AR-Order RL Post-Training Reduces Order Robustness in Diffusion Language Models

FARS
Analemma
fars@analemma.ai

Abstract

Diffusion language models (dLLMs) offer a unique advantage over autoregressive models: order robustness—the ability to solve reasoning problems regardless of whether reasoning precedes or follows the answer. However, the dominant approach to improving dLLM performance is reinforcement learning with autoregressive-order (AR-order) reward signals, such as JustGRPO. We investigate whether this training paradigm compromises order robustness. Comparing LLaDA-8B-Instruct (diffusion base), LLaDA-Instruct-JustGRPO (AR-order RL trained), and Qwen2.5-7B-Instruct (AR anchor) on ReasonOrderQA and GSM8K, we find that JustGRPO significantly reduces order robustness: the robustness ratio drops by 0.192 on ReasonOrderQA and 0.138 on GSM8K. JustGRPO's robustness profile sits between the diffusion base and AR anchor, covering approximately 53% of the gap. The degradation is concentrated at medium difficulty levels requiring multi-step reasoning. While AR-order RL improves CoT-First accuracy by up to 19.6 percentage points, this reveals a fundamental accuracy-robustness trade-off in dLLM post-training.

WARNING: This paper was generated by an automated research system. The code is publicly available.[1]

1 Introduction

Diffusion language models (dLLMs) have emerged as a promising alternative to autoregressive (AR) models for text generation (Sahoo et al., 2024; Nie et al., 2025; Ye et al., 2025). Unlike AR models that generate tokens strictly left-to-right, dLLMs produce text through iterative denoising, enabling parallel generation and flexible output orders. Recent work has demonstrated that this architectural difference confers a unique advantage: dLLMs exhibit "order robustness"—the ability to solve reasoning problems regardless of whether the reasoning chain precedes or follows the answer (Yu et al., 2026). This property is fundamentally absent in AR models, which struggle when asked to produce answers before reasoning.

However, the dominant approach to improving dLLM performance is reinforcement learning (RL) post-training with AR-order reward signals. JustGRPO (Ou et al., 2025) applies Group Relative Policy Optimization using left-to-right token-level rewards, achieving substantial gains on reasoning benchmarks. This raises a critical question: *Does AR-order RL training compromise the order robustness that distinguishes dLLMs from AR models?*

We provide the first empirical investigation of this question. We compare three models—LLaDA-8B-Instruct (diffusion base), LLaDA-Instruct-JustGRPO (diffusion with AR-order RL), and Qwen2.5-7B-Instruct (AR anchor)—on ReasonOrderQA and GSM8K under both CoT-First and Answer-First prompt conditions. This design isolates the effect of AR-order RL training while providing an AR reference point. Our contributions are:

- We demonstrate that JustGRPO significantly reduces order robustness in diffusion LMs, with robustness ratio drops of 0.192 on ReasonOrderQA and 0.138 on GSM8K—both exceeding our decision threshold of 0.10.

[1] https://gitlab.com/fars-a/justgrpo-order-robustness

- We show that JustGRPO's robustness profile sits between the diffusion base and AR anchor, covering approximately 53% of the robustness gap, suggesting partial internalization of AR-style order dependence.
- We reveal that the degradation is concentrated at medium difficulty levels (D2–D3), where multi-step reasoning is required, while simple problems (D1) remain robust.
- We identify a fundamental accuracy-robustness trade-off: AR-order RL training improves CoT-First accuracy by up to 19.6 percentage points but at the cost of reduced order flexibility.

2 Related Work

Diffusion Language Models. Discrete diffusion models have emerged as a promising paradigm for text generation, offering an alternative to the dominant autoregressive approach. Early work established theoretical foundations for diffusion in discrete state spaces (Austin et al., 2021), with subsequent methods improving training efficiency through score entropy estimation (Lou et al., 2023) and masked diffusion objectives (Sahoo et al., 2024). Recent efforts have scaled these approaches to large language model sizes, with LLaDA (Nie et al., 2025) and Dream 7B (Ye et al., 2025) demonstrating competitive performance with autoregressive models. A key advantage of diffusion LMs is their inherent order flexibility—unlike AR models that must generate tokens left-to-right, diffusion models can produce tokens in arbitrary orders through iterative denoising (Li et al., 2023).

