SCHRÖDINGER'S THINKING: REASONING VIA SUPERPOSITION IN DIFFUSION LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion large language models offer a promising alternative to sequential reasoning by predicting many parts of a solution in parallel. However, we find that their reasoning is still limited by an important form of sequential commitment: each prediction must collapse to a single discrete token, even when the future remains uncertain. We introduce reasoning via superposition, a framework that delays this commitment. Instead of selecting one token immediately, our approach represents an intermediate prediction as a continuous mixture of several plausible tokens. This allows multiple possible reasoning paths to remain active and jointly influence later predictions until the model has enough context to make a discrete decision. We further introduce reinforcement learning that teaches diffusion language models to reason through these superposed states. Across several diffusion language models and reasoning benchmarks, superposition improves reasoning accuracy at inference time without changing model parameters, and reinforcement learning further increases the gains to 7.3 points. Our analysis shows that preserving alternative hypotheses is especially important early in reasoning, when the correct future has not yet been determined.

1 INTRODUCTION

Autoregressive large language models (LLMs) achieve remarkable performance across a variety of complex tasks, yet they remain fundamentally constrained by their sequential nature Bengio et al. (2015); Vaswani et al. (2017); Radford et al. (2019). By generating text strictly left-to-right, these models must commit to individual tokens before a global reasoning structure is fully established. This limitation prevents current models from reasoning while keeping the future unresolved—an ability central to long-horizon planning Schacter et al. (2012); Botvinick & Toussaint (2012); Stachenfeld et al. (2017), underscoring the gap between sequential machine generation and fluid human-level reasoning.

A growing body of work has explored diffusion language models to enable global, non-causal reasoning Austin et al. (2021); Li et al. (2022); Strudel et al. (2022); Chang et al. (2022); Ni et al. (2026). By generating tokens in flexible orders, diffusion models promise to bypass the left-to-right bottleneck. However, while existing diffusion architectures focus on flexible generation orders, they are burdened because they still predict outputs in a strictly discrete language space Bie et al. (2025). Because diffusion LLMs must fix a single discrete token at every step, existing approaches force the model to prematurely commit to discrete words during the denoising process—even when no token is ready to be resolved, see Figure 1. This premature discretization cannot adapt to the uncertainty of early generation stages, forcing high-confidence guesses about distant, underspecified parts of the reasoning process and causing failures on complex tasks Gu et al. (2017); Wang et al. (2022); Zhou et al. (2022); Huang et al. (2025).

In this paper, we introduce an approach for preserving uncertainty over future states, allowing us to formulate a reasoning strategy that avoids premature commitment. Just as humans maintain a hazy set of plausible outcomes to guide intermediate steps, our approach maintains future tokens in *superposition*: instead of collapsing to a single discrete word, the model represents the future as a continuous linear combination of candidate tokens. Figure 1 visualizes our superpositioned reasoning on a complex math problem. However, reasoning via superposition is challenging because existing diffusion models are trained solely to target and output discrete tokens.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

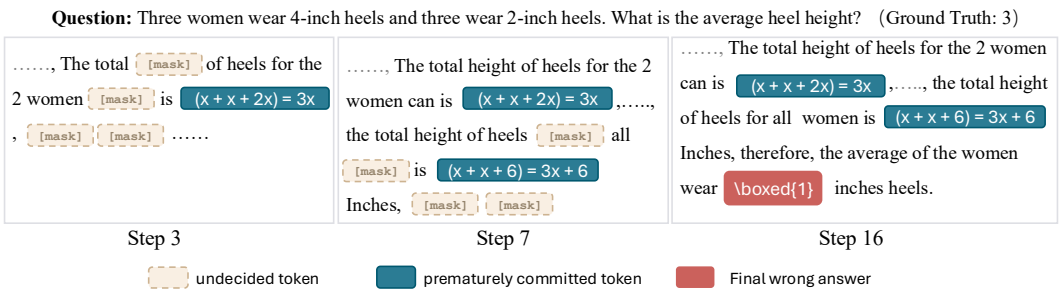


Figure 1: **Premature token commitment in masked diffusion decoding.** We show selected states from a masked diffusion model's reasoning trajectory. Once an uncertain prediction is committed to a discrete token, it becomes fixed and can steer subsequent reasoning toward an incorrect solution. Diffusion models, however, iteratively refine predictions using bidirectional context. This motivates our central hypothesis: delaying commitment to uncertain future tokens can lead to better reasoning.

