# LLaDA-R1: Scaling Reasoning at Inference Time with Diffusion-LLMs

| Radostin Cholakov  radicho@stanford.edu | Zeyneb N. Kaya  zeynebnk@stanford.edu | Joe Li  joeli@stanford.edu | Nicole Ma  manicole@stanford.edu |
| --- | --- | --- | --- |

# 

# Abstract

We present LLaDA-R1 (LLaDA Reasoner1), the first text diffusion model optimized for reasoning and computational efficiency at inference time. We present three key innovations: (1) a remasking optimization head that learns optimal token positions to remask at each step, (2) reinforcement learning with a novel reward function balancing reasoning accuracy and computational efficiency, and (3) an adaptive diffusion method post-trained with supervised fine-tuning to dynamically adjust diffusion steps. We challenge the assumption that autoregressive architectures are inherently superior for reasoning tasks and highlight the untapped potential of properly optimized diffusion-based approaches for complex problem-solving.

# 1. Introduction

Autoregressive models (ARMs) have dominated the landscape of language modeling in recent years. These models generate text one token at a time, with each new token conditioned on all previously generated tokens. This approach has proven remarkably effective, leading to the development of powerful models like GPT-4 and LLaMA (Minaee et al., 2024).

Recently, Nie et al. (2024) challenged this paradigm and introduced LLaDA (Large Language Diffusion with mAsking), a novel masked diffusion model that incorporates a discrete random masking process and trains a mask predictor to approximate its reverse process. Diffusion models, which have achieved remarkable success in image generation, represent a promising alternative paradigm. Unlike ARMs, diffusion models operate by gradually denoising data through an iterative process. While extensively studied in continuous domains like images, their application to discrete textual data presents unique challenges and opportunities. Unlike autoregressive approaches, diffusion-based methods can generate data in a non-sequential manner, potentially enhancing long-term planning capabilities, overcoming premise ordering limitations, offering better control over the generation process, and improving sampling efficiency (Sahoo et al., 2024).

**1.1 Remasking**

The masking-based diffusion architecture of LLaDA enables it to construct a probabilistic model with bidirectional dependencies while optimizing a variational lower bound on the log-likelihood. LLaDA outperforms several ARM baselines and excels at in-context learning and instruction-following abilities after SFT.

Despite these promising results, LLaDA’s reliance on specific remasking strategies reveals potential limitations. While Nie et al. explore deterministic remasking strategies similar to annealing techniques used in traditional LLMs, their evaluation is limited to only three configurations: random remasking, lowest confidence remasking, and a combination of lowest confidence with semi-autoregressive remasking. In particular, the limitations of LLaDA’s remasking strategies are emphasized by the papers’ insights that often, during the remasking processes, correctly predicted reasoning processes are masked out again, often regressing progress and masking important information due to ineffective masking strategies.

The choice of the remasking strategy significantly impacts model performance, yet only three methods were evaluated in the original work. For the LLaDA 8B Instruct model, directly applying the lowest confidence remasking strategy leads to a substantial performance drop, highlighting the sensitivity of diffusion-based language models to inference-time masking techniques.

**1.2 Diffusion and Denoising Steps**

The foundational work on Denoising Diffusion Probabilistic Models (DDPMs) introduced a forward process that gradually adds noise to the data and a reverse process that learns to remove noise step-by-step (Ho et al., 2020). Although DDPMs yield impressive results, one major drawback is their computational cost. The original sampling procedure requires a large number of reverse diffusion steps—on the order of hundreds to even thousands—to generate high-quality samples. This inefficiency has motivated a line of research focused on minimizing diffusion steps while maintaining or improving generation quality.

A significant advance toward efficient sampling is a non-Markovian diffusion process that shortens the reverse diffusion chain (Song et al., 2020). By relaxing the strict Markov assumption, DDIM enables a deterministic sampling path that effectively reduces inference time without heavily compromising sample fidelity. Building on these concepts, variance schedules and network architectures can be further refined to accelerate sampling while also improving the likelihood of generated samples (Nichol et al., 2021).