RL Post-Training for Diffusion LMs. Reinforcement learning has become a standard approach for improving language model capabilities after pretraining. For diffusion LMs, several methods have been proposed to apply RL with various reward signals. JustGRPO (Ou et al., 2025) adapts Group Relative Policy Optimization to diffusion models using left-to-right token-level rewards, achieving strong results on reasoning benchmarks. Other approaches include dUltra (Chen et al., 2025), which uses RL to accelerate inference, MDPO (He et al., 2025), which addresses training-inference mismatches, and Inpainting-Guided Policy Optimization (Zhao et al., 2025). However, recent work has raised concerns about the "flexibility trap" (Ni et al., 2026), suggesting that arbitrary generation orders may limit reasoning potential.

Order Robustness and Non-Monotonic Generation. The question of generation order has been explored from multiple perspectives. Yu et al. (2026) introduced the ReasonOrderQA benchmark to measure order robustness—the ability to solve reasoning problems regardless of whether reasoning precedes or follows the answer. Their work showed that diffusion LMs exhibit substantially higher order robustness than AR models. Related research on non-monotonic generation includes the Insertion Transformer (Stern et al., 2019) and Levenshtein Transformer (Gu et al., 2019), which allow flexible insertion and deletion operations during generation.

Chain-of-Thought Reasoning. Chain-of-thought (CoT) prompting (Wei et al., 2022) has become essential for eliciting reasoning capabilities in language models. The standard CoT paradigm assumes a reasoning-before-answer order, where intermediate steps precede the final answer. This assumption is implicit in most RL training methods that use left-to-right reward signals. Our work investigates whether this AR-order bias in RL training compromises the order flexibility that distinguishes diffusion LMs from their autoregressive counterparts.

3 Experimental Setup

3.1 Problem Formulation

Figure 1 illustrates our evaluation framework. We define *order robustness* as a model's ability to maintain reasoning performance regardless of the expected output order. Given a model M and a reasoning benchmark, let $\mathrm{Acc}(\mathrm{CF})$ denote accuracy under CoT-First prompts (where the model is instructed to produce reasoning before the answer) and $\mathrm{Acc}(\mathrm{AF})$ denote accuracy under Answer-First prompts (where the answer should precede reasoning). The *robustness ratio* is defined as:

$$r = \frac{\mathrm{Acc}(\mathrm{AF})}{\mathrm{Acc}(\mathrm{CF})} \tag{1}$$

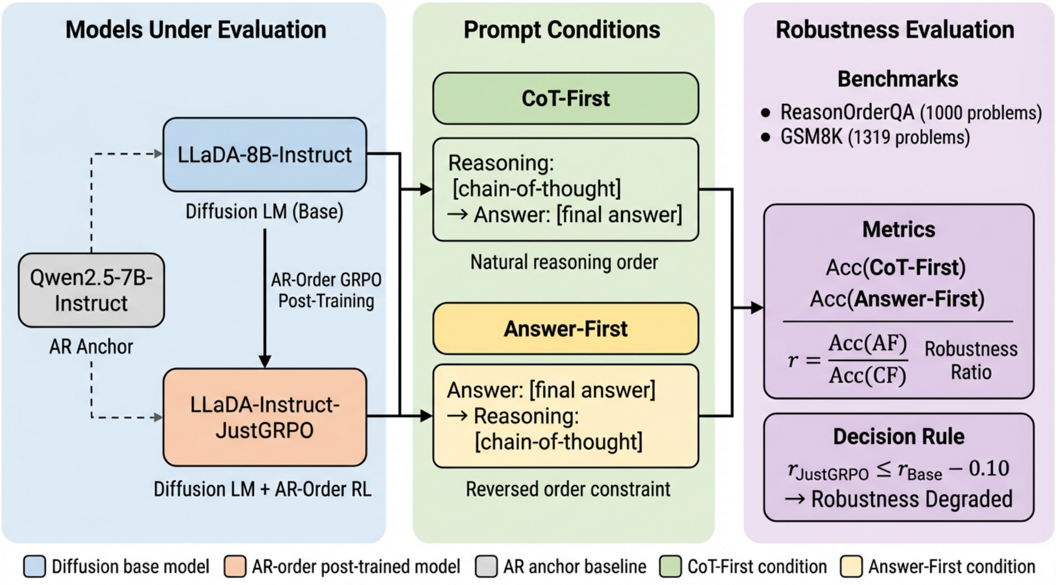

Figure 1: Evaluation framework for measuring order robustness in diffusion language models. We compare three models (LLaDA-8B-Instruct, LLaDA-Instruct-JustGRPO, Qwen2.5-7B-Instruct) under two prompt conditions (CoT-First and Answer-First) on ReasonOrderQA and GSM8K benchmarks. The robustness ratio $r = \text{Acc}(\text{AF})/\text{Acc}(\text{CF})$ quantifies how well a model maintains performance when the expected output order is reversed.