Our key insight is that diffusion LLMs contain rich, continuous reasoning priors that are naturally exposed when we delay discretization. We found that forcing early token collapse obscures these internal priors, collapsing the future thinking trace before a better path is identified. Our approach shows how to capitalize on this continuous space to keep multiple plausible reasoning trajectories active at once, allowing them to jointly shape intermediate steps. Furthermore, we introduce a reinforcement learning objective that operates directly on the superposed future, guaranteeing that the model learns to inherently optimize its planning and leverage this uncertainty, providing more stable reasoning than purely discrete training methods.

Since superposition is adaptive, the advantage of our method scales efficiently with the "density" of the future: as the model attempts to predict more tokens in a single step, maintains a larger number of possibilities helps more. Moreover, the modularity of our inference approach allows it to work with any diffusion LLM architecture. It seamlessly complements both full-attention models (such as LLaDA Nie et al. (2025)) and block-diffusion models (such as SDAR Cheng et al. (2025)), without requiring fundamental architectural changes.

Visualizations, empirical experiments, and structural analyses show that our superposition strategy significantly improves robust reasoning for several established benchmarks. Our method advances state-of-the-art diffusion LLMs by a large margin across diverse datasets, achieving up to a 6.7 point gain during inference alone. When combined with our reinforcement learning framework, accuracy improves by up to 7.3 points on benchmarks including MATH500 Hendrycks et al. (2021), GSM-Plus Li et al. (2024), GPQA-Diamond Rein et al. (2024), and MMLU-Pro Wang et al. (2024). Finally, qualitative visualizations demonstrate that our approach preserves semantically critical alternatives at key transition points. Together, these results suggest that robust non-causal intelligence lies not in better next-token prediction, but in allowing the model's internal thoughts to remain fluid.

2 Related Work

Discrete Diffusion Language Models. Discrete diffusion language models (DLMs) operate directly on the token vocabulary, bypassing continuous latent variables Li et al. (2022); Strudel et al. (2022); Gong et al. (2022); Yu et al. (2022); Mahabadi et al. (2024). A highly effective formulation is the masked DLM Lou et al. (2023); Zheng et al. (2023); Ou et al. (2024); Shi et al. (2024); Sahoo et al. (2024), which progressively corrupts text by replacing tokens with a `[MASK]` symbol. During inference, these models iteratively refine a fully masked sequence by predicting tokens in parallel, fixing high-confidence predictions, and re-masking uncertain ones. Current masked DLMs generally follow one of two structural paradigms: full-attention models (e.g., LLaDA Nie et al. (2025), Dream Ye et al. (2025)) that apply bidirectional attention across the entire sequence, and block-diffusion models (e.g., SDAR Cheng et al. (2025), Fast-dLLM v2 Wu et al. (2025a)) that generate sequential blocks while predicting intra-block tokens in parallel. While these paradigms extend to multimodal tasks Yang et al. (2025); Cheng et al. (2025); You et al. (2025), both fundamentally rely on multi-token parallel decoding to discrete ones, which is the mechanism that introduces premature discrete commitment. DLMs control generation speed through explicit decoding strategies. *Static decoding* Nie et al. (2025) unmasks a fixed number of high-confidence tokens per step, while *dynamic*

