Fast & Faithful: Diffusion Drift
Do Accelerated Diffusion Language Models Reason Faithfully?

Anirud Aggarwal[*] **Omkar Pathak**[*] **Nayana Gadde**[*]
University of Maryland
{anirud, omkarp07, ngadde9}@umd.edu

Abstract

Recent work has explored diffusion language models (DLMs) as an alternative to autoregressive (AR) generation for reasoning tasks, yet little is known about the faithfulness of their intermediate reasoning trajectories. This study introduces a preliminary framework for measuring Diffusion Chain-of-Thought (DoT) faithfulness and provides an initial empirical analysis using the LLaDA-8B model and its accelerated variant, dLLM-Cache. Using trajectory-level linear probes on the GSM8K benchmark, we examine how answer-relevant information emerges and evolves across diffusion steps, and how caching affects this process. Results show that correctness information appears early in the diffusion trajectory, accumulates over time, and remains largely preserved under acceleration with only modest degradation. While limited to a single acceleration method and probing-based evaluation, these findings provide early evidence that DLM reasoning dynamics can retain causal coherence under efficiency-oriented modifications. Future work will extend this framework with further diagnostics and acceleration methods to build a more complete understanding of faithfulness in diffusion-based reasoning.

1 Introduction

Chain-of-thought (CoT) prompting has become standard practice during evaluation and deployment of AR large language models (LLMs). Prior work has shown that both prompting the LLM to generate intermediate reasoning steps and explicitly training it to do so lead to performance improvements (Chen et al., 2025). However, a growing body of work demonstrates these explanations are often unfaithful, meaning the final answer is not causally mediated by the produced CoT (Jacovi and Goldberg, 2020; Lanham et al., 2023; Paul et al., 2024). This makes CoTs a poor basis for interpretability and undermines trust, as users may assume the rationale reflects the true internal computation.

DLMs extend diffusion-based generative modeling to discrete text (Li et al., 2022; Nie et al., 2025; Dong et al., 2023). Rather than sampling tokens left-to-right, DLMs iteratively denoise a masked sequence to produce tokens (we note there are many other approaches to diffusion language modeling). This sampling paradigm naturally exposes a rich state trajectory over time, including intermediate partial sequences that can be interpreted as "latent thoughts." Recent work has begun to exploit this structure for reasoning by diffusing over rationales before answers (Ye et al., 2024a) or by monitoring early answer convergence during the denoising process (Li et al., 2025). Yet it remains unclear whether the DoT explanations produced by DLMs are faithful.

At the same time, DLMs are computationally expensive due to repeated forward passes at each diffusion step. This has motivated a line of train-free acceleration techniques designed specifically for DLMs, including cross-step caching of representations (Liu et al., 2025b; et al., 2025), DLM-targeted post-training quantization (Xu and Yang, 2025), and step-adaptive early stopping (Li et al., 2025). While these methods report substantial speed and memory gains, they are evaluated almost exclusively on standard metrics (accuracy, perplexity, BLEU, etc.). This leaves open whether they alter the causal dependence between reasoning traces and answers. Do such accelerations preserve or erode the faithfulness of DoTs?

This work takes a first step toward answering this question by studying DoT faithfulness and its sensitivity to train-free acceleration methods. Building on existing causal and intervention-based notions of faithfulness developed for AR LLMs (Jacovi and Goldberg, 2020; Lanham et al., 2023; Paul et al., 2024), we adapt these ideas to the diffusion setting. In particular, reasoning unfolds across tokens and

[*] Equal contribution.

