Decipher Deep Math: Numeric Rounding Behaviors in LLMs

Anonymous authors

I. Introduction

Approximation constitutes a fundamental component of empirical scientific discovery [1], and its significance is further amplified by the emergence of hypotheses generated by large language models (LLMs) [2]. However, mechanistic interpretability studies that examine the capacity of LLMs to perform estimation remain less explored. Building on prior work [3], we investigate whether LLMs exhibit human-like rounding strategies (e.g., rounding to the nearest multiple of 5 or 10). We conduct a systematic analysis using linear probes [4], across LLMs of varying sizes and architectures—autoregressive (AR), diffusion (Diff), and state-space (SS). Specifically, we examine tasks involving ”near 5” and ”near 10” estimation to characterize their rounding behavior. Our findings suggest early stopping opportunities [10] may enable reduced computational cost while still arriving at rounded outputs. Future work includes pursuing causal analyses to identify neurons implicated in rounding operations. Additionally, we plan to develop applications that leverage rounding-based strategies for faster inference.

II. Results

We design a task that requires detecting whether a given number is close to a multiple of n. Specifically, we set n=10 and n=5, and train linear probes to predict this property from hidden states. Figure 1 reports classification accuracy across layers, and Table I, II presents error rates at different rounding distances (measured at the best-performing layer) for a range of models: Qwen3-4B-Instruct-2507 (NT) [5], Qwen3-4B-Thinking-2507 (T), Dream-org/Dream-v0-Instruct-7B [6], as well as the state-space models Mamba-1.4B and Mamba-2.8B [7].

Both autoregressive (AR) and diffusion (Diff) models exhibit an early emergence of strong accuracy. In the later layers, Diff achieves the best accuracy, whereas AR achieve second best accuracy. This pattern is consistent across tasks and model configurations (thinking vs. non-thinking, architectural differences). Furthermore, we hypothesize that AR and Diff models achieve early performance peaks through rapid initial processing, followed by refinement attempts in intermediate layers. This observation underscores the presence of internal shortcuts and highlights opportunities for early stopping strategies. In contrast, the Mamba models show lower and oscillating accuracies across layers, independent of model size or task type, with no evidence of refinement. Table I, II further indicates that the Mamba models display binary behavior across all tasks, achieving either complete correctness or total failure.

Across most AR and Diff experiments, the highest error rates occur at rounding distance 1 (numbers one unit away from multiples of 5 or 10), regardless of task type. This pattern is significant because it reveals how models process numerical proximity: numbers like 9, 11, 14, or 16 are most likely to be misclassified in rounding tasks, while exact multiples (distance 0) and numbers further away (distance 2+) are classified more accurately. We hypothesize that models struggle particularly with these boundary cases due to digit-based number encoding [8] and calibration biases [9]. The error pattern has an exception for Qwen3-4B-Instruct-2507 in Near-5 tasks where errors increase with distance. Also, Qwen3-4B-Thinking and Dream-7B show superior Near-10 performance, while Qwen3-4B-Instruct-2507 excels in Near-5 tasks, indicating model-specific rounding strategies that may inform improved approximation.

TABLE I
Comprehensive Near-5 Analysis: Performance and Error Patterns at the best layer. Acc = Accuracy; Err = Error rate

Model	Peak Acc	Best Layer	Err (0)	Err (1)	Err (2)
Qwen3-4B-Instruct-2507	0.939	2	0.4%	5.5%	9.4%
Qwen3-4B-Thinking-2507	0.917	6	7.2%	14.6%	2.5%
Dream-7B	0.970	26	4.2%	4.8%	0.5%
Mamba-2.8B-HF	0.597	1	0.0%	0.0%	100.0%
Mamba-1.4B-HF	0.597	1	0.0%	0.0%	100.0%

TABLE II
Comprehensive Near-10 Analysis: Performance and Error Patterns at the Best Layer. Best: Best Layer

Model	Peak Acc	Best	Err (0)	Err (1)	Err (2)	Err (3)	Err (4+)
Qwen3-4B-Instruct-2507	0.967	8	4%	12%	1%	1%	0%
Qwen3-4B-Thinking-2507	0.987	7	1%	3%	3%	0%	1%
Dream-7B	0.988	24	2%	5%	0%	0%	0%
Mamba-2.8B-HF	0.697	1	100%	100%	0%	0%	0%
Mamba-1.4B-HF	0.697	2	100%	100%	0%	0%	0%