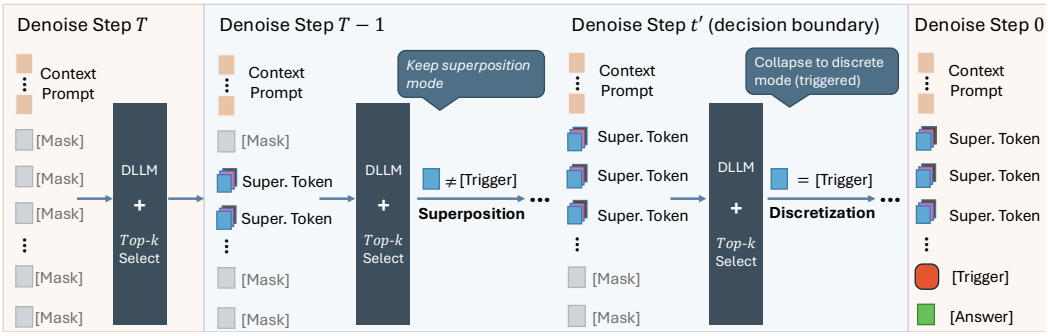


Figure 2: **Superposition inference in diffusion language models.** We illustrate the denoising process of a diffusion language model with superposition inference. Starting from a fully masked sequence (Denoise Step T), the model progressively fills masked positions using top-K selection. Before the decision boundary (T–$1 \to t'$), selected positions are represented as *superposition tokens*—continuous mixtures of multiple high-probability discrete tokens—which preserve uncertainty and maintain multiple future hypotheses. At the decision boundary, a trigger token determines whether the model remains in superposition mode or collapses to discrete decoding. Once the trigger is emitted, decoding proceeds in standard discrete mode to produce readable final outputs at Step 0.

decoding Cheng et al. (2025) unmasks all tokens exceeding a confidence threshold. Both approaches accelerate inference by predicting multiple tokens simultaneously, but aggressive parallelism forces the model to commit to tokens under higher uncertainty, compounding early errors.

Continuous-Space Reasoning. Recent studies question whether discrete token spaces are optimal for long-horizon planning Chen et al. (2024); Su et al. (2025); Xu et al. (2025); Ruan et al. (2025). In autoregressive models, reasoning is often encoded in continuous hidden states rather than surface-level tokens Yang et al. (2024); Biran et al. (2024), with some evidence pointing to implicit parallel hypotheses Shalev et al. (2024). To explicitly leverage continuous reasoning, prior work generally falls into two categories. *Inference-time methods* use soft token mixtures to delay discrete commitment in frozen models Zhang et al. (2025); Wu et al. (2025b), though they are often sensitive to heuristics. *Post-training methods* explicitly teach models to reason in continuous latent spaces, such as using final-layer hidden states (Coconut Hao et al. (2024)) or continuous distillation (Codi Shen et al. (2025)). Unlike these autoregressive approaches, our work introduces continuous superposition reasoning natively into the parallel decoding process of diffusion models, bridging both inference-time flexibility and end-to-end training.

3 Method

Given a sequence of tokens $x = (x_1, \ldots, x_N)$, our goal is to model the joint distribution $p(x)$ using a discrete denoising diffusion language model. Unlike autoregressive models, which factorize generation left-to-right, diffusion language models (DLMs) generate tokens in parallel by iteratively refining a corrupted sequence, `[MASK]` tokens, into a clean one. At each step t, the model $p_\theta(x \mid x^{(t)})$ predicts discrete tokens for a selected subset of masked positions $\mathcal{P}^{(t)}$, which are then fixed for all subsequent steps.

For a clean input sequence $x^{(0)}$, the forward diffusion process randomly replaces tokens with a special mask token `[MASK]`. Let $t \in [0, 1]$ denote a continuous noise level. The forward noising process factorizes over token positions and is defined as:

$$q(x^{(t)} \mid x^{(0)}) = \prod_{i=1}^{N} \mathrm{Cat}\Big(x_i^{(t)};\ (1-t)\,\delta_{x_i^{(0)}} + t\,\delta_{\texttt{[MASK]}}\Big), \tag{1}$$

where $x_i^{(t)}$ denotes the token at position i after noising. As t increases, a larger fraction of tokens are replaced by `[MASK]`, with the fully masked sequence obtained at the terminal diffusion step.

The reverse denoising process is parameterized by a neural network p_θ, which predicts the original tokens conditioned on a partially masked sequence: $p_\theta(x^{(t-1)} \mid x^{(t)})$. In practice, masked diffusion

language models are trained using a simplified objective that applies cross-entropy loss only on masked positions.

However, standard DLMs suffer from *premature commitment*: at every step, reasoning is forced to collapse one additional token into a single discrete token choice. This discretization on tokens that are not yet ready to be decoded destroys the model's ability to maintain uncertainty and explore parallel reasoning paths. To overcome this, we introduce *Superposition Inference*, which delays discrete commitment by operating in the continuous embedding space, and *Superposition Reinforcement Learning (SuperRL)*, which fine-tunes the model to actively leverage this latent ambiguity.