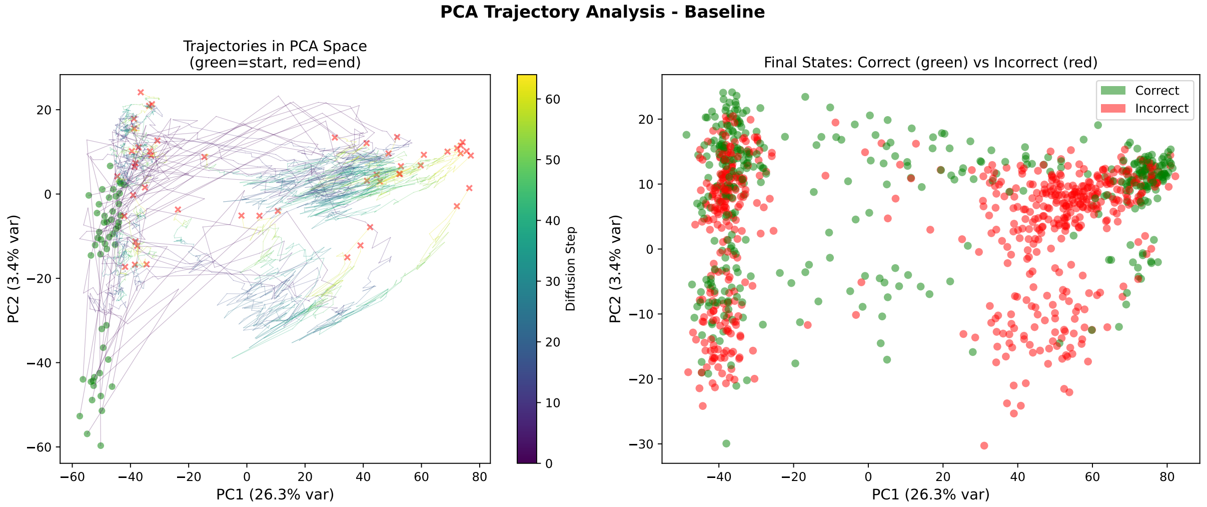

Figure 1: Overview of reasoning dynamics in diffusion language models. PCA projections of hidden-state trajectories from the baseline LLaDA-8B model show that correct (green) and incorrect (red) examples gradually diverge as denoising progresses, illustrating how answer-relevant structure emerges over diffusion time. This pattern underlies analysis of DoT faithfulness and forms the basis for measuring faithfulness throughout a diffusion trajectory.

the diffusion time dimension, providing an additional axis for exploration. We empirically compare an uncompressed AR and DLM baseline to their accelerated variants on reasoning benchmarks, analyzing how each acceleration method affects their reasoning traces. Our key contributions are: (1) a conceptual framework for DoT faithfulness that connects AR causal definitions to the iterative denoising process; and (2) the first, to our knowledge, empirical study of the faithfulness of DoT.

2 Related Works

Faithfulness and CoT in AR LLMs While specific definitions vary, a faithful CoT is one where the model's stated reasoning truly reflects the model's internal decision process, rather than merely supplying a plausible answer. Jacovi and Goldberg's (2020) work provides clear guidelines on distinguishing plausibility from faithfulness. A faithful CoT is not only plausible, but causally responsible for the model's final prediction. Lanham et al.'s (2023) work formalizes faithfulness by intervening on intermediate reasoning steps and examining the change in final model answer. Through interventions such as early-answering, mistake injecting, and more, they find that while CoTs boost accuracy, they are often unfaithful. However, DLMs have shown robustness to errors in CoTs (Ye et al., 2024b; Cetin et al., 2025), which may introduce a confounding factor in such interventions. That is, did a DoT recover from the injected mistake in order to still predict the correct response? Rather than intervening on the intermediate reasoning traces, other approaches such as Turpin et al. (2023), make counterfactual edits to the input prompts only and find AR LLMs present plausible but consistently unfaithful CoTs.

Still other works attempt to attribute LLMs' internal states to portions of the CoT. That is, if the model reasons step-by-step, then intermediate steps should strongly influence the next steps and the final answer. Recent work by (Helbling et al., 2025) presents a train-free method of obtaining high-quality saliency maps in Diffusion Transformers for image generation, a method that shows plausibility in an application to DLMs. While these assist in understanding the model's internal workings, they demonstrate correlation rather than causation.

DLMs and reasoning Recent work treats DLM denoising trajectories as explicit reasoning processes. Ye et al. (2024b) propose Diffusion of Thoughts, which iteratively refines intermediate "thought" states over diffusion steps and decodes the final state to text, naturally supporting self-correction and a compute–accuracy trade-off via the diffusion schedule. Huang et al. (2025) introduce DCoLT, which frames reverse-diffusion steps as latent "thinking actions" and optimizes the full trajectory with outcome-based reinforcement learning, leveraging bidirectional, non-causal attention and unconstrained intermediate states to encourage lateral exploration. Kang et al. (2025) propose LaDiR, a latent diffusion model that compresses reasoning into VAE "blocks of thought" and refines them with blockwise bidirectional attention, enabling parallel exploration of candidate trajectories and adaptive test-time compute. Shao et al. (2025) present Diffuse Thinking, a hybrid framework where a DLM proposes diverse intermediate thoughts in parallel and an AR LLM selects among them, using diffusion primarily as an efficient generator of candidate reasoning paths.

