus-hybrid checkpoint comparison. Third, it turns the remaining phases into bounded evidence: corpus and task checks test scope, while prompt interventions and internal measurements generate mechanistic hypotheses without being presented as causal proof.

2 Related work

Position bias in long contexts Liu et al. (2024) established the lost-in-the-middle curve in retrieval and multi-document QA. This task derives its questions from Natural Questions (Kwiatkowski et al., 2019), while LongBench, ZeroSCROLLS, and LooGLE broaden long-context evaluation across tasks, languages, and reasoning demands (Bai et al., 2024; Shaham et al., 2023; Li et al., 2024). Subsequent work measures effective context length (Hsieh et al., 2024b), associates the curve with attention allocation (Hsieh et al., 2024a), and proposes reordering, compression, retrieval, and context-use interventions (An et al., 2024; Peysakhovich and Lerer, 2023; Jiang et al., 2024; Xu et al., 2024; Agrawal et al., 2024). Irrelevant context can itself distract a model, motivating our use of the same distractors at every tested position (Shi et al., 2023). We focus on whether the curve changes across sequence mixers.

Sequence mixers Dense self-attention (Vaswani et al., 2017) mixes token pairs under a causal mask and commonly encodes order with rotary embeddings or linear attention biases (Su et al., 2024; Press et al., 2022). Alternative long-sequence mixers include structured state spaces, recurrent weighted key-value models, and long convolutions (Gu et al., 2022; Peng et al., 2023; Poli et al., 2023). Mamba and Mamba-2 carry information through selective state-space recurrences (Gu and Dao, 2024; Dao and Gu, 2024). Hybrid models such as Griffin, Jamba, and Nemotron-H combine recurrent or state-space blocks with attention (De et al., 2024; Lieber et al., 2024; NVIDIA, 2025). Waleffe et al. (2024) release the pure and hybrid 8B Mamba-2 checkpoints used in our tightest checkpoint comparison.

Cross-architecture long-context evidence Architecture comparisons do not imply that recurrent models are uniformly position-robust. Huang (2025) find practical long-context failures across recurrent, hybrid, and Transformer designs, and a concurrent preprint reports a U-shaped Mamba recall

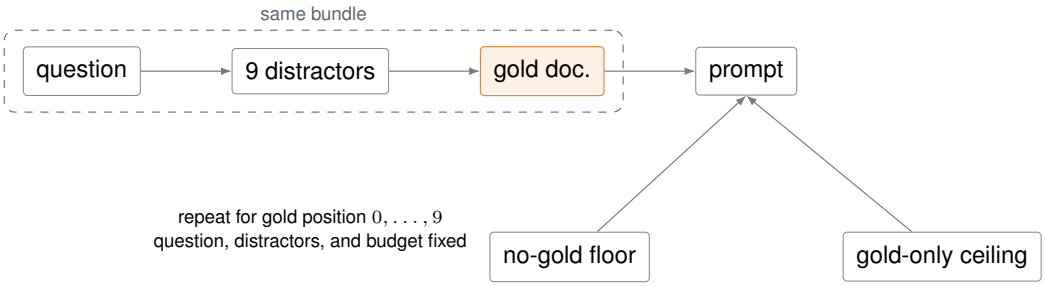


Figure 2: Paired position intervention. Each question contributes one fixed bundle and ten prompts; only the gold document's position changes. The no-gold and gold-only anchors bound guessing and answerability.

curve on structured memory tasks (Airlangga et al., 2025). These results differ from our Mamba QA curve and reinforce a central boundary of this study: position profiles depend on the task and prompt, not only the mixer. Work that reformulates Mamba recurrence as implicit attention offers a complementary route for future mechanism tests (Ali et al., 2025).

Attention sinks and probes Xiao et al. (2024) showed that causal transformers can concentrate attention on initial tokens, and Sun et al. (2024) connected unusually large internal activations to concentrated attention. We test whether this attention-sink statistic covaries with primacy, treating the result as correlational because a valid sink-blocking ablation is absent. Our position probe follows the principle that a diagnostic classifier needs a control task before its accuracy can be interpreted as evidence about a representation (Hewitt and Liang, 2019).

3 Measurement and statistics

Task and anchors We use the 2,655 released questions from the Lost-in-the-Middle 10-document QA setting (Liu et al., 2024), which is built from Natural Questions (Kwiatkowski et al., 2019). Each question has one gold document and nine distractors. We build ten prompts per question, moving only the gold document through positions 0 to 9 while retaining the same distractors and token budget. A no-gold floor measures guessing or memorized knowledge, and a gold-only ceiling measures answerability under perfect retrieval. The fixed split uses seed 240521 and contains 800 exploratory and 1,855 confirmatory questions. Every reported ten-document QA result uses the exploratory allocation or a subset of it. The confirmatory allocation remains unopened.