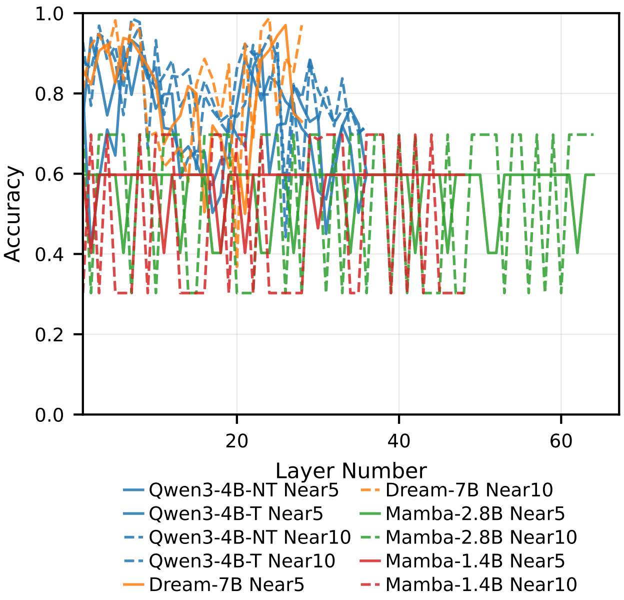

Fig. 1. Linear Probe Accuracy Across All Models and All Digit-Tasks.

[1] [2]

References

[1] S. Lutz, “Generalizing empirical adequacy I: multiplicity and approximation,” *Synthese*, vol. 191, no. 10, pp. 3195–3225, 2014.

[2] A. Abdel-Rehim, H. Zenil, O. Orhobor, M. Fisher, R. J. Collins, E. Bourne, G. W. Fearnley, E. Tate, H. X. Smith, L. N. Soldatova, and R. D. King, “Scientific hypothesis generation by large language models: laboratory validation in breast cancer treatment,” *Journal of the Royal Society Interface*, vol. 22, no. 227, 2025, doi: 10.1098/rsif.2024.0674.

[3] M. Li, J. Yang, and X. Ye, “Children’s number line estimation strategies: evidence from bounded and unbounded number line estimation tasks,” *Frontiers in Psychology*, vol. 15, Nov. 2024, doi: 10.3389/fpsyg.2024.1421821.

[4] G. Alain and Y. Bengio, “Understanding intermediate layers using linear classifier probes,” in *ICLR Workshop Track*, 2016.

[5] Qwen Team, “Qwen3 Technical Report,” *arXiv preprint arXiv:2505.09388*, 2025. [Online]. Available: https://arxiv.org/abs/2505.09388

[6] J. Ye, Z. Xie, L. Zheng, J. Gao, Z. Wu, X. Jiang, Z. Li, and L. Kong, “Dream 7B: Diffusion large language models,” *arXiv preprint arXiv:2508.15487*, 2025.

[7] A. Gu and T. Dao, “Mamba: Linear-time sequence modeling with selective state spaces,” *arXiv preprint arXiv:2312.00752*, 2023.

[8] A. A. Levy and M. Geva, “Language models encode numbers using digit representations in base 10,” *arXiv preprint arXiv:2402.00157*, 2024.

[9] C. Lovering, M. Krumdick, V. D. Lai, V. Reddy, S. Ebner, N. Kumar, R. Koncel-Kedziorski, and C. Tanner, “Language Model Probabilities are Not Calibrated in Numeric Contexts,” in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds. Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 29218–29257. doi: 10.18653/v1/2025.acl-long.1417. [Online]. Available: https://aclanthology.org/2025.acl-long.1417/

[10] S. Teerapittayanon, B. McDanel, and H. T. Kung, “BranchyNet: Fast inference via early exiting from deep neural networks,” *arXiv preprint arXiv:1709.01686*, 2017. [Online]. Available: http://arxiv.org/abs/1709.01686

[1] We open source the results and tools in https://github.com/ctseng777/Decipher-Deep-Math-in-Rounding

[2] We use Claude to help with scripting and ChatGPT with Latex formatting