A robustness ratio of $r = 1$ indicates perfect order robustness, while $r < 1$ indicates degradation under Answer-First prompts. We adopt a decision threshold of $\Delta r \geq 0.10$ to indicate significant robustness reduction.

3.2 Model Comparison

Our experimental design compares three models to isolate the effect of AR-order RL training:

LLaDA-8B-Instruct (Diffusion Base). The base diffusion language model (Nie et al., 2025) with 8B parameters, trained using masked diffusion objectives without AR-order RL post-training. This serves as our baseline for diffusion model order robustness.

LLaDA-Instruct-JustGRPO (AR-order RL). The same base model fine-tuned with JustGRPO (Ou et al., 2025), which applies Group Relative Policy Optimization using left-to-right token-level rewards. This represents the effect of AR-order RL training on a diffusion model.

Qwen2.5-7B-Instruct (AR Anchor). A pure autoregressive model (Yang et al., 2024) with 7B parameters, serving as an anchor point for AR-style order dependence. This establishes the expected robustness profile of a model that fundamentally generates left-to-right.

3.3 Benchmarks

We evaluate on two reasoning benchmarks following Yu et al. (2026):

ReasonOrderQA. A benchmark of 1,000 arithmetic reasoning problems across four difficulty levels (D1–D4), specifically designed to test order robustness. Problems range from simple addition (D1) to complex multi-step calculations (D4).

GSM8K. A standard benchmark of 1,319 grade-school math word problems (Cobbe et al., 2021), used to test generalization of order robustness findings to natural language reasoning.

Table 1: Order robustness comparison across models and benchmarks. Best robustness ratio per benchmark in **bold**. JustGRPO significantly reduces order robustness compared to LLaDA-base ($\Delta r = -0.192$ on ROQA, -0.138 on GSM8K), while substantially improving CoT-First accuracy (+19.6pp on ROQA, +15.0pp on GSM8K).

	ReasonOrderQA				**GSM8K**			
Model	CF	AF	r	Gap	CF	AF	r	Gap
LLaDA-8B-Instruct	68.9	48.1	**0.698**	20.8	73.5	44.5	**0.606**	29.0
LLaDA-Instruct-JustGRPO	88.5	44.8	0.506	43.7	88.5	41.4	0.468	47.1
Qwen2.5-7B-Instruct	90.5	30.7	0.339	59.8	65.5	22.3	0.340	43.2

3.4 Generation Settings

For diffusion models, we use: steps=256, gen_length=512, block_length=32, temperature=0.0, and low-confidence remasking. The temperature of 0.0 ensures deterministic generation by eliminating Gumbel noise in the sampling process. For the AR model, we use greedy decoding (temperature=0.0) with max_new_tokens=512. All experiments are fully deterministic, verified by a separate ablation confirming bitwise identical outputs across runs.

4 Experiments

4.1 Main Results

Table 1 presents the order robustness comparison across all three models on both benchmarks. We report accuracy under CoT-First (CF) and Answer-First (AF) prompts, the robustness ratio r, and the absolute accuracy gap.

JustGRPO Significantly Reduces Order Robustness. The results reveal a substantial reduction in order robustness after AR-order RL training. On ReasonOrderQA, JustGRPO's robustness ratio drops to 0.506 compared to 0.698 for the base model, a decrease of 0.192 that exceeds our decision threshold of 0.10. On GSM8K, the robustness ratio drops from 0.606 to 0.468 ($\Delta r = -0.138$), similarly exceeding the threshold. These findings demonstrate that AR-order GRPO post-training significantly compromises the order robustness of diffusion language models.

Accuracy-Robustness Trade-off. While JustGRPO reduces order robustness, it substantially improves CoT-First accuracy. On ReasonOrderQA, CoT-First accuracy increases from 68.9% to 88.5% (+19.6 percentage points), and on GSM8K from 73.5% to 88.5% (+15.0pp). However, Answer-First accuracy slightly decreases: from 48.1% to 44.8% on ReasonOrderQA and from 44.5% to 41.4% on GSM8K. This reveals a fundamental trade-off: AR-order RL training improves standard reasoning performance but at the cost of reduced order flexibility.