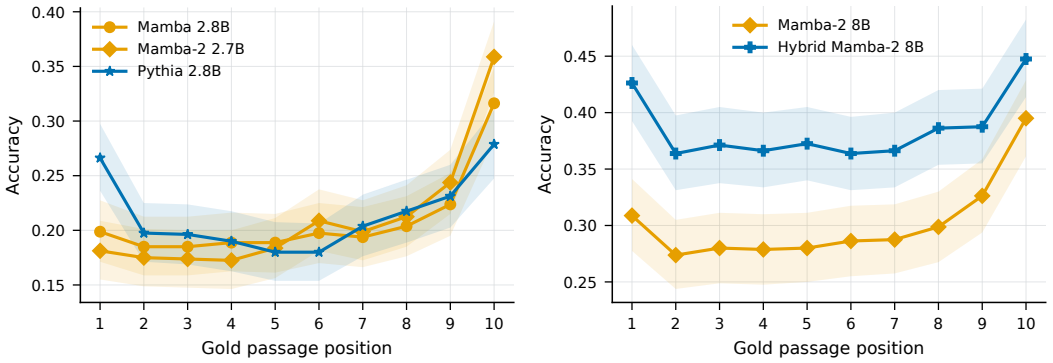
Edges and statistics Let a_S, a_M, and a_E be mean accuracy at start positions $\{0, 1\}$, middle positions $\{4, 5\}$, and end positions $\{8, 9\}$. We define primacy as $a_S - a_M$ and recency as $a_E - a_M$. The question is the unit of analysis. Each of 10,000 bootstrap samples draws complete ten-position question bundles with replacement, following the bootstrap principle of resampling the independent unit (Efron, 1979), and model comparisons reuse the same sampled questions. We report percentile 95% confidence intervals and apply Holm's sequential correction within each primacy-and-recency pair for a model or paired model contrast (Holm, 1979); across all phases this paper reports several dozen individual edge estimates, and correction is applied per pair rather than globally across the paper. The intervals quantify variation across questions under the fixed checkpoints, prompts, and decoding configuration, not variation across training seeds or model checkpoints.

Reproducibility and controls The harness vendors the official prompt builder and `best_subspan_em` scorer at a pinned upstream commit. It also records normalized exact match, following the normalization convention popularized by SQuAD (Rajpurkar et al., 2016), and first-line extraction. Generation is greedy with temperature 0, at most 32 new tokens, fixed seeds, BF16 where supported, pinned model revisions, and a fixed Mamba execution path within each controlled contrast. Each phase fixes one execution path within its controlled comparison, although the Transformers and Megatron backends differ across phases. Committed standard QA sweeps preserve prompts, raw generations, failures, token counts, software versions, and decoding settings. The

Table 1: Pile-trained models on 800 exploratory questions. Paired-bootstrap 95% CIs and Holm-corrected p-values within each model's primacy-and-recency pair; bold p indicates $\alpha = 0.05$ significance.

	Primacy			Recency			Anchors	
Model	Δ	95% CI	p	Δ	95% CI	p	Floor	Ceiling
`pythia-2.8b`	0.052	$[0.032, 0.072]$	$\mathbf{<10^{-4}}$	0.075	$[0.054, 0.097]$	$\mathbf{<10^{-4}}$	0.091	0.640
`mamba-2.8b`	-0.001	$[-0.016, 0.014]$	0.91	0.077	$[0.058, 0.097]$	$\mathbf{<10^{-4}}$	0.115	0.615
`mamba2-2.7b`	-0.018	$[-0.034, -0.003]$	**0.02**	0.105	$[0.086, 0.124]$	$\mathbf{<10^{-4}}$	0.128	0.619

Δ is anchor-minus-middle accuracy; Floor/Ceiling are the no-gold/gold-only anchors from Fig. 2.



(a) Pile-trained models near 2.8B parameters

(b) Matched pure and hybrid Mamba-2 models at 8B

Figure 3: Accuracy as the answer-bearing document moves through the context on 800 exploratory questions. Pythia rises at the beginning, while both smaller Mamba variants do not; the matched 8B hybrid also rises more than the pure model, but their paired difference remains uncertain. Shaded bands show 95% question-bootstrap confidence intervals.

synthetic retrieval check is represented by summaries and plots only; raw generations and an environment manifest are not available in the repository snapshot. The released key-value task provides an executed positive control: the pipeline detects a $+0.38$ Pythia edge-versus-middle effect. For the sham-gold control, accuracy was 0.1235 versus a 0.0950 floor, a difference of 0.0285; primacy was $+0.0050$ with 95% CI $[-0.0325, 0.0425]$, and recency was -0.0175 with 95% CI $[-0.0500, 0.0125]$. Distractor-order permutation accuracies were 0.2850, 0.3000, and 0.2767, with a maximum spread of 0.0233. Both certification controls passed their gates on the 200-question exploratory Pythia sample only, and they do not test the 8B Megatron checkpoints.

