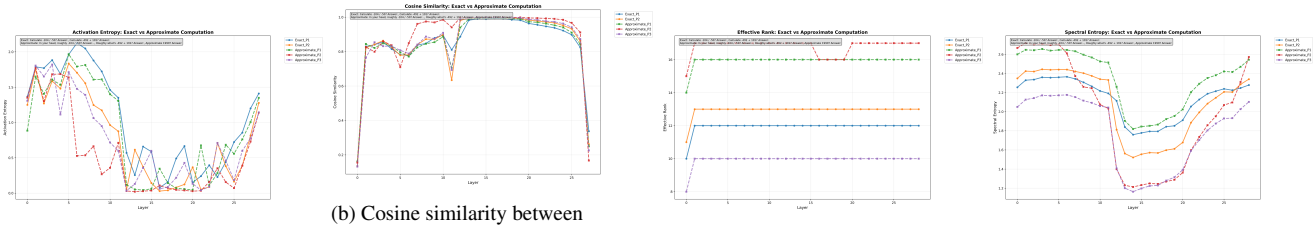
# Dream Diary: Case Study on Diffusion LLM's Arithmetic Behavior

Mechanistic interpretability studies of autoregressive (AR) models are abundant, while studies on diffusion models (DLLM) remain less explored. In this study, we investigate the arithmetic behaviors of Dream-v0-Instruct-7B (Dream)[5].

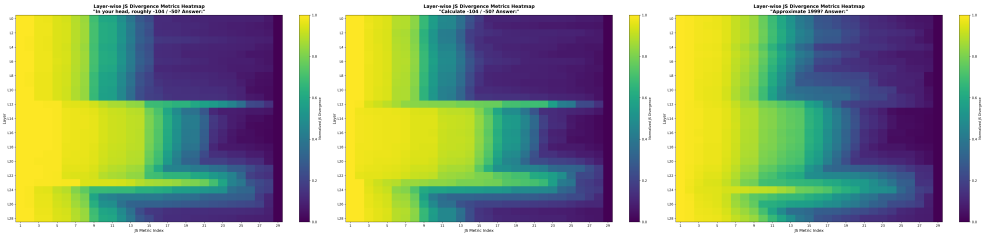
We evaluate Dream against 5 arithmetic cases with curated prompts[2]: exact fraction (Exact_P1; *Calculate -104 / -50? Answer:*), exact addition (Exact_P2; *Calculate -492 + 191? Answer:*), approximated fraction (Approximate_P1; *In your head, roughly -104 / -50? Answer:*), approximated addition (Approximate_P2; *Roughly what's -492 + 191? Answer:*) and rounding (Approximate_P3; *Approximate 1999? Answer:*). We observe a distinct turning point around layer 11 in both activation entropy and spectral entropy, echoing earlier findings of early answer convergence [2], though in our case this occurs later, around 40% of the decoding depth rather than within the first quarter. The entropy patterns are similar between exact-arithmetic cases and approximated-arithmetic cases; Jensen-Shannon divergences [3] measured at each layer also suggest a similar energy distribution between the two cases. The effective rank study [4] shows similar dimensionalities of approximated fraction and approximated addition cases compared to their exact arithmetic counterparts. These findings imply that the approximation does not save computing costs for DLLM. This is contrary to how a human uses an approximation for easier computation [6]. However, the rounding case shows lower effective rank and lower JS divergence in the mid layers. We hypothesize the model reaches the answer faster due to the simplicity of the case.

Future work includes causal study of DLLM to isolate the arithmetic neurons [1], particularly approximation operations, extending the evaluation to larger benchmarks to gain statistical significance and providing mechanistic interpretability study tools to the community.



(a) Activation entropy in each layer (b) Cosine similarity between consecutive layers (c) Effective ranks in each layers (d) Spectral entropy in each layer

Figure 1: Cross layer comparisons



(a) Approximate fraction (b) Exact fraction (c) Approximate rounding

Figure 2: Heatmap of Jensen–Shannon divergence in each layer

[1] B. R. Christ, Z. Gottesman, J. Kropko, and T. Hartvigsen. Math neurosurgery: Isolating language models' math reasoning abilities using only forward passes, 2025. [2] P. Li, Y. Zhou, D. Muhtar, L. Yin, S. Yan, L. Shen, Y. Liang, S. Vosoughi, and S. Liu. Diffusion language models know the answer before decoding, 2025. [3] M. L. Men ´endez, J. A. Pardo, L. Pardo, and M. d. C. Pardo. The jensen-shannon divergence. Journal of the Franklin Institute, 334(2):307–318, 1997. [4] O. Roy and M. Vetterli. The effective rank: A measure of effective dimensionality. In 2007 15th European Signal Processing Conference, pages 606–610, 2007. [5] J. Ye, Z. Xie, L. Zheng, J. Gao, Z. Wu, X. Jiang, Z. Li, and L. Kong. Dream 7b, 2025. [6] Jiang DL, Ye S, Zhao L, Gu B. Do reductions in search costs for partial information on online platforms lead to better consumer decisions? Evidence of cognitive miser behavior from a natural experiment. Information Systems Research, 2025, pp. 1–19. doi:10.1287/isre.2022.0432

We ackowledge the use of AI for scripting and Latex formatting