Another prominent strategy is progressive distillation, which transfers knowledge from a large DDPM into a smaller, more efficient model capable of generating samples in fewer steps (Salimans et al., 2022). The key idea is to iteratively train the shallower network to mimic the output distribution of the deeper diffusion chain, progressively reducing the chain length. Through this repeated distillation process, the model learns to achieve comparable sample quality with a fraction of the original steps. In a related vein, DPM-Solver addresses the sampling challenge by reformulating the reverse diffusion process as an Ordinary Differential Equation (ODE) and applying specialized numerical methods for rapid solution (Lu et al., 2022). By leveraging higher-order solvers, DPM-Solver attains high-quality generation in as few as ten steps.

Recent research has also explored content-aware approaches that dynamically adjust the number of diffusion steps based on generation complexity. Yu and Barad (2024) demonstrated that different text prompts require varying numbers of denoising steps to achieve optimal quality, developing an NLP model that predicts the minimum necessary steps for each prompt. Their approach showed that over 70% of prompts required significantly fewer steps than the standard default, resulting in substantial computational savings while maintaining or improving output quality.

# 2. Methodology

In this study, we introduce LLaDA-R1 (LLaDA Reasoner1), an optimized approach to efficiently scale LLaDA for reasoning through multiple key methods: (a) remasking optimization, (b) reinforcement learning, and (c) adaptive diffusion.

**2.1 Remasking Optimization**

The limitations of LLaDA’s remasking strategies and its influence on the performance highlight it as a key aspect to improving its reasoning capabilities. In LLaDA-R1, we modify the LLaDA architecture with an additional linear head to predict the best output positions to remask at each diffusion step. The intuition behind this additional parameter is to learn ideal positions to remask at a given step to yield the best performance while requiring the least number of diffusion steps. It reduces the remasking of important information that reverse progress. Its weights are randomly initialized, aligning with the original random remasking approach in LLaDA, and are updated in the subsequent reinforcement learning stage. During inference time, the additional output logits from this head are passed through the softmax operation to compute the optimal masks to remask for the step.

**2.2 Reinforcement Learning**

While the original LLaDA shows substantial gains through supervised fine-tuning for instructions, reinforcement learning has not been applied to any text diffusion model before. We present a reinforcement learning approach that optimizes both for reasoning ability and computational efficiency at the same time. We work with the training set of the GSM8K Benchmark. We develop a reward function that combines two objectives:

1. A standard accuracy reward that determines whether the model output is correct
2. A penalty component for the number of diffusion steps taken.

The penalty encourages the model to reach the best performance while optimizing the number of diffusion steps required to reach it. Rather than applying a fixed reduction in steps or distilling knowledge from a larger model, our reinforcement learning-based framework enables adaptive optimization that responds to both the content characteristics and computational constraints, and we combine these two objectives with Group Relative Policy Optimization (GRPO). By incorporating step count directly into the reward signal, LLaDA-R1 learns to dynamically optimize sampling steps while maintaining generation accuracy and quality, effectively addressing the need for highly-efficient test-time scaling.

**2.3 Adaptive Diffusion**

To further optimize the computational efficiency and number of diffusion steps used, we develop an adaptive diffusion method in LLaDA-R1 that overcomes the fixed diffusion steps applied in LLaDA. Our approach leverages supervised fine-tuning to train an additional head in the diffusion model. We create synthetic labels by running the denoising process on 50 calibration examples from the train set of GSM8K for 128 steps and save the hidden states from LLaDA on each step. We use a verifier to check if the correct answer for the reasoning task is contained in the response from that step and record that as a positive label. Otherwise, we record a negative label. Then we train a neural network which takes a hidden state as an input and optimizes a binary classification objective. During inference, LLaDA uses the additional head on each step to determine if it should continue denoising or terminate denoising on the next step.