4 Architecture comparisons

Primary 2.8B comparison We compare `pythia-2.8b` (Biderman et al., 2023) with `mamba-2.8b` (Gu and Dao, 2024) and `mamba2-2.7b` (Dao and Gu, 2024). All are base models pretrained on the Pile (Gao et al., 2020), have similar parameter counts, and use related tokenizers. Table 1 and Figure 3 show a positive primacy edge for Pythia and no positive primacy edge for either Mamba model. The paired Mamba-minus-Pythia primacy interactions are -0.053 and -0.070 (both Holm $p < 10^{-4}$), while Mamba-1 versus Mamba-2 is uncertain ($+0.017$, 95% CI $[-0.005, 0.038]$). All three models have positive recency edges. These comparisons match corpus and approximate scale, but not depth, positional encoding, state size, or checkpoint family, and each architecture is represented by a single checkpoint rather than a sample of checkpoints.

Matched 8B checkpoint comparison Our tightest released-checkpoint comparison uses NVIDIA's pure and hybrid 8B Mamba-2 checkpoints (Waleffe et al., 2024), which share training data, tokenizer, scale, depth, and positional setup. The hybrid contains 7% attention layers and a different MLP composition, so the comparison is not an attention-only intervention. Within models, the pure model's primacy edge is uncertain ($+0.008$, $p = 0.36$) and the hybrid's is positive ($+0.027$, $p = 0.007$).

Table 2: Tightly matched 8B released-checkpoint comparison on 800 exploratory questions. The paired interaction tests the composite pure-versus-hybrid difference, not attention alone.

Contrast	Primacy	95% CI	Holm p
pure Mamba-2	$+0.008$	$[-0.009, 0.024]$	0.362
hybrid	$+0.027$	$[0.008, 0.047]$	0.007
hybrid minus pure	$+0.019$	$[-0.006, 0.044]$	0.144

One estimate being significant and the other not does not establish that the models differ. The paired hybrid-minus-pure primacy effect is $+0.019$ (95% CI $[-0.006, 0.044]$, Holm $p = 0.144$), while the recency effect is -0.028 (95% CI $[-0.055, -0.001]$, Holm $p = 0.089$). This direction supports the 2.8B observation but the composite checkpoint contrast does not isolate a causal attention effect.

5 Scope checks

Scale Across five approximate size pairs, the Pythia-minus-Mamba primacy gap is near zero at the two smallest scales and is $+0.062$, $+0.069$, and $+0.053$ in the three larger pairs (Figure 5). The smallest Pythia ceiling is only 0.009, so this is a descriptive capability-and-family trend, not evidence for a depth threshold or a causal scale transition.

Training corpus Holding the Mamba-2.8B architecture fixed, the Pile-minus-SlimPajama primacy effect is -0.007 (95% CI $[-0.030, 0.016]$, $p = 0.568$) and the recency effect is -0.016 (95% CI $[-0.044, 0.012]$, Holm $p = 0.502$) (Soboleva et al., 2023). The floor differs substantially (0.270 versus 0.114), so this comparison finds no edge-shape change but does not rule out training-data effects (Figure 5).

Synthetic retrieval On 50 RULER `niah_single_1` examples at 2K tokens (Hsieh et al., 2024b), Pythia again has a primacy edge ($+0.12$, 95% CI $[0.05, 0.20]$). Both Mamba models score 1.0 at every depth, so this check reproduces the transformer pattern but cannot compare mixer curves on the synthetic task (Figure 6).

Descriptive system comparison We also run `Llama-3.1-8B` (Dubey et al., 2024), `Qwen2.5-7B` (Yang et al., 2024), and `Nemotron-H-8B` (NVIDIA, 2025). All three show positive primacy edges ($+0.076$, $+0.041$, $+0.067$). Only Llama versus Qwen has a significant paired primacy interaction ($+0.036$, Holm $p = 0.016$). Because the systems differ on architecture, data, tokenizer, token count, depth, and positional encoding, these results establish prevalence, not an architectural explanation (Figure 8). The fact that only the Llama-versus-Qwen contrast clears correction cannot rank architecture against the other changing factors, and the non-significant contrasts with Nemotron-H are not evidence of equivalence.

6 Mechanistic evidence and intervention

Attention sinks track primacy Using the initial-token attention-share statistic of Xiao et al. (2024), we measure almost no final-layer token-0 sink in Pythia-160M (0.003), while Pythia-2.8B reaches 0.455. Pythia-410M has a small final-layer sink (0.034) and no significant primacy edge. Figure 7 shows a scale-linked association, but the measurements share model scale as a confound and no valid sink ablation was completed.

Utilization, not storage, is a hypothesis A balanced linear probe on frozen mid-depth states predicts edge versus middle position modestly above its shuffled-label control for Pythia (0.603 versus 0.484, a 12-point lift) and Mamba (0.648 versus 0.508, a 14-point lift). Thus position information remains linearly decodable, at a modest margin over the shuffled baseline, even when Mamba has no positive primacy edge. This is consistent with a utilization hypothesis, but the effect size is modest and probe decodability does not prove that the model stores or uses the answer-bearing content.