3.1 Superposition Inference

We propose to mitigate this premature commitment error via committing to a cloud of possible tokens if have to. As shown in Figure 2, the insight of our method is that intermediate reasoning states should not be forced onto the vertices of the vocabulary simplex (discrete tokens) but should instead inhabit the interior of the embedding space, where uncertainty over multiple hypotheses can be preserved.

At diffusion step t, for a selected position i, the model produces logits $\ell_i^{(t)}$. Let $\mathcal{S}_i^{(t)}$ be the top-K candidate tokens. We define a *superposition token* $\tilde{e}_i^{(t)}$ as a continuous mixture of these candidates:

$$\tilde{e}_i^{(t)} = \sum_{k=1}^{K} w_{ik}^{(t)} e(v_{ik}), \quad w_{ik}^{(t)} = \frac{\exp(\ell_{ik}^{(t)}/\tau)}{\sum_{j=1}^{K} \exp(\ell_{ij}^{(t)}/\tau)} \tag{2}$$

where $e(\cdot)$ is the embedding function and τ controls the sharpness of the mixture.

Instead of sampling a discrete token, we update the position with $\tilde{e}_i^{(t)}$. Crucially, these continuous embeddings persist across subsequent denoising steps $t' < t$, serving as the observed context $\tilde{X}^{(t)}$. This allows multiple hypotheses to coexist and actively shape the evolving reasoning trajectory.

Switching to Discrete Decoding. Superposition inference is designed for intermediate reasoning. To guarantee a human-readable final output, we introduce a model-driven trigger. If the model predicts a designated trigger token v_{trig} at any selected position, we terminate superposition and revert to standard discrete decoding for all remaining steps.

3.2 Superposition-Based Reinforcement Learning

While Superposition Inference enables continuous reasoning at inference time, pre-trained diffusion language models are not explicitly trained to propagate or refine superposed states across diffusion steps. To address this mismatch, we introduce *Superposition Reinforcement Learning* (SuperRL). By leveraging the superpositioned rollouts detailed in Section 3.1 and building on the trajectory-aware TraceRL framework Wang et al. (2025), SuperRL adapts the model parameters to actively internalize superposition-based reasoning.

Trajectory Likelihood. A key design choice in applying RL to superposition inference is defining the likelihood of a trajectory. While the rollout trajectory is generated using continuous top-K superposition embeddings, we empirically find that calculating the policy likelihood directly over these continuous mixtures is suboptimal. Instead, rather than matching the model's full softmax output against the superposed top-K distribution, we evaluate the likelihood strictly with respect to the top-1 discrete token hypothesis $v_i^{(t)}$ selected at each step. Given a diffusion rollout trajectory $\tau = \{\tau(1), \ldots, \tau(T)\}$, we define the sequence likelihood by evaluating these discrete top-1 tokens against the corresponding superposition-induced context $\tilde{X}^{(t)}$:

$$\log \pi_\theta(\tau) = \sum_{t=1}^{T} \sum_{i \in \mathcal{P}^{(t)}} \log p_\theta\Big(v_i^{(t)} \mid \tilde{X}^{(t)}\Big) \tag{3}$$

where $\tilde{X}^{(t)}$ contains the prompt tokens, persistent superposition embeddings from previous steps, and the remaining masks.

Table 1: Inference-time comparison between standard discrete decoding (Base) and Superposition Inference (**SuperInfer**) across multiple diffusion language models and reasoning benchmarks. SuperInfer is applied at inference time only, without modifying model parameters. Numbers in parentheses indicate the absolute performance change relative to the Base model, with blue (+) denoting improvements and red (−) denoting degradations. Overall, SuperInfer consistently improves reasoning accuracy across models and datasets.

