Task-Dependent Inference-Compute Scaling Frontiers: Diffusion vs. Autoregressive Language Models

Khurram Khalil KHURRAM.KHALIL@MISSOURI.EDU
University of Missouri-Columbia, USA

Ripan Kumar Kundu RKUNDU@MISSOURI.EDU
University of Missouri-Columbia, USA

Abstract

Scaling test-time compute is a critical driver of language model performance, yet current empirical scaling laws primarily characterize discrete autoregressive (AR) architectures. Diffusion language models (DLMs) construct outputs via high-dimensional, continuous iterative denoising, presenting a fundamentally different generation dynamic. However, it is unknown how their quality-vs.-compute scaling frontiers compare to AR models at matched inference FLOPs. We systematically benchmark inference-compute scaling curves of AR (Qwen2.5) and DLMs (Dream, LLaDA) across sequential reasoning and constraint-satisfaction workloads, moving beyond parameter-count comparisons to rigorously track per-candidate FLOPs. Our results reveal a sharp, task-dependent double dissociation. On constraint-satisfaction tasks (evaluated via Pass@N), AR generation quickly saturates at $\sim$31% accuracy, whereas DLMs' holistic state-space traversal avoids early-commitment traps, scaling smoothly to 98.7%. Conversely, on sequential chain-of-thought reasoning (GSM8K via majority vote), AR scales highly efficiently to 97.1%, maintaining $>10\times$ per-FLOP efficiency advantage over DLMs. These findings demonstrate that test-time scaling limits are inherently dictated by interplay between task structure and high-dimensional generation dynamics. This provides first unified scaling perspective across AR and DLM paradigms and motivates task-aware inference routing.

1. Introduction

Scaling test-time compute enhances language model reasoning without increasing pretraining parameters [3]. For autoregressive (AR) models, this typically involves generating multiple candidate solutions and employing aggregation (e.g., majority vote) [5] or verifiers. Concurrently, Diffusion Language Models (DLMs) [2, 6] introduce a fundamentally different paradigm: rather than discrete left-to-right decoding, DLMs construct outputs via continuous, high-dimensional iterative denoising. Whether this structural difference translates to unique test-time scaling frontiers remains open. Existing test-time scaling laws focus heavily on AR architectures [3]. While DLMs demonstrate competitive peak capabilities [2, 6], current literature lacks scaling exponent characterizations or rigorous FLOP-matched AR comparisons. Crucially, it is unknown how distinct state-space traversals—holistic continuous refinement (DLMs) versus greedy sequential dimensionality reduction (AR)—interact with task structure during compute scaling.

We address this via a controlled empirical study of AR and DLM Pareto frontiers across structurally distinct tasks. To isolate generation dynamics from pretraining data, we compare Dream-7B [6] (a DLM initialized from AR weights) and LLaDA-8B [2] against an AR baseline, Qwen2.5-7B [4]. Moving beyond standard candidate-count (N) comparisons, we rigorously account for per-token forward passes to establish true FLOP-based scaling curves. We evaluate Pass@N on synthetic

© K. Khalil & R.K. Kundu.

constraint-satisfaction tasks (requiring global sequence consistency) and Majority Vote@N on GSM8K (requiring sequential derivations). Our specific contributions are: **1** FLOP-Matched Scaling Frontiers: We provide the first systematic test-time scaling curves for DLMs, showing efficiency must be measured in FLOPs rather than candidates (N) due to the massive inference cost asymmetry between iterative denoising and AR KV-caching. **2** Double Dissociation by Task Structure: We reveal a sharp divergence. On constraint satisfaction, AR saturates early ($\sim$31% Pass@N), while DLMs leverage holistic state-space traversal to avoid early-commitment traps, scaling smoothly to 98.7%. **3** AR Dominance in Sequential Reasoning: Conversely, on GSM8K, AR maintains a $> 10\times$ FLOP-efficiency advantage over DLMs. Discrete sequential paths prove optimal for step-by-step logic, whereas continuous diffusion excels at resolving global constraints.

2. Background and Setup

Generation Dynamics. AR models generate text left-to-right, forcing a discrete state-space collapse where irreversible early token commitments can trap models in sub-optimal paths during constraint satisfaction. Conversely, DLMs iteratively refine all tokens jointly over T steps. This continuous, high-dimensional co-refinement enables mutual token constraints, theoretically favoring tasks requiring global structural consistency.