Training-free prompt intervention On 200 exploratory questions, the Mamba primacy estimate is 0.000 under the baseline layout and $+0.103$ under the bookended layout ($p < 10^{-4}$ within that cell).

Under the question-first layout, Mamba primacy is $+0.020$ ($p = 0.366$) and recency is $+0.010$, compared with baseline recency of $+0.070$. The bookend result is not a mitigation result: much of the larger edge comes from lower middle-position accuracy. The variants jointly change query placement, repetition, and token layout, and no paired between-condition contrast was computed, so they are not a clean test of fixed-state compression.

Prompt sensitivity The per-template Pythia primacy estimates are $+0.070$ under the baseline template and $+0.005$ under the instructional template in the 200-question mechanism subset. These are per-template exploratory estimates rather than a paired template-difference test, and the instructional-template confidence interval crosses zero. The result warns against treating either the sign or magnitude as a template-invariant architectural constant.

7 Discussion and limitations

The most secure result is the 2.8B exploratory family-by-position interaction. The matched 8B direction, sink association, and probe results motivate a mixer-dependent evidence-use hypothesis, while prompt sensitivity shows that architecture is not the only determinant.

Several limitations set the confirmatory boundary. All analyses use the exploratory split, so the planned 1,855-question test must be run under a frozen analysis before confirmatory language is warranted. The 2.8B and scale comparisons retain depth, positional-encoding, state-size, and capability confounds. Each architecture in this comparison is also represented by a single checkpoint, so an observed difference could reflect that specific training run rather than a general property of its architecture family. The matched 8B paired contrast crosses zero after Holm correction; a later clean re-run at commit `33d6bb5` (clean source tree, both repositories) reproduced the summary byte-for-byte, so the null result is not an artifact of the earlier dirty-tree run. The committed sham-gold and distractor-order certification controls passed on exploratory Pythia, but they do not test the 8B Megatron checkpoints. A 50-generation blinded audit sample exists, but human labels were not completed. Finally, the attempted hybrid sink block did not alter the model because its custom attention path did not honor the hook; it is not an ablation result. The next decisive study should freeze exclusions and contrasts, repair the hybrid intervention, and evaluate the held-out split once.

Broader impact Position-sensitive evidence use can make retrieval-augmented systems unreliable even when retrieval succeeds. An evaluation that exposes this failure can support better ordering strategies and more honest model selection. The same phenomenon can also cause systems to privilege facts because of placement rather than relevance, so reported long-context accuracy should not be treated as uniform evidence access. This work releases diagnostics and outputs, not a new generative model.

8 Conclusion

On the exploratory Lost-in-the-Middle split, Pythia-2.8B and two Pile-trained Mamba models have different primacy profiles under a paired ten-position evaluation. A matched 8B comparison is directionally consistent but statistically uncertain, so the present evidence supports a family-by-position interaction without establishing that attention causes it. The held-out split and extension of the certification controls beyond Pythia provide a clear path from this exploratory workshop result to a confirmatory study.

References

D. Agrawal, S. Gao, and M. Gajek. Can't remember details in long documents? You need some R&R. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 12692-12704, 2024.

M. C. Airlangga, H. AlQuabeh, M. S. Nwadike, and K. Inui. Emergence of primacy and recency effect in Mamba: A mechanistic point of view. *arXiv:2506.15156*, 2025.

A. A. Ali, I. Zimerman, and L. Wolf. The hidden attention of Mamba models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 1516-1534, 2025.

S. An, Z. Ma, Z. Lin, N. Zheng, J.-G. Lou, and W. Chen. Make your LLM fully utilize the context. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Y. Bai, X. Lv, J. Zhang, et al. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 3119-3137, 2024.

S. Biderman, H. Schoelkopf, Q. Anthony, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, volume 202 of PMLR, pages 2397-2430, 2023.

T. Dao and A. Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *International Conference on Machine Learning*, volume 235 of PMLR, pages 10041-10071, 2024.

S. De, S. L. Smith, A. Fernando, et al. Griffin: Mixing gated linear recurrences with local attention for efficient language models. *arXiv:2402.19427*, 2024.

A. Dubey, A. Jauhri, A. Pandey, et al. The Llama 3 herd of models. *arXiv:2407.21783*, 2024.

B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1-26, 1979.

L. Gao, S. Biderman, S. Black, et al. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv:2101.00027*, 2020.

A. Gu and T. Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *Conference on Language Modeling*, 2024.

A. Gu, K. Goel, and C. Ré. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022.

J. Hewitt and P. Liang. Designing and interpreting probes with control tasks. In *Proceedings of EMNLP-IJCNLP*, pages 2733-2743, 2019.

S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65-70, 1979.