Model	Variant	MATH500	GSM-Plus (test-mini)	GPQA-Diamond	MMLU-Pro
LLaDA-8B-Instruct	Base	37.1	69.1	20.7	31.4
	+ SuperInfer (Ours)	39.4 (+2.3)	70.9 (+1.8)	25.3 (+4.6)	33.8 (+2.4)
MMaDA-8B-MixCoT	Base	13.4	28.3	24.2	20.9
	+ SuperInfer (Ours)	15.0 (+1.6)	28.0 (-0.3)	26.8 (+2.6)	22.0 (+1.1)
SDAR-1.7B-Chat	Base	48.8	61.2	21.9	20.6
	+ SuperInfer (Ours)	55.5 (+6.7)	62.8 (+1.6)	23.3 (+1.4)	25.0 (+4.4)
SDAR-4B-Chat	Base	62.4	76.7	22.7	37.6
	+ SuperInfer (Ours)	67.2 (+4.8)	81.2 (+4.5)	25.3 (+2.6)	39.9 (+2.3)

Policy Optimization. We optimize the superposition policy using Proximal Policy Optimization (PPO) Schulman et al. (2017), which constrains updates relative to a reference policy π_{old} to ensure stable training. The PPO objective is given by:

$$L(\theta) = \mathbb{E}_{\tau\sim\pi_{\text{old}}}\Big[\min\Big(r_\theta(\tau)\,A(\tau), \text{clip}\big(r_\theta(\tau), 1-\epsilon, 1+\epsilon\big)\,A(\tau)\Big)\Big]. \tag{4}$$

where $A(\tau)$ denotes the advantage estimate computed from the trajectory-level reward $R(\tau)$, which is calculated based on the correctness of the output answer.

The probability ratio is defined using the trajectory likelihood evaluated from the Superposition Inference rollout:

$$\log r_\theta(\tau) = \log \pi_\theta(\tau) - \log \pi_{\text{old}}(\tau) = \sum_{t=1}^{T} \sum_{i\in\mathcal{P}^{(t)}} \Big[\log p_\theta\Big(v_i^{(t)} \mid \tilde{X}^{(t)}\Big) - \log p_{\text{old}}\Big(v_i^{(t)} \mid \tilde{X}^{(t)}\Big)\Big] \tag{5}$$

Here, $\tilde{X}^{(t)}$ is the diffusion input induced by the Superposition Inference rollout, and the logits for both π_θ and π_{old} are recomputed on the same rollout states for likelihood evaluation.

Gradient Computation. The resulting policy gradient takes the form:

$$\nabla_\theta L(\theta) = \mathbb{E}_{\tau\sim\pi_{\text{old}}}\left[A(\tau)\sum_{t=1}^{T}\sum_{i\in\mathcal{P}^{(t)}} \nabla_\theta \log p_\theta\Big(v_i^{(t)} \mid \tilde{X}^{(t)}\Big)\right], \tag{6}$$

This gradient corresponds to a policy update over discrete token probabilities evaluated along continuous rollout states. Intuitively, it directly upweights token hypotheses that contribute to effective superposition trajectories that end up with the right answer. By optimizing this objective, SuperRL fine-tunes the model's vector field to navigate the *Superposition Concept Space*. After training, the model shall be able to propagate, manage, and refine latent ambiguity across steps rather than collapsing uncertainty prematurely.

4 Experiments & Results

4.1 Experiment Setup

Benchmarks. We evaluate our method on a diverse set of reasoning benchmarks covering mathematical, scientific, and multi-domain question answering. Specifically, we report results on MATH500 Hendrycks et al. (2021), GSM-Plus (test-mini) Li et al. (2024), GPQA-Diamond Rein et al. (2024), and MMLU-Pro Wang et al. (2024). For MMLU-Pro, we uniformly sample 10% of examples from each domain.

Models. We conduct experiments on several representative diffusion language models, including LLaDA-8B-Instruct Nie et al. (2025), MMaDA-8B-MixCoT Yang et al. (2025), SDAR-1.7B-Chat,

Table 2: Comparison between inference-time superposition inference (**SuperInfer**) and training-time superposition reinforcement learning (**SuperRL**) on MATH500 and GSM-Plus. Post-training diffusion LLMs to reason with superpositioned tokens via reinforcement learning further improves reasoning accuracy beyond inference-time superposition alone.