More broadly, this line of work connects to latent chain-of-thought and diffusion-based reasoning, where multi-step inference is carried out in hidden representations rather than explicit rationales (He et al., 2024; Chen et al., 2025). Across these approaches, trajectories are treated as structured objects to refine, optimize, or diversify. However, evaluations largely emphasize answer accuracy, diversity, or qualitative interpretability, leaving open whether intermediate diffusion-time "thoughts" are causally necessary for the final prediction, and how inference-time modifications (e.g., caching, quantization, early stopping) alter that dependence. We address this gap by analyzing DLM trajectories through a causal lens and measuring how train-free accelerations modulate the relationship between diffusion-time states and final outputs.

Table 1: Model performance and efficiency on GSM8K. We compare the baseline model (LLaDA-8B) against dLLM-Cache. Accuracy is measured with exact matches, and latency is averaged over 500 samples.

Model	Accuracy	Latency (Speedup)
LLaDA-8B	39.70%	6.31s (1.00×)
dLLM-Cache	36.80%	3.54s (1.78×)

Train-free acceleration for diffusion LMs and LLMs Caching-based, train-free acceleration has recently emerged as a powerful technique for diffusion models in vision, particularly DiT-style architectures, delivering large speedups with minimal quality degradation (Aggarwal et al., 2025; Liu et al., 2025a; Wimbauer et al., 2023). These ideas have since been adopted in diffusion language models, where methods such as dLLM-Cache (Liu et al., 2025b) and dKV-Cache (et al., 2025) reuse or partially update representations across diffusion steps to reduce redundant computation. However, these methods currently lack a clear analysis of how caching alters the underlying reasoning process.

Similarly, Post-Training Quantization (PTQ), now commonplace for AR LLMs (Frantar et al., 2022; Yao et al., 2022), struggles with DLMs, where iterative denoising amplifies quantization noise (Xu and Yang, 2025). Although tailored quantizers (Xu and Yang, 2025) and step-adaptive methods like Prophet (Li et al., 2025) have emerged to reduce costs, their effect on the relationship between DoTs and final answers remains unknown. Our work aims to evaluate these acceleration techniques through the lens of faithfulness.

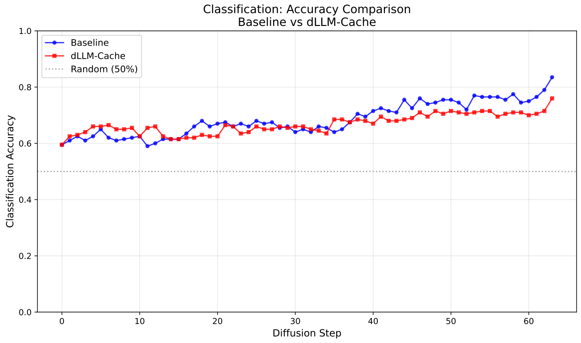

Figure 2: Classification probe accuracy over diffusion steps. The probe predicts answer correctness from mean-pooled hidden states. Both models show increasing accuracy in later steps, with the baseline achieving higher peak accuracy (83.5% vs 76.0% at step 63).

3 Method

We investigate the faithfulness of DoTs and how train-free acceleration techniques affect this property in three stages: (1) formalizing faithfulness in DLMs, (2) developing trajectory-level diagnostics to measure it, and (3) applying these diagnostics to uncompressed and accelerated DLMs.

Faithfulness in Diffusion Language Models We define a DLM as *faithful* if its intermediate denoising states encode the causal information necessary for producing the final answer, and if perturbing these states produces predictable, semantically coherent changes in the output. In other words, a DoT is faithful *if and only if its intermediate trajectory lies on the model's minimal causal path* from input to prediction.

Unlike AR models, DLMs denoise bidirectionally and can overwrite errors in later steps. This property complicates the use of standard CoT interventions such as mistake injection or token deletion (Lanham et al., 2023). Faithfulness in DLMs therefore requires examining the temporal evolution of latent representations rather than the surface text alone. Effective metrics must account for redundancy and smoothing over diffusion time, such that robustness is not mistaken for causal irrelevance.

Table 2: Summary of probe performance metrics. Classification accuracy measures the percentage of correctly classified answer predictions by the probe; R^2 and RMSE correspond to regression probes directly predicting numeric answers.