J. Huang. How well can a long sequence model model long sequences? Comparing architectural inductive biases on long-context abilities. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 29-39, 2025.

C.-Y. Hsieh, Y.-S. Chuang, C.-L. Li, Z. Wang, L. Le, A. Kumar, J. Glass, A. Ratner, C.-Y. Lee, R. Krishna, and T. Pfister. Found in the middle: Calibrating positional attention bias improves long context utilization. In *Findings of the Association for Computational Linguistics: ACL*, pages 14982-14995, 2024.

C.-P. Hsieh, S. Sun, S. Kriman, et al. RULER: What's the real context size of your long-context language models? In *Conference on Language Modeling*, 2024.

H. Jiang, Q. Wu, X. Luo, et al. LongLLMLingua: Accelerating and enhancing LLMs in long context scenarios via prompt compression. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 1658-1677, 2024.

T. Kwiatkowski, J. Palomaki, O. Redfield, et al. Natural Questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452-466, 2019.

O. Lieber, B. Lenz, H. Bata, et al. Jamba: A hybrid Transformer-Mamba language model. *arXiv:2403.19887*, 2024.

J. Li, M. Wang, Z. Zheng, and M. Zhang. LooGLE: Can long-context language models understand long contexts? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 16304-16333, 2024.

N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157-173, 2024.

NVIDIA. Nemotron-H: A family of accurate and efficient hybrid Mamba-Transformer models. *arXiv:2504.03624*, 2025.

A. Peysakhovich and A. Lerer. Attention sorting combats recency bias in long context language models. *arXiv:2310.01427*, 2023.

B. Peng, E. Alcaide, Q. Anthony, et al. RWKV: Reinventing RNNs for the Transformer era. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 14048-14077, 2023.

M. Poli, S. Massaroli, E. Nguyen, et al. Hyena hierarchy: Towards larger convolutional language models. In *International Conference on Machine Learning*, volume 202 of PMLR, pages 28043-28078, 2023.

O. Press, N. A. Smith, and M. Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*, 2022.

P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of EMNLP*, pages 2383-2392, 2016.

U. Shaham, M. Ivgi, A. Efrat, J. Berant, and O. Levy. ZeroSCROLLS: A zero-shot benchmark for long text understanding. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 7977-7989, 2023.

F. Shi, X. Chen, K. Misra, N. Scales, D. Dohan, E. H. Chi, N. Schärli, and D. Zhou. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, volume 202 of PMLR, pages 31210-31227, 2023.

D. Soboleva, F. Al-Khateeb, R. Myers, J. R. Steeves, J. Hestness, and N. Dey. SlimPajama: A 627B token cleaned and deduplicated version of RedPajama. Cerebras dataset card and release, 2023.

J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

M. Sun, X. Chen, J. Kolter, and Z. Liu. Massive activations in large language models. In *Conference on Language Modeling*, 2024.

A. Vaswani, N. Shazeer, N. Parmar, et al. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

R. Waleffe, W. Byeon, D. Riach, et al. An empirical study of Mamba-based language models. *arXiv:2406.07887*, 2024.

G. Xiao, Y. Tian, B. Chen, S. Han, and M. Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, 2024.

F. Xu, W. Shi, and E. Choi. RECOMP: Improving retrieval-augmented LMs with compression and selective augmentation. In *International Conference on Learning Representations*, 2024.

A. Yang, B. Yang, B. Zhang, et al. Qwen2.5 technical report. *arXiv:2412.15115*, 2024.

A Supplementary evidence and artifacts

This appendix maps every experimental phase to its executed evidence and includes the most informative committed plots that do not fit in the main paper. All figures are generated from the cited JSON summaries by one deterministic plotting script. The canonical figure files are SVG; the PDF companions embedded here are produced from the same plotting calls because standard LaTeX engines do not include SVG directly.

A.1 Phase and artifact ledger

Table 3 separates executed measurements from their allowed interpretation. The standard ten-position sweep contains ten prompts per question and uses the same question bundle at every position.

The dataset snapshot contains 2,655 rows and has SHA-256 checksum `192a05b27af2b09eec33ca0c94bb5cf82bcaf70d78b3bdff1258df34bf37aab9`. The fixed split allocates 800 rows to exploration and 1,855 to confirmation. The confirmatory rows have not been evaluated.

Table 3: Executed evidence by phase. Sample counts are questions unless otherwise stated.