Model	Variant	MATH500	GSM-Plus
LLaDA-8B	Base	37.1	69.1
	+ SuperInfer (Ours)	39.4 (+2.3)	70.9 (+1.8)
	+ SuperRL (Ours)	42.8 (+5.7)	71.5 (+2.4)
SDAR-4B	Base	62.4	76.7
	+ SuperInfer (Ours)	67.2 (+4.8)	81.2 (+4.5)
	+ SuperRL (Ours)	69.7 (+7.3)	83.1 (+6.4)

and SDAR-4B-Chat Cheng et al. (2025). These models span different architectural designs and scales, allowing us to assess the generality of superposition inference across diffusion-based reasoning models.

Hyperparameter Settings. Unless otherwise specified, all experiments use dynamic decoding and are conducted in a zero-shot setting. We use dynamic decoding with a confidence threshold of 0.9 for all models. We set the temperature to 1.0, use top-k sampling with $k = 16$, and top-p sampling with $p = 1.0$. For Superposition Inference, we use $K = 4$ superposition tokens at each selected position. We further analyze the effect of varying K in Section 4.3. For Superposition Reinforcement Learning, we train on the Math-train dataset Wang et al. (2025) using minibatch policy optimization with a learning rate of 1×10^{-6}, where each update uses 64 prompts with 16 trajectories per prompt (64×16 samples).

4.2 Main Results

We first evaluate SuperInfer as a training-free inference-time mechanism. Table 1 reports results across four reasoning benchmarks. SuperInfer consistently improves accuracy over standard discrete diffusion decoding without modifying model parameters. These gains span both full-attention architectures, including LLaDA-8B-Instruct and MMaDA-8B-MixCoT, and block-diffusion architectures, including SDAR-1.7B-Chat and SDAR-4B-Chat.

The improvements are particularly pronounced on challenging benchmarks such as MATH500 and GPQA-Diamond, reaching up to 6.7 percentage points. These results support preserving uncertainty during decoding: delaying discrete commitment allows alternative hypotheses to inform subsequent predictions, improving reasoning through a simple inference-time modification.

We next examine whether training can further improve reasoning with superposition. We apply SuperRL to LLaDA-8B-Instruct and SDAR-4B-Chat, covering full-attention and block-diffusion architectures. Table 2 shows that SuperRL improves upon SuperInfer on both MATH500 and GSM-Plus. For LLaDA-8B, MATH500 accuracy increases from 39.4 with SuperInfer to 42.8 with SuperRL, a gain of 5.7 percentage points over Base. For SDAR-4B, it increases from 67.2 to 69.7, exceeding Base by 7.3 points. The corresponding gains on GSM-Plus show that the benefits extend across both mathematical reasoning benchmarks.

These results demonstrate the complementary benefits of inference-time superposition and training-time optimization. SuperInfer improves reasoning by preserving multiple candidate hypotheses without changing model parameters, while SuperRL trains the model to use superposed representations more effectively. The additional gains across both architectures indicate that superposition provides a useful inference mechanism and a representation that can be further improved through reinforcement learning.

4.3 Analysis

Effect of Decoding Multiple Tokens per Diffusion Step. We evaluate the effect of decoding parallelism on MATH500 using LLaDA-8B and SDAR-4B, comparing standard discrete decoding (Base) with Superposition Inference (+SuperInfer, $K = 4$). As summarized in Table 3, we consider two schedules: *static decoding*, which predicts a fixed number of tokens per diffusion step, varied over $\{1, 2, 4\}$; and *dynamic decoding*, which selects positions according to a confidence threshold,

Table 3: **Robustness to decoding parallelism on MATH500 (Pass@1).** Static decoding varies the number of tokens predicted per step; dynamic decoding varies the confidence threshold. Bold denotes the better result within each model and setting.

		Static: tokens / step			Dynamic: confidence threshold				
Model	**Method**	1	2	4	0.9	0.8	0.7	0.6	0.5
LLaDA-8B	Base	36.2	34.2	22.2	37.1	36.9	35.8	34.2	28.6
	+SuperInfer	**38.4**	**36.8**	**28.8**	**39.4**	**38.4**	**36.4**	**35.5**	**32.1**
SDAR-4B	Base	65.6	36.7	4.0	62.4	61.7	57.0	59.4	**55.9**
	+SuperInfer	**67.4**	**56.6**	**20.6**	**67.2**	**65.2**	**64.8**	**62.5**	55.5