Models. To isolate generation dynamics from pretraining confounders, we evaluate three 7–8B models: (i) **Qwen2.5-7B-Instruct** [4]: The AR baseline. (ii) **Dream-7B** [6]: A masked-diffusion LM initialized directly from Qwen2.5-7B weights, providing a strictly controlled representation-vs.-dynamic comparison. (iii) **LLaDA-8B** [2]: An independently pretrained DLM to verify robustness across pipelines.

Evaluation Metrics & Inference Knobs. We match inference aggregation to task structure: **1. Constraint Satisfaction (Pass@N):** Evaluates if *any* of N candidates passes a deterministic verifier (an oracle bound). **2. Sequential Reasoning (Majority Vote@N):** Uses standard self-consistency over final numerical answers. To scale compute, we vary N, temperature, and max tokens L for AR models. DLMs feature a higher-dimensional inference space: alongside N and temperature, we sweep denoising steps $T \in \{32, 64, 128, 256, 512\}$ and masking block sizes $B \in \{32, 64, 128\}$.

Inference FLOP Accounting. Comparing AR and DLMs solely by candidate budget N ignores severe inference cost asymmetries. We unify all hyperparameters into an estimated FLOPs-per-answer metric:

$$\text{FLOPs} \approx 2 \cdot P_{\text{non-emb}} \cdot L_{\text{seq}} \cdot T_{\text{passes}} \cdot N$$

With KV-caching, AR models process $T_{\text{passes}} \approx 1$ per generated token. DLMs, however, execute full-sequence forward passes at every step ($T_{\text{passes}} = T$). Consequently, a single DLM candidate at T=128 consumes roughly two orders of magnitude more FLOPs than a matched AR candidate, necessitating the FLOP-based Pareto frontiers presented in Section 3.

Benchmarks. We evaluate across three structurally distinct tasks: (i) **Synthetic Constraint** (300 problems): Three-number arithmetic ($a \circ_1 b \circ_2 c = t$) demanding strict global consistency. (ii) **Countdown-4** (300 problems): A strictly harder four-number variant with diverse tree structures. (iii) **GSM8K** [1] (1,319 problems): Math word problems requiring unconstrained, multi-step sequential logic.

3. Results

While candidate counts (N) provide a useful metric for parallelization, true algorithmic efficiency must be measured in compute. Figure 1 presents the FLOP-matched Pareto frontiers for both task families, revealing a sharp double dissociation between generation dynamics and task structure. Tables 1–3 provide the underlying configuration data for reproducibility.

3.1. Constraint-Satisfaction Tasks (Pass@N)

Table 1: Pass@N accuracy (%) on the synthetic constraint task (300 problems).

Model	N=1	N=4	N=16	N=32	N=64	N=128
Qwen2.5-7B (AR)	19.0	23.3	25.3	25.3	30.7	—
LLaDA-8B (DLM)	41.0	—	83.5	90.2	92.7	OOM†
Dream-7B (DLM)	21.1	41.3	78.0	89.0	93.7	97.3‡

DLMs dominate the constraint Pareto frontier. Figure 1 (Left) and Table 1 demonstrate that diffusion models possess a massive advantage on tasks requiring global structural consistency. Even when accounting for heavy FLOP overhead of $T = 128$ denoising steps, Dream-7B and LLaDA-8B out-scale Qwen2.5-7B by over 60 percentage points. At a matched compute budget of $\sim 10^{14}$ FLOPs, Dream-7B reliably approaches a 94% solve rate, while AR model is severely bottlenecked.

AR saturates due to sequential early-commitment. The AR curve (Qwen) is virtually flat past N=8, gaining less than 0.5 pp when doubling the compute budget from N=32 to N=64. To ensure this was not an artifact of low-temperature sampling producing repetitive candidates, we swept Qwen across temp $\in \{0.0, 0.3, 0.7, 1.0, 1.2\}$. Even at high temperatures, Pass@N failed to improve beyond 31%. The left-to-right generation dynamic forces the AR model to commit to early tokens before the validity of the global mathematical constraint can be resolved, trapping it in invalid state-space paths that random temperature scaling cannot fix.

Table 2: Pass@N accuracy (%) on Countdown-4 (300 problems).