Phase	Executed comparison	Committed source	Interpretation in this paper
1	200 QA; 50 key-value	`artifacts/phase1`	Calibration and positive control
2	3 models, $n = 800$ each	`artifacts/phase2`	Primary family interaction
3	pure and hybrid 8B, $n = 800$	`artifacts/phase3`	Tightly matched composite contrast
4	5 size pairs, $n = 800$ each	`artifacts/phase4`	Descriptive scale trend
5	2 corpora, $n = 800$ each	`artifacts/phase5`	Corpus scope check
6	3 models, 50 at two lengths	`artifacts/phase6`	Synthetic task scope check
7	100-200 per intervention; 1,200 states	`artifacts/phase7-mechanisms`	Hypothesis generation only
8	3 systems, $n = 800$ each	`artifacts/phase8`	Descriptive prevalence check

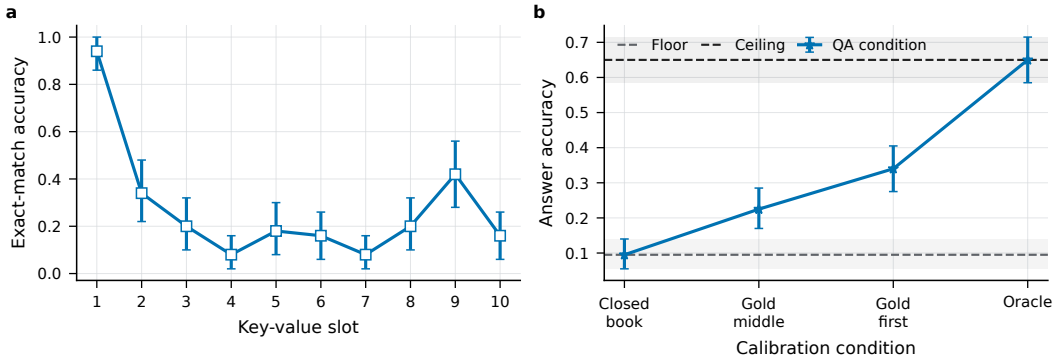


Figure 4: Phase 1 calibration. The left panel shows key-value retrieval by slot. The right panel shows the no-gold floor, middle-document condition, first-document condition, and gold-only ceiling for Pythia-2.8B.

A.2 Calibration and the executed positive control

Phase 1 calibrates the primary Pythia checkpoint before the cross-family comparison. The no-gold floor is 0.095, the gold-only ceiling is 0.650, and moving the gold document from the middle to the first position changes accuracy by $+0.115$ (95% CI $[0.055, 0.175]$). The key-value control produces a $+0.38$ edge-versus-middle effect, demonstrating that the position-intervention pipeline can recover an imposed positional signal. This calibration figure is a single-position contrast (first slot minus middle slot) on the 200-question Phase 1 subset, and is not directly comparable to the primacy statistic reported in Table 1, which averages positions $\{0, 1\}$ against $\{4, 5\}$ on the 800-question exploratory split; the two floor/ceiling pairs and the two edge sizes accordingly differ.

A.3 Scale and corpus scope checks

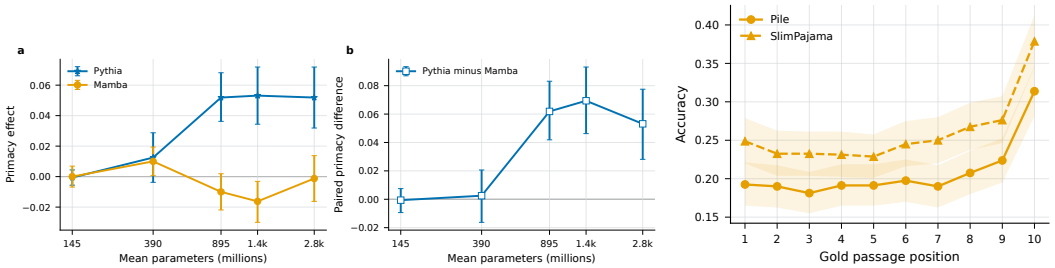
Figure 5 reports the complete Phase 4 and Phase 5 summaries. The scale trend is inseparable from task capability at the smallest sizes. The corpus comparison changes the accuracy level but does not detect an edge-shape difference under the fixed Mamba-2.8B architecture.

A.4 Synthetic retrieval

The Phase 6 forest plot in Figure 6 separates the QA estimates from 2K-token needle retrieval. Open markers denote the synthetic task and filled markers denote QA. Mamba and Mamba-2 saturate the needle task, so the absence of an edge effect for those models is a ceiling result rather than evidence of better middle-context behavior. Only the summary and plots for this phase are committed; raw generations and an environment manifest are not available in the repository snapshot.

A.5 Mechanism-oriented measurements and prompt interventions

Figure 7 combines the valid Phase 7 measurements. The sink statistic and probe results are associations, not interventions. The attempted hybrid attention-sink block is excluded because the model's custom attention path did not honor the hook, so the intervention did not change the computation.



(a) Approximate size pairs **(b) Fixed architecture, different corpora**

Figure 5: Additional scope checks. Markers and colors retain the paper-wide encoding: attention-based Pythia models are blue stars, while state-space models are orange and use shape to distinguish checkpoint variants.