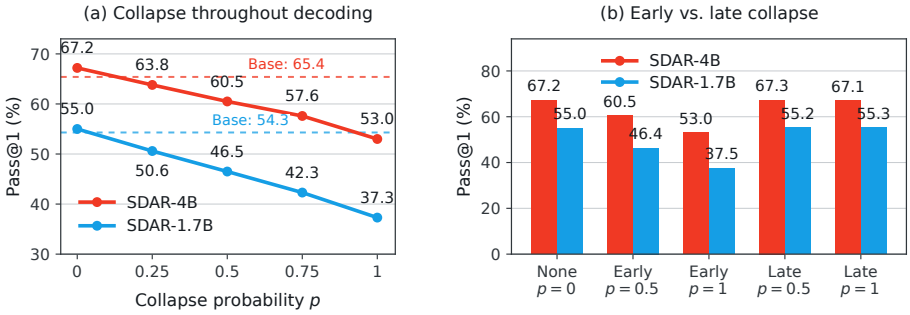


Figure 3: **Controlled collapse of superposition states on MATH500.** At each position selected for commitment, p is the probability of replacing its superposition embedding with a sampled discrete token. **(a)** Collapse is applied throughout decoding; dashed lines mark the standard discrete-decoding Base accuracy. **(b)** *Early* and *Late* restrict this intervention to the first and second halves of the response block, respectively; positions outside the selected half retain normal Superposition Inference. Within that half, $p = 0.5$ collapses each selected position with probability 0.5, whereas $p = 1$ collapses every selected position. *None* ($p = 0$) denotes Superposition Inference without forced collapse. Early collapse causes substantially larger accuracy losses than late collapse. All values are MATH500 Pass@1 percentages.

varied over $\{0.9, 0.8, 0.7, 0.6, 0.5\}$. Lower thresholds permit more positions to be decoded in parallel. We report Pass@1 accuracy, with the remaining settings following Section 4.1.

Under static decoding, increasing the number of tokens per step from 1 to 4 reduces Base accuracy from 36.2 to 22.2 for LLaDA-8B and from 65.6 to 4.0 for SDAR-4B. SuperInfer also loses accuracy, but retains 28.8 and 20.6, respectively, at 4 tokens per step, exceeding Base by 6.6 and 16.6 points. Under dynamic decoding, SuperInfer improves accuracy in most settings; for example, at threshold 0.7, SDAR-4B improves from 57.0 to 64.8. These results suggest that preserving multiple hypotheses can mitigate the accuracy loss from parallel token commitment, particularly under aggressive static schedules, although it does not eliminate this loss in every setting.

Effect of Premature Superposition Collapse. We conduct a controlled intervention on MATH500 using SDAR-4B and SDAR-1.7B with Superposition Inference ($K = 4$), measuring Pass@1 accuracy. At each position selected for commitment, we independently replace its superposition embedding with a sampled discrete token with probability p; otherwise, the position retains its superposition embedding. For Figure 3(a), this intervention is applied throughout decoding, with $p \in \{0, 0.25, 0.5, 0.75, 1\}$. Thus, $p = 0$ is unmodified Superposition Inference, while $p = 1$ forces every selected position to collapse. The dashed lines provide the standard discrete-decoding Base reference; the remaining settings follow Section 4.1.

Accuracy decreases monotonically as p increases. For SDAR-4B, it falls from 67.2 at $p = 0$ to 63.8, 60.5, 57.6, and 53.0 at $p = 0.25$, 0.5, 0.75, and 1, respectively. SDAR-1.7B exhibits the same trend, decreasing from 55.0 to 37.3. This controlled degradation supports the interpretation that retaining superposition states across denoising steps contributes to reasoning accuracy.