# 3. Performance

We evaluate performance on the GSM8K Benchmark with LLaDa-R1 8B and compare our results to baseline performance with the original LLaDA-Instruction 8B.

|  | Avg. Diffusion Steps | Max. Diffusion Steps | Accuracy |
| --- | --- | --- | --- |
| LLaDa | 128 | 128 | 0.68 |
| LLaDa-R1 | **118** | 256 | **0.72** |

*Table 1. Performance on a subset of the GSM8K benchmark test set.*

Through our adaptive diffusion technique, we demonstrate that for harder tasks, we can spend longer time reasoning (hence higher maximum diffusion steps than LLaDa), but for easier tasks, we use 8% fewer diffusion steps. Moreover, >50% of diffusion steps could be skipped depending on the input task with marginal impacts on final performance. This improves inference-time efficiency substantially. Furthermore, we observed that our methods often identified the best time to stop at precisely when the model first obtained the correct final answer. Our approaches present effective methods for dynamically adjusting diffusion steps and allowing efficient inference time reasoning.

We observe that our optimized reasoning methodology was not only able to outperform the baseline model of equal size, but also do so in a manner that did so with greater efficiency. Our combination of optimized and novel methodologies takes the greatest advantage of the strengths and potential of text diffusion models, and shows their potential in complex reasoning tasks.

# 4. Conclusion and Discussion

We introduce LLaDA-R1, the first text diffusion model optimized dually for reasoning and computational inference efficiency. Our approach incorporates three key innovations: a remasking optimization head, reinforcement learning with a step-penalty reward, and adaptive diffusion through confidence convergence stopping. Results on the GSM8K benchmark demonstrate that LLaDA-R1 outperforms both the original LLaDA and comparable autoregressive models while reducing diffusion steps by up to ~50% with minimal performance impact.

These findings challenge the assumption that autoregressive models are inherently superior for reasoning tasks and highlight the unique advantages of diffusion-based approaches when properly optimized. The non-sequential generation capabilities of diffusion models prove beneficial for complex problems requiring long-term planning and logical deduction. Future work should explore scaling to larger models, extending these techniques to other reasoning domains, and further refining the balance between accuracy and efficiency.

# Acknowledgements

Thank you to Etched, Mercor, and Cognition AI for hosting the 2025 inference-time compute hackathon, as well as Coreweave and Anthropic for providing us with compute resources, to develop these methods and the environment to explore new frontiers of AI.

# 

# References

Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., & Zhu, J. (2022). DPM-Solver: A Fast ODE Solver for Diffusion Probabilistic Model Sampling in Around 10 Steps. *ArXiv, abs/2206.00927*.

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M.A., Socher, R., Amatriain, X., & Gao, J. (2024). Large Language Models: A Survey. *ArXiv, abs/2402.06196.*

Nichol, A., & Dhariwal, P. (2021). Improved Denoising Diffusion Probabilistic Models. *ArXiv, abs/2102.09672*.

Nie, S., Zhu, F., You, Z., Zhang, X., Ou, J., Hu, J., ... & Li, C. (2025). Large Language Diffusion Models. *arXiv preprint arXiv:2502.09992.*

Sahoo, S.S., Arriola, M., Schiff, Y., Gokaslan, A., Marroquin, E.M., Chiu, J.T., Rush, A., & Kuleshov, V. (2024). Simple and Effective Masked Diffusion Language Models. *ArXiv, abs/2406.07524.*

Salimans, T., & Ho, J. (2022). Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*.

Song, J., Meng, C., & Ermon, S. (2020). Denoising Diffusion Implicit Models. *ArXiv, abs/2010.02502*.

Yu, J., & Barad, H. (2024). Step Saver: Predicting Minimum Denoising Steps for Diffusion Model Image Generation. *ArXiv, abs/2408.02054*.