Model	Classification Accuracy		R^2 Score	
	Mean	Best (Step)	Mean	Best (Step)
LLaDA-8B	68.45%	83.50% (63)	-45.25	-16.67% (1)
dLLM-Cache	66.84%	76.00% (63)	-38.24	-15.19% (1)

Trajectory-Level Diagnostics We propose two complementary diagnostics tailored to the diffusion setting, and include experiments utilizing the first, leaving the second to future work.

(a) Linear Probing. For each diffusion timestep t, we train lightweight linear probes to predict the model's final answer from the intermediate latent state x_t. The accuracy of these probes indicates how early answer-relevant information emerges in the trajectory. If early timesteps already yield near-oracle predictions—or if such information appears only in the final steps—this suggests that the visible DoT does not reflect a genuine reasoning sequence. Instead, it may represent early answer storage followed by decorative denoising. Probe-sharpness curves over time thus reveal whether reasoning unfolds iteratively or is merely post hoc.

(b) Counterfactual Perturbation. We inject targeted semantic perturbations into intermediate latent states x_t, such as flipping a logical predicate or altering a partial computation. We then measure whether these changes (1) persist and influence the final output or (2) are overwritten by later denoising steps. High counterfactual sensitivity implies that x_t is causally upstream of the output, whereas insensitivity suggests that the trajectory segment is non-causal or redundant.

Comparing Uncompressed and Accelerated DLMs In this study, we focus our empirical analysis on the baseline DLM and its accelerated variant employing cross-step caching (dLLM-Cache). Caching reuses intermediate representations across diffusion steps to reduce redundant computation, potentially skipping reasoning-relevant updates. This setting allows us to examine whether such acceleration alters the causal relationship between intermediate denoising states and final outputs.

For both models, we record the full trajectory $\{x_t\}_{t=1}^{T}$ and apply our primary diagnostic—trajectory-level probing—to assess the emergence and persistence of answer-relevant information over diffusion time. These initial results establish a foundation for understanding how caching affects reasoning dynamics and faithfulness.

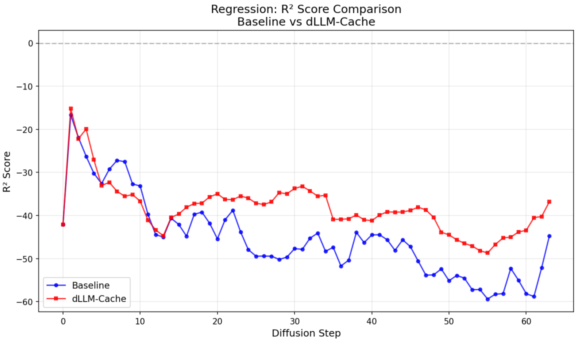

Figure 3: Regression probe R^2 scores over diffusion steps. Negative R^2 indicates poor linear predictability of numeric answers. Both models show negative R^2 throughout, with dLLM-Cache achieving slightly better (less negative) scores, particularly in early steps.

While the present work centers on dLLM-Cache, future research will extend this framework to include additional acceleration strategies, such as post-training quantization and step-adaptive early stopping, as well as the remaining diagnostic (counterfactual perturbation). These forthcoming analyses will enable a more comprehensive characterization of how different train-free accelerations influence the causal coherence of diffusion-based reasoning.

4 Experimental Details

Our primary model is LLaDA-8B, a discrete diffusion-based large language model trained for general text and reasoning tasks (Nie et al., 2025). We use the official implementation and pretrained checkpoints, interpreting the denoising trajectory as a Diffusion Chain-of-Thought (DoT). Unless otherwise specified, we adopt the default linear noise schedule with $T = 64$ reverse steps. All model weights remain frozen; only inference-time procedures are modified.

We conduct experiments on the GSM8K math word problem dataset (Cobbe et al., 2021), which contains 7.5k training and 1.3k test examples. We follow the LM-evaluation harness and official LLaDA/dLLM-Cache setups, formatting each example as a question followed by an "Answer:" prompt. For diffusion inference, we adopt the GSM8K prompt templates and decoding scripts provided in the dLLM-Cache repository for