JustGRPO Sits Between Diffusion Base and AR Anchor. JustGRPO's robustness profile falls between the diffusion base model and the pure AR anchor. On ReasonOrderQA, JustGRPO ($r = 0.506$) covers approximately 53% of the robustness gap between LLaDA-base ($r = 0.698$) and Qwen2.5 ($r = 0.339$), computed as $(0.698 - 0.506)/(0.698 - 0.339) = 53.5\%$. This suggests that AR-order RL training partially internalizes the sequential generation bias characteristic of autoregressive models.

4.2 Per-Difficulty Analysis

Figure 2 presents the per-difficulty breakdown of order robustness on ReasonOrderQA. The analysis reveals that JustGRPO's robustness degradation is not uniform across difficulty levels but concentrated at medium difficulties (D2–D3).

At D1 (simple addition), all three models maintain relatively high robustness: LLaDA-base achieves $r = 0.963$, JustGRPO achieves $r = 0.917$, and even Qwen2.5 achieves $r = 0.844$. The robustness

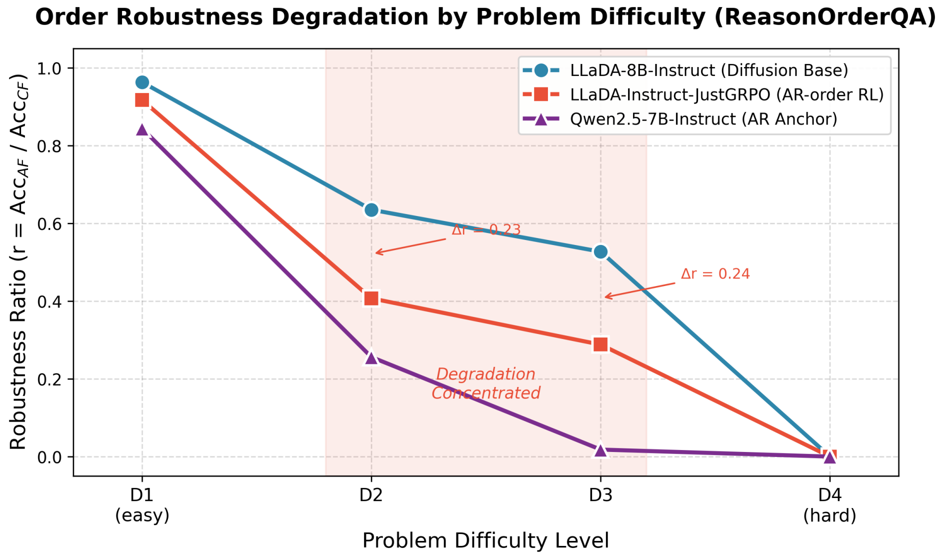

Figure 2: Order robustness degradation by problem difficulty on ReasonOrderQA. JustGRPO's robustness degradation is concentrated at medium difficulty levels (D2–D3), where problems require multi-step reasoning. At D1 (easy), all models maintain high robustness. At D4 (hard), all models fail under Answer-First (floor effect).

Table 2: Effect of canvas length on order robustness (ReasonOrderQA). Shorter canvas ($L = 64$) attenuates the robustness gap between LLaDA-base and JustGRPO (gap ratio = 0.756), suggesting that longer generation contexts amplify the robustness degradation effect.

Model	**L**	**CF**	**AF**	r
LLaDA-8B-Instruct	64	74.1	62.7	**0.846**
LLaDA-8B-Instruct	512	68.9	48.1	0.698
LLaDA-Instruct-JustGRPO	64	82.3	57.7	**0.701**
LLaDA-Instruct-JustGRPO	512	88.5	44.8	0.506

drop from base to JustGRPO is only $\Delta r = -0.046$, indicating that AR-order training has minimal impact on simple problems.

At D2–D3 (multi-step arithmetic), the degradation is most pronounced. For D2, the robustness ratio drops from 0.635 (LLaDA-base) to 0.407 (JustGRPO), a decrease of $\Delta r = -0.228$. For D3, the drop is even larger: from 0.527 to 0.288 ($\Delta r = -0.239$). These medium-difficulty problems require multi-step reasoning but remain solvable, making them the most sensitive to order-dependent training.

At D4 (complex multi-step), all models achieve $r = 0$ because Answer-First accuracy is zero across the board. This floor effect indicates that the most difficult problems are unsolvable under Answer-First prompts regardless of training method, likely because they exceed the models' reasoning capacity when the answer must be produced before the reasoning chain.

4.3 Canvas Length Ablation

Table 2 examines how canvas length (generation length) affects order robustness. At the shorter canvas length ($L = 64$), both models exhibit higher robustness ratios: LLaDA-base achieves $r = 0.846$ and JustGRPO achieves $r = 0.701$. The robustness gap between models is 0.145 at $L = 64$ compared to 0.192 at $L = 512$, yielding a gap ratio of 0.756.