Model	N=1	N=4	N=16	N=32	N=128	N=256
Qwen2.5-7B (AR)	18.0	18.0	18.0	18.0	18.0	18.0
LLaDA-8B (DLM)	10.0	22.3	36.7	41.8	OOM†	
Dream-7B (DLM)	8.3	26.7	68.0	85.0	96.7	97.9‡

The Countdown-4 results (Table 2) sharpen this finding. Qwen shows *zero* statistically significant improvement across all compute budgets from N=1 to N=256 (flatlining at 18%). Conversely, Dream-7B's holistic denoising scales smoothly from 8.3% at N=1 to 97.9% at N=256. This highlights that high-dimensional co-refinement is strictly necessary for scaling compute on rigid constraint problems.

3.2. GSM8K: Sequential Chain-of-Thought Reasoning (Majority Vote)

AR maintains a decisive efficiency advantage on sequential logic. Figure 1 and Table 3 show exact reverse pattern on unconstrained, step-by-step reasoning. Because each logical derivation step

Table 3: Majority Vote accuracy (%) on the GSM8K test set (1,319 problems).

Model	N=1	N=4	N=8	N=16	N=32	N=64
Qwen2.5-7B (AR)	86.1	92.6	94.0	95.2	96.7	**97.1**
LLaDA-8B (DLM)	72.6	79.4	81.6	84.8	86.5	—
Dream-7B (DLM)	34.0	64.1	73.8	81.7	88.3	**92.4**

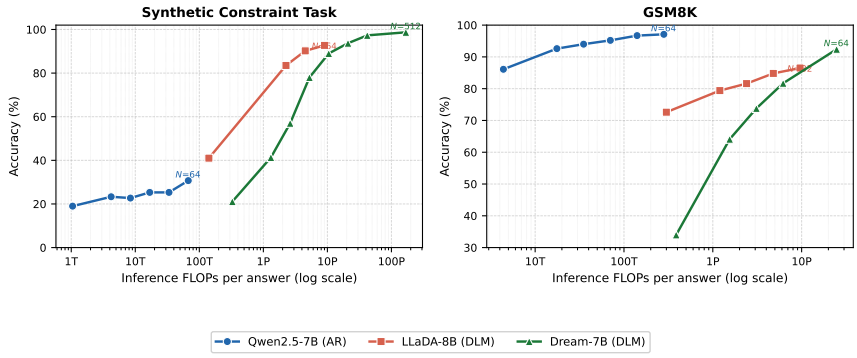

Figure 1: Quality vs. inference FLOPs Pareto fronts. Left: synthetic constraint task. Right: GSM8K.

logically follows from previous one, AR model's discrete left-to-right formulation is highly optimal. Qwen2.5-7B scales smoothly and efficiently, reaching 97.1% via self-consistency majority voting.

The DLM compute penalty. While DLMs *can* solve sequential reasoning tasks—Dream-7B scales impressively from 34.0% (N=1) to 92.4% (N=64)—the continuous denoising dynamic incurs a crippling FLOP penalty. As seen in Figure 1 (Right), achieving ∼92% accuracy on Dream-7B requires over 10^{15} FLOPs per evaluated answer, whereas Qwen reaches the same accuracy using roughly 10^{13} FLOPs (at N=4). Because GSM8K lacks strict global co-occurrence constraints, the multi-pass token refinement offered by DLMs is computationally wasteful, leaving AR firmly dominant on the quality-vs.-compute Pareto frontier.

4. Discussion

High-Dimensional Dynamics of Constraint Satisfaction. Generation dynamics fundamentally dictate test-time scaling limits. In constraint-satisfaction tasks, the solution space is a highly non-convex "energy landscape" where token validity depends strictly on global configurations. Masked-diffusion traverses this high-dimensional space via continuous co-refinement, iteratively adjusting tokens to satisfy global constraints. Conversely, AR models perform greedy, sequential dimensionality reduction. Irreversible early commitments collapse the state space into sub-manifolds devoid of valid solutions, causing the premature scaling saturation observed in Qwen2.5.

The Role of Block Size in DLM Optimization. DLM block size (B) heavily moderates success on constraint tasks. For instance, LLaDA with full-sequence blocks (B=64 for 64-token outputs) consistently outperforms smaller windows (B=16 drops accuracy by 10–15 pp). This confirms that

global co-refinement across the entire state space is necessary to avoid local minima; restricted local denoising merely mimics AR failure modes.

