The Parallax Principle

Position-Marginalized Decoding for Order-Robust Language Models

Badrinath Chandana[1], Ishan Jindal[2], Mahesh Chandran[2]

[1] Mohamed bin Zayed University of Artificial Intelligence [2] Fujitsu Research of India

NeurIPS 2026 Workshop
LP4FM Spotlight

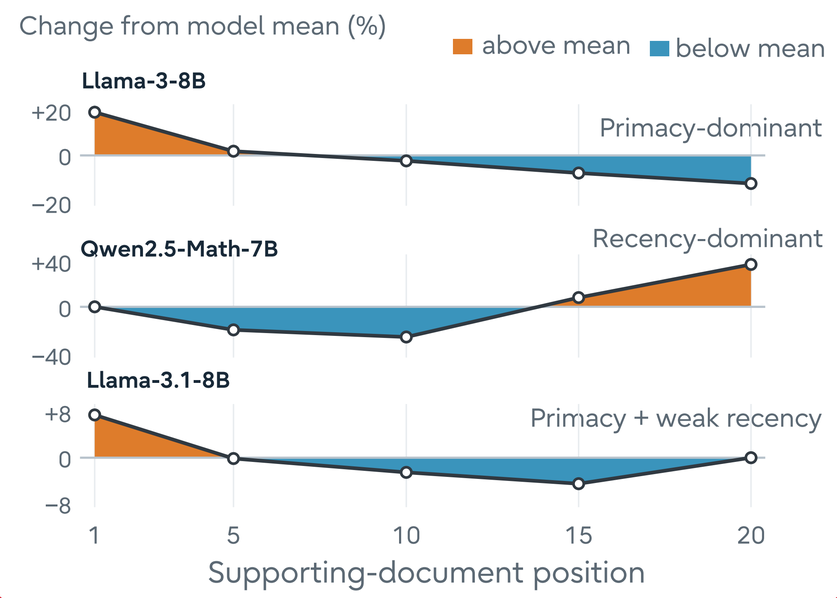
Introduction

Language models can give different answers when the same evidence appears in a different order. Some models favor early evidence, others late evidence. A single preferred ordering cannot cover every model and task, even when the document contents and question remain fixed.

Core Empirical Dilemma:
Models exhibit conflicting positional biases (e.g., strong Primacy in Llama-3/Qwen3 vs. strong Recency in Qwen2.5-Math). Consequently, heuristic reranking rules (e.g., "place gold first") fail to generalize across diverse LLMs

Can we make answers robust to evidence order without training the model?

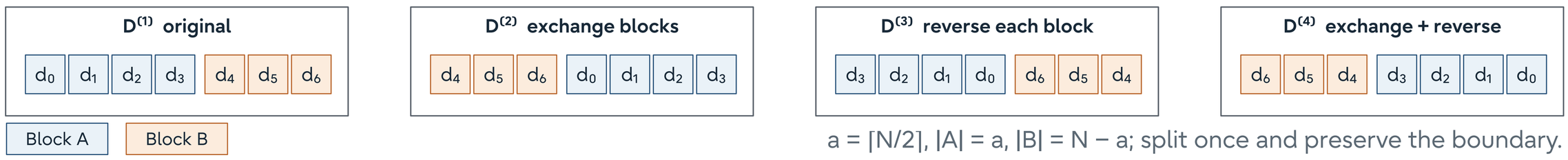
The Bias



The Solution: Position-Marginalized Decoding

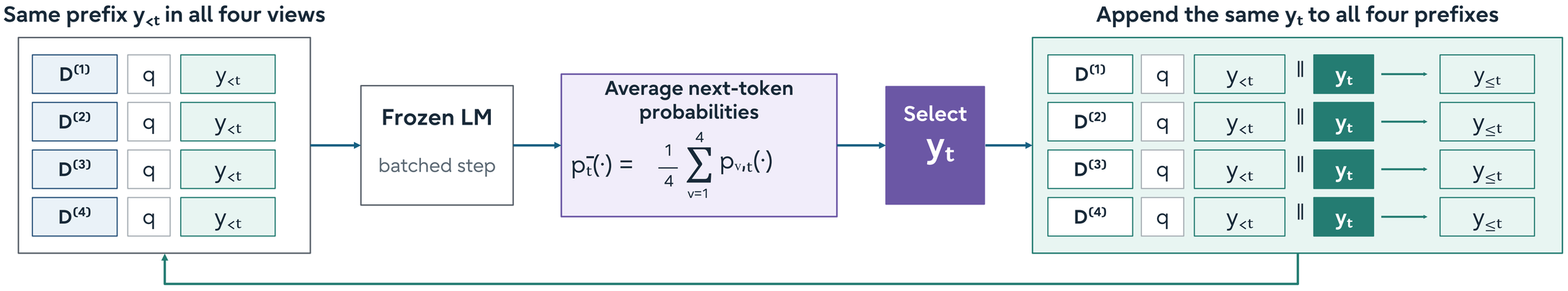
(a) Four complementary views from one fixed split