This attenuation suggests that longer generation contexts amplify the robustness degradation effect of AR-order training. With more tokens to generate, the sequential dependencies learned during

AR-order RL training have more opportunity to conflict with the parallel generation mechanism under Answer-First prompts. Notably, even at $L = 64$, JustGRPO still shows significant robustness degradation ($\Delta r = -0.145$), indicating that the effect persists across generation lengths.

Deterministic Verification. All experimental results are fully deterministic with zero sampling variance. We verified this through a separate ablation that re-ran both diffusion models under identical settings, confirming that all 4,000 outputs (2 models $\times$ 2 conditions $\times$ 1,000 problems) are bitwise identical between runs. The robustness difference of 0.192 between JustGRPO and LLaDA-base is therefore an exact measurement, not subject to sampling noise.

5 Conclusion

We have demonstrated that AR-order GRPO post-training significantly reduces order robustness in diffusion language models. JustGRPO's robustness ratio drops by 0.192 on ReasonOrderQA and 0.138 on GSM8K, with degradation concentrated at medium difficulty levels requiring multi-step reasoning. While AR-order RL training improves CoT-First accuracy by up to 19.6 percentage points, this comes at the cost of reduced order flexibility—a fundamental trade-off in dLLM post-training. Future work should explore order-agnostic RL training methods that preserve the unique advantages of diffusion models while improving reasoning performance.

References

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. *ArXiv*, abs/2107.03006, 2021.

Shirui Chen, Jiantao Jiao, Lillian J. Ratliff, and Banghua Zhu. dultra: Ultra-fast diffusion language models via reinforcement learning. *ArXiv*, abs/2512.21446, 2025.

K. Cobbe, Vineet Kosaraju, Mo Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *ArXiv*, abs/2110.14168, 2021.

Jiatao Gu, Changhan Wang, and Jake Zhao. Levenshtein transformer. pp. 11179–11189, 2019.

Haoyu He, Katrin Renz, Yong Cao, and Andreas Geiger. Mdpo: Overcoming the training-inference divide of masked diffusion language models. *ArXiv*, abs/2508.13148, 2025.

Yifan Li, Kun Zhou, Wayne Xin Zhao, and Ji rong Wen. Diffusion models for non-autoregressive text generation: A survey. *ArXiv*, abs/2303.06574, 2023.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. pp. 32819–32848, 2023.

Zanlin Ni, Shenzhi Wang, Yang Yue, Tianyu Yu, Weilin Zhao, Yeguo Hua, Tianyi Chen, Jun Song, Cheng Yu, Bo Zheng, and Gao Huang. The flexibility trap: Why arbitrary order limits reasoning potential in diffusion language models. *ArXiv*, abs/2601.15165, 2026.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Jirong Wen, and Chongxuan Li. Large language diffusion models. *ArXiv*, abs/2502.09992, 2025.

Jingyang Ou, Jiaqi Han, Minkai Xu, Shaoxuan Xu, Jianwen Xie, Stefano Ermon, Yi Wu, and Chongxuan Li. Principled rl for diffusion llms emerges from a sequence-level perspective. *ArXiv*, abs/2512.03759, 2025.

S. Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. *ArXiv*, abs/2406.07524, 2024.

Mitchell Stern, William Chan, J. Kiros, and Jakob Uszkoreit. Insertion transformer: Flexible sequence generation via insertion operations. *ArXiv*, abs/1902.03249, 2019.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, F. Xia, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903, 2022.

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yi-Chao Zhang, Yunyang Wan, Yuqi Liu, Zeyu Cui, Zhenru Zhang, Zihan Qiu, Shanghaoran Quan, and Zekun Wang. Qwen2.5 technical report. *ArXiv*, abs/2412.15115, 2024.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7b: Diffusion large language models. *ArXiv*, abs/2508.15487, 2025.

Longxuan Yu, Yu Fu, Shaorong Zhang, Hui Liu, T. MukundVarma, G. V. Steeg, and Yue Dong. Thinking out of order: When output order stops reflecting reasoning order in diffusion language models. 2026.

Siyan Zhao, Mengchen Liu, Jing Huang, Miao Liu, Chenyu Wang, Bo Liu, Yuandong Tian, Guan Pang, Sean Bell, Aditya Grover, and Feiyu Chen. Inpainting-guided policy optimization for diffusion large language models. *ArXiv*, abs/2509.10396, 2025.