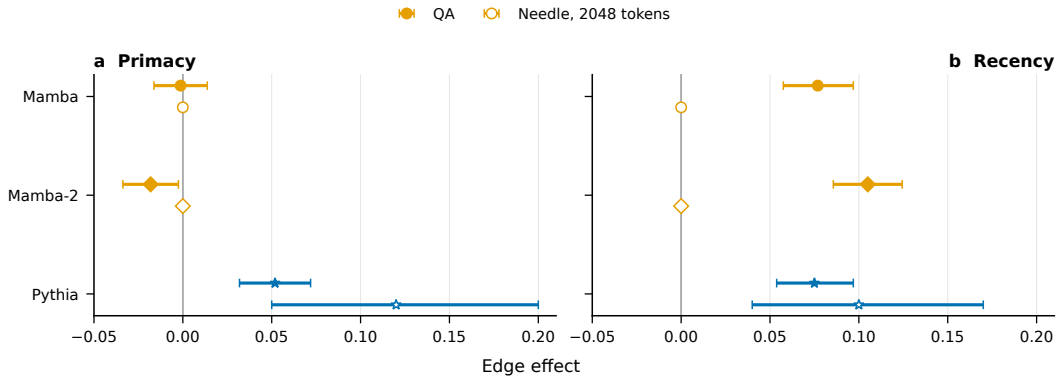


Figure 6: Primacy and recency across the main QA task and 2K-token RULER needle retrieval. Horizontal intervals are 95% question-bootstrap confidence intervals.

The prompt variants jointly alter query position, repetition, and token layout and therefore do not isolate one mechanism.

A.6 Additional 7B-8B systems

Figure 8 shows the full position curves for the descriptive system comparison. Marker shape and line style distinguish systems without adding colors. These checkpoints differ on several factors beyond the sequence mixer, so the figure is evidence of prevalence rather than a controlled architecture result.

B Integrity and provenance notes

The primary standard sweep yields 9,600 generations per model: 8,000 position prompts, 800 no-gold prompts, and 800 gold-only prompts. The paired bootstrap resamples complete question bundles and preserves prompt and distractor identity across positions. No generation failure or scored null is reported in the primary summaries. Prompt-length checks keep the compared position prompts within one token.

The Phase 3 result was independently re-run under a clean source tree at commit `33d6bb5`, reproducing the original summary exactly; see `artifacts/phase3/rerun-33d6bb5/REPORT.md`. A 50-generation blinded audit sample exists, but the human labels were not completed. The committed negative and distractor-order controls passed on the exploratory Pythia sample only and do not test the 8B Megatron checkpoints.

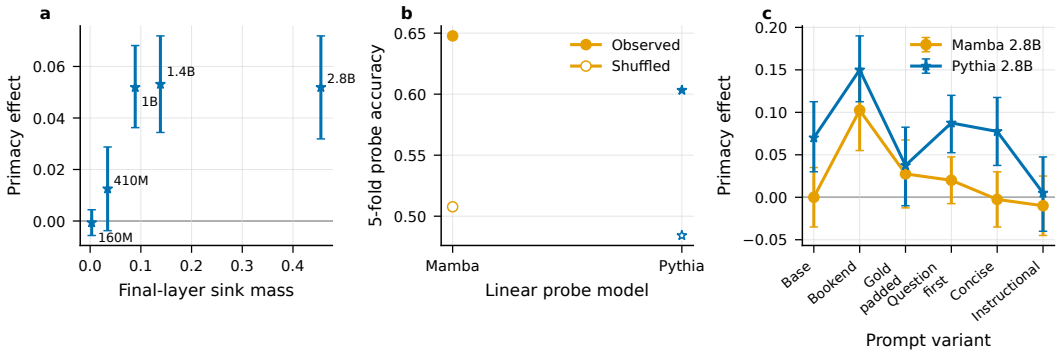


Figure 7: Phase 7 mechanism-oriented evidence. The sink panel compares final-layer initial-token attention share with primacy. The probe panel includes shuffled-label controls. The prompt panel reports edge effects for the executed query and template variants.

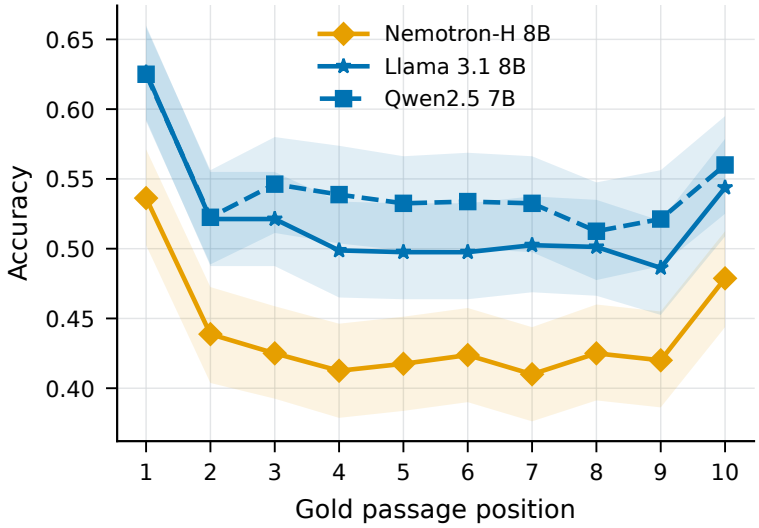


Figure 8: Position curves for Nemotron-H-8B, Llama-3.1-8B, and Qwen2.5-7B on the 800-question exploratory split.