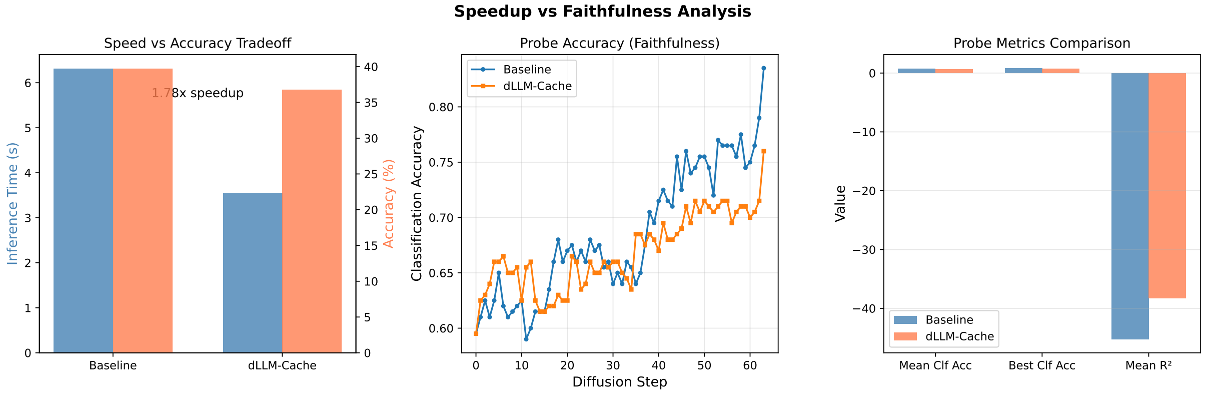

Figure 4: Speed-faithfulness tradeoff for dLLM-Cache. The model achieves 1.78× speedup while preserving most faithfulness metrics, suggesting caching maintains the causal structure of reasoning.

LLaDA-8B. We reserve 1,000 randomly sampled training instances as a validation set for hyperparameter selection and probe calibration. For evaluation, we extract the final numeric answer as the last integer or decimal in the model's output and compare it with the gold solution provided by GSM8K; examples with non-numeric answers (fewer than 1%) are excluded. Each sample is decoded for up to 512 tokens, with temperature values selected from $\{0.0, 0.5, 1.0\}$ based on validation performance.

All experiments are implemented in PyTorch 2.x and run on NVIDIA A5000 GPUs under `bf16` precision. We report mean and standard error across three random seeds, measuring both exact-match accuracy and average wall-clock time per example.

Baseline and Accelerated Model. We evaluate the unmodified LLaDA-8B baseline alongside its accelerated variant using dLLM-Cache (Liu et al., 2025b). dLLM-Cache reuses intermediate representations across diffusion steps to reduce redundant computation. We adopt the authors' default adaptive block-level caching configuration and record intermediate diffusion states for trajectory-level analysis. For faithfulness evaluation, we extract mean-pooled hidden states from the final transformer layer at each timestep and train lightweight linear probes: classification probes to predict correctness and regression probes to estimate numeric outputs, trained using logistic and Ridge regression, respectively, with hyperparameters selected via cross-validation.

Future Extensions. Although this study focuses on LLaDA-8B and dLLM-Cache, our framework generalizes to other acceleration methods. Future work will include *post-training quantization* (Xu and Yang, 2025), applying weight-only 4-bit quantization with calibration on GSM8K examples, and a *step-adaptive early exit* scheme (Li et al., 2025), which halts diffusion once answer stability and confidence thresholds are met. We also plan to extend our faithfulness diagnostics to include counterfactual perturbations, providing a more comprehensive characterization of how inference-time acceleration impacts diffusion-based reasoning.

5 Results

We analyze the faithfulness of DoT reasoning using the LLaDA-8B diffusion language model and its accelerated variant, dLLM-Cache, on the GSM8K mathematical reasoning benchmark. Our evaluation addresses two main questions: (1) to what extent intermediate diffusion states encode information about final answers, and (2) how inference-time acceleration affects this encoding. We focus on linear probe analyses that quantify the emergence and persistence of answer-relevant information across diffusion steps, providing an initial assessment of how caching influences the causal structure of diffusion-based reasoning.

Model Performance and Efficiency Table 1 summarizes task accuracy and inference latency for the baseline LLaDA-8B and the cached variant. As expected, dLLM-Cache provides a substantial speedup with a minor accuracy reduction ($39.7\% \rightarrow 36.8\%$). These results, consistent with prior findings (Liu et al., 2025b), frame our analysis of whether such acceleration preserves the causal relationship between intermediate reasoning states and final answers.