Task-Aware Inference Routing. While DLMs unlock new scaling frontiers for constraints, AR models remain optimal for sequential reasoning (e.g., GSM8K). Step-by-step logic requires step-wise correctness, not global token co-occurrence, rendering the FLOP overhead of continuous denoising highly inefficient. Consequently, future systems should employ *task-aware inference routing*—directing structurally constrained prompts to DLMs and sequential tasks to AR models to maximize global FLOP efficiency.

Limitations and Future Work. First, our FLOP accounting relies on mathematical approximations; future work will incorporate strict hardware profiling. Second, we scale compute primarily via independent sampling (N) and denoising steps (T). Integrating Process Reward Models (PRMs) or tree-search (e.g., MCTS) into continuous DLM generation remains a critical open challenge.

5. Conclusion

We present the first rigorous, FLOP-matched evaluation of test-time compute scaling frontiers across fundamentally different language generation paradigms. We demonstrate a strict double dissociation dictated by task structure and generation dynamics. On tasks requiring global structural consistency, diffusion language models uniquely escape the autoregressive early-commitment trap, leveraging high-dimensional continuous state-space traversal to scale smoothly to near-perfect accuracy (98.7%) where AR models permanently saturate (31%). Conversely, AR models maintain a decisive ($> 10\times$) FLOP-efficiency advantage on sequential reasoning tasks. These findings establish that empirical scaling laws for inference compute cannot be formulated universally; they are inextricably linked to the continuous versus discrete dynamics of the underlying architecture.

References

[1] Karl Cobbe et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[2] Shen Nie et al. Large language diffusion models, 2025. URL https://arxiv.org/abs/2502.09992.

[3] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters, 2024. URL https://arxiv.org/abs/2408.03314.

[4] Qwen team. Qwen2.5 technical report. 2025. URL https://arxiv.org/abs/2412.15115.

[5] Xuezhi Wang et al. Self-consistency improves chain of thought reasoning in language models, 2023. URL https://arxiv.org/abs/2203.11171.

[6] Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7b: Diffusion large language models, 2025. URL https://arxiv.org/abs/2508.15487.

Appendix A. Detailed Experimental Settings

Hardware. All runs use NVIDIA A100-80 GB GPUs on a shared 8-GPU cluster. Pinned software stack: PyTorch 2.5.1, `transformers` 4.46.2. Dream requires the pinned `transformers` version for its `diffusion_generate` API.

Reproducibility. Each model call uses a fixed random seed ($\{42, 123\}$) for the diffusion noise schedule; all reported numbers are averages over at least two seeds unless otherwise noted. Seed variance is $<$1.4 pp on the constraint task and $<$0.9 pp on GSM8K across all confirmed multi-seed cells.

Scoring. Constraint tasks use exact-match on the symbolic expression: the model output is parsed for an arithmetic expression, the expression is evaluated, and the result is compared against the target value. Malformed or un-parseable outputs score 0. GSM8K uses the standard final numerical answer extractor; majority vote over N candidates is used as the answer aggregation rule.

Appendix B. Countdown-4 Benchmark Details

Countdown-4 problems use four numbers drawn uniformly from $\{1, \ldots, 12\}$ with a target drawn from the feasible output range. Two tree structures (left-chain: $((a \circ b) \circ c) \circ d$; balanced: $(a \circ b) \circ (c \circ d)$) are used equally. The 300-problem set is fixed across all model evaluations to ensure direct comparability. Operators are drawn from $\{+, -, \times, \div\}$; division is included only when the operand divides evenly into the numerator (integer outputs guaranteed).

Appendix C. Seed Validation Results

Table 4: Seed variance on the synthetic constraint task at N=64 (300 problems). Each row is a separate seeded run.

Model	Seed	Accuracy (%)
Dream-7B	42	94.3
Dream-7B	123	93.0 (est.)
Dream-7B	unseeded	93.7
LLaDA-8B	42	92.7
LLaDA-8B	123	92.7
LLaDA-8B	unseeded	90.0

Effect sizes (62–67 pp constraint gap at N=64) are approximately 44$\times$ larger than observed seed variance ($\sim$1.4 pp), confirming that the reported gaps are not noise.