B.1 Checkpoint provenance

Table 5 records the immutable revisions for the checkpoints that define the primary, matched, corpus, and additional-system comparisons. The complete scale-suite registry appears in `src/mixing_matters/models.py`.

Elapsed time and aggregate compute were not logged consistently enough to report a defensible total. Checkpoint repository identifiers and immutable revisions are recorded in `src/mixing_matters/models.py` and the per-run manifests. Some evaluated converted checkpoint mirrors do not declare complete license metadata in their model cards, so this snapshot does not claim a complete asset-license audit.

Table 4: Execution environments recorded by the phase artifacts.

Phases	Recorded accelerator	Scope
1 and 5	NVIDIA A10G	Calibration and corpus check
2-4	NVIDIA A40	Primary, matched, and scale comparisons
6	NVIDIA T4	Synthetic retrieval
7	NVIDIA L40S and T4	Prompt, sink, and probe studies
8	NVIDIA L40S	Additional 7B-8B systems

Table 5: Selected checkpoint repositories and pinned revisions.

Paper key	Repository identifier	Revision
Pythia-2.8B	`EleutherAI/pythia-2.8b`	`2a259cdd96a4`
Mamba-2.8B	`state-spaces/mamba-2.8b-hf`	`96c48e0292b6`
Mamba-2 2.7B	`AntonV/mamba2-2.7b-hf`	`ef542707386f`
Pure Mamba-2 8B	`nvidia/mamba2-8b-3t-4k`	`b915550c63ba`
Hybrid Mamba-2 8B	`nvidia/mamba2-hybrid-8b-3t-4k`	`35e8852e2240`
SlimPajama Mamba-2.8B	`state-spaces/mamba-2.8b-slimpj`	`a7bdd41af90c`
Nemotron-H-8B	`nvidia/Nemotron-H-8B-Base-8K`	`94ea861e008c`
Llama-3.1-8B	`meta-llama/Llama-3.1-8B`	`d04e592bb4f6`
Qwen2.5-7B	`Qwen/Qwen2.5-7B`	`d14972939875`

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 1 identify the exploratory family-by-position comparison, its depth, positional-encoding, recurrent-state-size, and checkpoint-family confounds, the uncertain matched 8B contrast, the prompt dependence, and the correlational status of the mechanism evidence.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 7 and Appendix B state the exploratory-only scope, the residual confounds, the uncertain paired 8B result, the Pythia-only certification-control scope, the missing audit labels, the synthetic-retrieval artifact gap, and the invalid sink intervention.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.

- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper is empirical and contains no theoretical results.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [No]

   Justification: Section 3 and Appendix A disclose the dataset, split, inference settings, scorer, phase samples, and analysis, but the synthetic retrieval check lacks committed raw generations and an environment manifest, so exact reconstruction of every reported result is not guaranteed.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.

- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The dataset and checkpoints are public and the repository contains the harness, runbooks, and raw outputs, but a fully anonymized submission archive has not yet been prepared.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: Section 3 gives the split, fixed generation settings, edge definitions, and bootstrap procedure; Appendix A gives every executed phase sample; no model training is performed beyond the frozen-state linear probe, whose layer is fixed a priori.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Reported edges and controlled contrasts include 95 percent question-level paired-bootstrap intervals from 10,000 resamples, with Holm correction within each primacy-and-recency pair, as defined in Section 3.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [No]

Justification: Appendix B identifies the A10G, A40, L40S, and T4 accelerators used by phase, but elapsed time and aggregate compute were not logged consistently enough to report a defensible total.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work is an analysis of released public models on a released public benchmark, with no human subjects and no new data collection.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Section 7 discusses how the diagnostics can support model evaluation and ordering strategies, and how position bias can make systems privilege evidence because of placement rather than relevance.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

Justification: The paper releases analysis code and evaluation outputs, not any high-risk model or scraped dataset.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [No]

Justification: The paper credits the benchmark, scorer, and model families and records exact checkpoint identifiers and revisions, but some evaluated converted checkpoint mirrors do not declare complete license metadata in their model cards, so a complete license audit is not claimed.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The repository documents the evaluation harness, per-phase runbooks, artifact ledger, plotting script, and source paths for the reported results; the outstanding anonymized submission archive is disclosed in the open-access answer.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.

- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper involves no crowdsourcing and no human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper involves no human subjects and therefore required no IRB review.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: Large language models are the objects of study, not a component of the method; no LLM was used as an original or non-standard part of the methodology.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.