Classification Probes We evaluate whether intermediate diffusion states encode final-answer correctness via linear probes at each diffusion step (Figure 2). Accuracy above chance indicates that

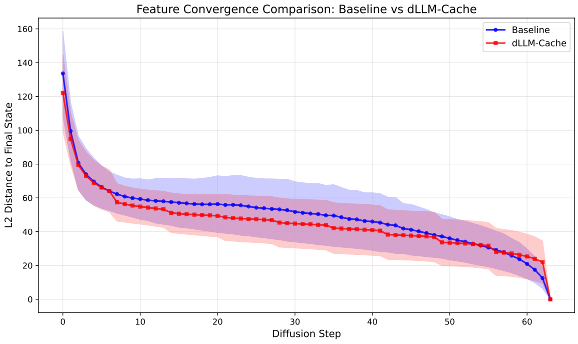

Figure 5: The L2 distance to the final predicted answer, showing the convergence of hidden states toward the final representation over diffusion steps.

hidden states contain information predictive of the final answer even before final answer convergence.

Classification accuracy rises steadily across diffusion steps, peaking at 83.5% for the baseline and 76.0% for dLLM-Cache at the final step. Averaged over all steps, caching reduces accuracy by about 1.6 points, indicating modest but consistent degradation in correctness encoding.

Even early diffusion steps (0–31) show above-chance classification accuracy: 63.9% for the baseline and 64.2% for dLLM-Cache. This indicates that "correctness" information emerges well before the final answer is decoded. Later steps accuracy improves to 73.0% (baseline) and 69.5% (cached), consistent with the accumulation of reasoning evidence during denoising. The smaller early–late gap under caching (5.3 vs. 9.1 points) suggests a compressed information timeline, with correctness signals distributed more uniformly across diffusion time.

Regression Probes We next test whether intermediate states encode numeric answer values using regression probes, summarized by R^2 and RMSE in Table 2. Both models exhibit negative R^2 scores across all steps, confirming that linear regression cannot recover numeric answers from hidden states. This is expected given the complexity of mapping high-dimensional representations to precise numeric values, and is consistent with findings in prior work (Paul et al., 2024).

The baseline R^2 peaks early (-16.7 at step 1) and declines thereafter (mean -45.3), indicating that numeric information becomes less linearly separable over time. dLLM-Cache yields slightly higher R^2 (mean -38.2) but remains strongly negative, suggesting that this minor improvement reflects task difficulty rather than greater faithfulness.

Impact of Acceleration on Faithfulness To quantify how caching affects the relationship between intermediate states and final answers, we compare probe performance between baseline and cached models. Figure 3 shows the difference in R² scores across steps. While both models show negative R² throughout, dLLM-Cache consistently achieves slightly better (less negative) R² scores, particularly in early steps. However, this pattern reverses for classification accuracy, where the baseline outperforms dLLM-Cache at nearly every step.

The divergence between classification and regression probe results suggests that caching affects different aspects of the hidden state representation. Classification probes, which measure whether correctness information is encoded, show degradation under caching. Regression probes, which measure numeric precision, show slight improvement, though both remain far from successful prediction. This indicates that caching may preserve coarse-grained answer information while altering fine-grained correctness signals.

Trajectory Analysis We further analyze the diffusion trajectories themselves to understand how reasoning unfolds. Figure 1 shows 2D PCA projections of hidden state trajectories, colored by final answer correctness. Correct and incorrect trajectories show some separation in the latent space, particularly in later diffusion steps, supporting the classification probe findings that correctness information accumulates over time.

Distance-to-final analysis (Figure **??**) reveals that hidden states converge toward the final representation as diffusion progresses, with convergence accelerating in later steps. This pattern is consistent across both models, suggesting that caching preserves the overall trajectory structure despite introducing some degradation in probe performance.

Feature Divergence Between Models To directly measure how caching alters hidden state representations, we compute L2 distance and cosine similarity between baseline and cached model states at each diffusion step (Figure 6). States diverge gradually over the diffusion process, with cosine similarity dropping from near 1.0 at step 0 to approximately 0.85-0.90 by step 63. This divergence is consistent with the probe performance differences observed, suggesting that caching introduces systematic changes to the hidden state space that affect faithfulness metrics.