Example: N = 7, A = (d_0, … , d_3), B = (d_4, d_5, d_6)



a = ⌈N/2⌉, |A| = a, |B| = N − a; split once and preserve the boundary.

(b) Synchronized next-token decoding



next step: all four views share the updated prefix $y_{\le t}$

Every document has mean position (N − 1)/2 across the four views. This does not imply arbitrary permutation invariance.

1 Reorder

Split the documents once into A and B. Create four complementary orders. Every view retains the full context.

2 Average

Use a frozen LM for all four views. Average next-token probabilities at the same generated prefix.

3 Synchronize

Select one token from the average. Append it to every view, then repeat. The generated prefix stays identical.

Results

NQ: 20 documents, evidence at {1, 5, 10, 15, 20}.

Transfer: HotpotQA, MuSiQue, 2WikiMultiHopQA, SearchQA, MS MARCO.

1. **Mean accuracy:** +0.8–4.4 pp across seven backbones.
2. **Worst placement:** +1.9–9.4 pp.
3. **85.6–98.8%** less position-induced variance.
4. **15/15** transfer settings preserve or improve the worst placement.

Mean BS-EM (%). Worst Δ in pp. VR = variance reduction. Relative latency.

Backbone	Standard	PMD	Worst Δ	VR (%)	Latency
Llama-3-8B	61.96	**65.64**	+9.40	96.66	1.45×
Llama-3.1-8B	68.56	**69.36**	+1.90	86.03	1.45×
Llama-3.2-3B	56.96	**61.32**	+7.80	96.71	1.39×
Qwen3-4B	60.80	**63.06**	+5.10	94.82	1.52×
Qwen3-8B	62.63	**63.74**	+1.89	85.56	1.49×
Gemma-4-E4B	59.92	**62.38**	+4.70	98.78	1.92×
Qwen3.5-4B	49.60	**51.92**	+4.10	92.40	1.85×

Cost: four streams, four KV caches, ≈4× compute.
Measured latency: 1.39–1.92× Standard (H100, 64 tokens).

- ❑ **Optimal Overall Trade-off:** Highest mean BS-EM and lowest measured wall-clock latency among all position-robust decoders.
- ❑ **Fast Execution:** 1.45×–1.52× latency on single-H100, far outperforming RoToR (2.33×–4.35×) and PINE (2.86×–5.88×).

Pareto Dominance Frontier

PMD establishes a strictly superior accuracy–latency frontier, delivering robust position invariance with minimal serving overhead and zero internal model modification.

85.6–98.8%
less position-induced variance

+1.9–9.4 pp
higher worst-placement accuracy

15/15 transfer settings preserve or improve worst-placement accuracy

HotpotQA, MuSiQue, 2WikiMultiHopQA, SearchQA, MS MARCO

References

1. Liu, N. F.; Lin, K.; Hewitt, J.; Paranjape, A.; Bevilacqua, M.;Petroni, F.; and Liang, P. 2024. Lost in the Middle: How Language Models Use Long Contexts. Transactions of the Association for Computational Linguistics, 12.
2. Yoon, S.; Ahn, D.; Lee, Y.; Jung, M.; Jang, H.; and Hwang,S.-w. 2025. RoToR: Towards More Reliable Responses for Order-Invariant Inputs. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 18739–18760.
3. Yu, Y.; Jiang, H.; Luo, X.; Wu, Q.; Lin, C.-Y.; Li, D.; Yang, Y.;Huang, Y.; and Qiu, L. 2025. Mitigate Position Bias in LLMs via Scaling a Single Hidden States Channel. In Findings of the Association for Computational Linguistics: ACL 2025,6092–6111.
4. Zhang, Z.; Chen, R.; Liu, S.; Yao, Z.; Ruwase, O.; Chen,B.; Wu, X.; and Wang, Z. 2024. Found in the middle: How language models use long contexts better via plug-and-play positional encoding. Advances in Neural Information Processing Systems, 37: 60755–60775.