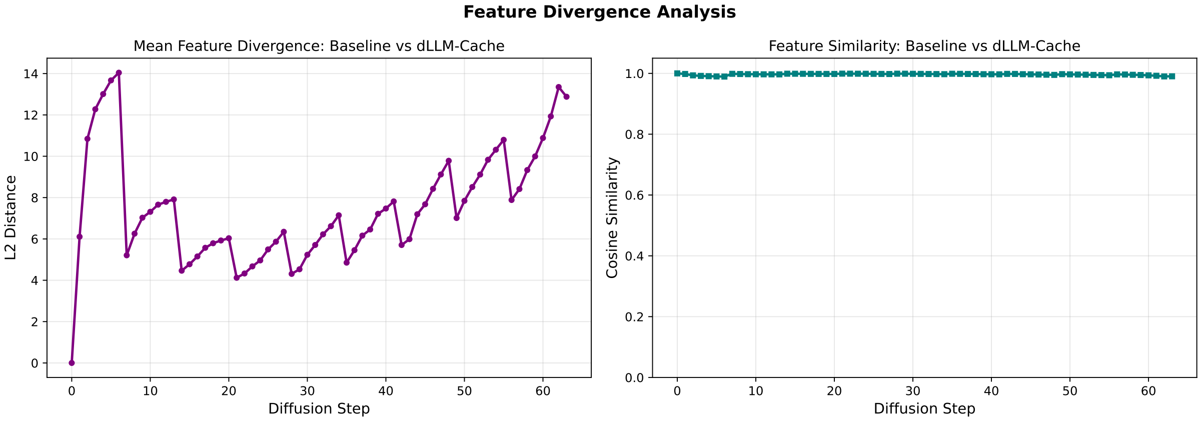

Figure 6: Hidden state divergence between baseline and cached models. Cosine similarity decreases over diffusion steps, indicating that caching introduces systematic changes to representations.

Speed-Faithfulness Tradeoff Figure 4 visualizes the efficiency-faithfulness tradeoff achieved by dLLM-Cache. The cached model achieves 1.78× speedup while maintaining 92.7% of baseline accuracy and 97.6% of baseline classification probe performance (mean accuracy). This suggests that caching provides a favorable tradeoff, preserving most of the faithfulness signal while substantially reducing inference time. The modest degradation in probe performance (1.6-7.5 percentage points) indicates that the causal relationship between intermediate states and final answers is largely preserved under acceleration.

Summary Our results reveal several key findings on DoT faithfulness and the impact of acceleration:
1. Early information emergence: Classification probes show above-chance accuracy in early steps, indicating that correctness information emerges early in the diffusion process.
2. Accumulation over time: Classification and trajectory analyses show that information accumulates during diffusion, peaking at the final step.
3. Caching preserves structure: Caching modestly reduces probe accuracy but preserves overall trajectory patterns and reasoning structure.
4. Regression limitations: Regression probes fail to predict numeric answers, indicating numeric values are not linearly encoded in hidden space.
5. Efficiency-faithfulness tradeoff: Caching provides a favorable efficiency–faithfulness tradeoff, maintaining most reasoning fidelity while accelerating inference, marking it a benign method for faster diffusion inference.

These findings suggest that DoTs encode meaningful reasoning structure and acceleration methods like caching preserve this structure with acceptable degradation. However, the linear probe methodology reveals limitations in numeric prediction, indicating that faithfulness may be better measured through correctness prediction rather than precise value estimation. Future evaluations should move beyond linear probes, such as our proposed counterfactual perturbations, to capture richer forms of causal faithfulness.

6 Conclusion

This work presented an empirical analysis of faithfulness in diffusion-based reasoning using LLaDA-8B and its accelerated variant, dLLM-Cache. Through trajectory-level probing, we examined how information about final answers emerges over diffusion steps and found that correctness signals appear early, strengthen over time, and remain largely preserved under caching despite minor degradation.

While our study provides an initial view of how acceleration affects reasoning dynamics, it is limited in scope. We focused on a single acceleration method and employed linear probes as a first-step diagnostic rather than a full causal analysis. Future iterations will incorporate counterfactual perturbation tests and additional acceleration techniques to broaden the evaluation and refine our understanding of causal faithfulness in diffusion models.

Future work will extend this framework in order to deepen empirical understanding of how trainfree acceleration methods influence the internal reasoning structure of diffusion language models. By expanding both methodological coverage and diagnostic rigor, we aim to establish a clearer picture of the trade-offs between computational efficiency and the preservation of faithful reasoning trajectories.

References

Anirud Aggarwal, Abhinav Shrivastava, and Matthew Gwilliam. 2025. Evolutionary caching to accelerate your off-the-shelf diffusion model. *Preprint*, arXiv:2506.15682.

Edoardo Cetin, Tianyu Zhao, and Yujin Tang. 2025. Large language models to diffusion finetuning. *arXiv preprint arXiv:2501.15781*.

Xinghao Chen, Anhao Zhao, Heming Xia, Xuan Lu, Hanlin Wang, Yanjun Chen, Wei Zhang, Jian Wang, Wenjie Li, and Xiaoyu Shen. 2025. Reasoning beyond language: A comprehensive survey on latent chain-of-thought reasoning. *arXiv preprint arXiv:2505.16782*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. In *Proceedings of the 38th International Conference on Machine Learning*.

Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, and 1 others. 2023. Dream-LLM: Synergistic multimodal comprehension and creation. *arXiv preprint arXiv:2309.11499*.

Ma et al. 2025. dKV-Cache: The cache for diffusion language models. *arXiv preprint*. Title and details as reported in secondary sources.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2022. GPTQ: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*.

Liqi He, Zuchao Li, Xiantao Cai, and Ping Wang. 2024. Multi-modal latent space learning for chain-of-thought reasoning in language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18180–18187.

Alec Helbling, Tuna Han Salih Meral, Ben Hoover, Pinar Yanardag, and Duen Horng Chau. 2025. Conceptattention: Diffusion transformers learn highly interpretable features. *Preprint*, arXiv:2502.04320.

Zemin Huang, Zhiyang Chen, Zijun Wang, Tiancheng Li, and Guo-Jun Qi. 2025. Reinforcing the diffusion chain of lateral thought with diffusion language models. *arXiv preprint arXiv:2505.10446*.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

to be completed Kang and 1 others. 2025. Ladir: Latent diffusion enhances LLMs for text reasoning. *arXiv preprint arXiv:2510.04573*.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, and 1 others. 2023. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*.

Pengxiang Li, Yefan Zhou, and 1 others. 2025. Diffusion language models know the answer before decoding. *arXiv preprint arXiv:2508.19982*.

Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori Hashimoto. 2022. Diffusion-LM improves controllable text generation. *Advances in Neural Information Processing Systems*.

Jiacheng Liu, Chang Zou, Yuanhuiyi Lyu, Junjie Chen, and Linfeng Zhang. 2025a. From reusing to forecasting: Accelerating diffusion models with taylorseers. *arXiv preprint arXiv:2503.06923*.

Zhiyuan Liu, Yicun Yang, Yaojie Zhang, Junjie Chen, Chang Zou, Qingyuan Wei, Shaobo Wang, and Linfeng Zhang. 2025b. dLLM-Cache: Accelerating diffusion large language models with adaptive caching. *arXiv preprint arXiv:2506.06295*.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. 2025. Large language diffusion models. *arXiv preprint arXiv:2502.09992*.

Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. 2024. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning. *Findings of the Association for Computational Linguistics: EMNLP 2024*.

to be completed Shao and 1 others. 2025. Diffuse thinking: Exploring diffusion language models as efficient thought proposers for reasoning. *arXiv preprint*. Please insert full author list and arXiv ID from the paper.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, pages 74952–74965. Curran Associates, Inc.

Felix Wimbauer, Bichen Wu, Edgar Schoenfeld, Xiaoliang Dai, Ji Hou, Zijian He, Artsiom Sanakoyeu, Peizhao Zhang, Sam Tsai, Jonas Kohler, and 1 others. 2023. Cache me if you can: Accelerating diffusion models through block caching. *arXiv preprint arXiv:2312.03209*.

Chen Xu and Dawei Yang. 2025. DLLMQuant: Quantizing diffusion-based large language models. *arXiv preprint arXiv:2508.14090*.

Zhewei Yao, Reza Yazdani Aminabadi, Minjia Zhang, Xiaoxia Wu, Conglong Li, and Yuxiong He. 2022. Zeroquant: Efficient and affordable post-training quantization for large-scale transformers. *arXiv preprint arXiv:2206.01861*.

Jiacheng Ye, Shansan Gong, Liheng Chen, Lin Zheng, Jiahui Gao, Han Shi, Chuan Wu, Xin Jiang, Zhenguo Li, Wei Bi, and Lingpeng Kong. 2024a. Diffusion of thoughts: Chain-of-thought reasoning in diffusion language models. *arXiv preprint arXiv:2402.07754*.

Jiacheng Ye, Shansan Gong, Liheng Chen, Lin Zheng, Jiahui Gao, Han Shi, Chuan Wu, Xin Jiang, Zhenguo Li, Wei Bi, and Lingpeng Kong. 2024b. Diffusion of thoughts: Chain-of-thought reasoning in diffusion language models. In *Advances in Neural Information Processing Systems*.