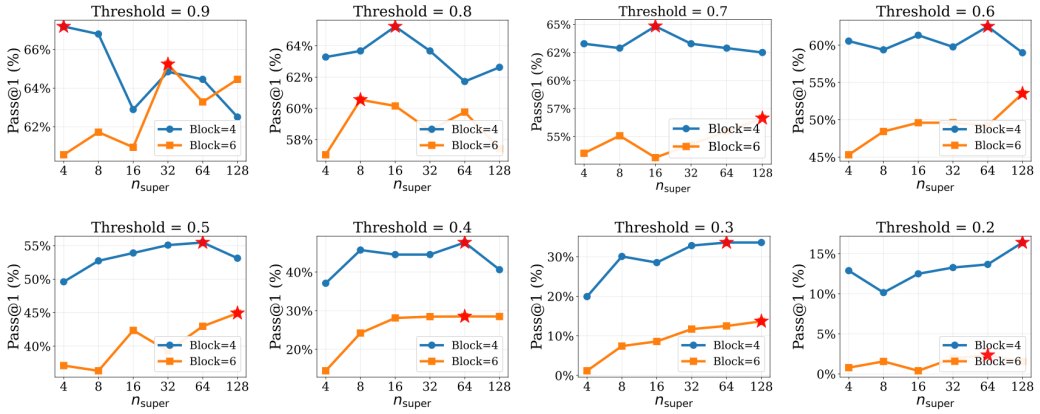


Figure 4: **Superposition capacity under increasingly aggressive dynamic decoding.** We vary the confidence threshold and block length to control decoding parallelism and uncertainty. The x-axis shows the number of superposition tokens (n_{super}), and the y-axis reports Pass@1 accuracy. Red stars indicate the best-performing capacity in each setting. As the confidence threshold decreases, the best-performing tested n_{super} generally shifts toward larger values; block length 6 often favors larger capacities than block length 4.

Early vs. Late Collapse. Using the same models, benchmark, and metric, we next restrict the intervention to the first half (*Early*) or second half (*Late*) of the response block, as shown in Figure 3(b). We test $p \in \{0.5, 1\}$ within the selected half; positions outside it retain normal Superposition Inference. The $p = 0$ condition serves as the reference without forced collapse.

The two intervention regions have markedly different effects. At $p = 1$, Early collapse reduces accuracy from 67.2 to 53.0 for SDAR-4B and from 55.0 to 37.5 for SDAR-1.7B, whereas Late collapse yields 67.1 and 55.3, respectively. The same asymmetry holds at $p = 0.5$: Early collapse yields 60.5 and 46.4, compared with 67.3 and 55.2 for Late collapse. These results indicate that preserving superposition in the first half of the response block is particularly important, consistent with the hypothesis that premature commitments can constrain subsequent reasoning.

Visualization of Reasoning Trajectories. Figure 5 compares six selected baseline-wrong/SuperInfer-correct cases from GSM-Plus, MATH500, and MMLU-Pro using SDAR-4B and SDAR-1.7B. In (a), (c), (d), and (e), we replay earlier baseline commitments, replace the selected token's feedback with a top-K mixture ($K = 4$) while keeping its written token unchanged, and continue with SuperInfer. Panels (b) and (f) compare completed-response excerpts.

Local errors carry into later calculations: the incorrect DVD count 2712 is reused in the total 4982 in (a), the sign error yields $c = -1.91$ and $b + c = -0.61$ in (c), and 4666 propagates to the answer 70.23 in (d). SuperInfer continuations recover the correct final answers 3018, 3.21, and 66, respectively. Omitted income in (b) and a repeated acceleration factor in (f) show similar error propagation. In (e), recovery occurs at answer selection despite an incorrect intermediate speed.

Effect of Sampling Temperature. We further analyze the robustness of Superposition Inference under different sampling temperatures. Temperature controls the stochasticity of token prediction: higher temperatures flatten the output distribution and make decoding less deterministic, thereby increasing uncertainty during the denoising process. As shown in Table 4, Superposition Inference remains effective across a broad range of temperatures on MATH500.

The improvement is especially pronounced at higher temperatures. For example, when $\tau = 2.0$, LLaDA-8B drops sharply from 37.1 at $\tau = 1.0$ to 12.0 under standard decoding, whereas SuperInfer still achieves 32.1. SDAR-4B improves from 45.0 to 63.4 with SuperInfer at $\tau = 2.0$. These results suggest that SuperInfer is particularly helpful in more stochastic and challenging decoding regimes, where discrete decoding is more likely to commit prematurely to unstable token hypotheses.

Interaction Between Dynamic Decoding and Superposition Capacity. We evaluate SDAR-4B-Chat under dynamic decoding, sweeping the confidence threshold from 0.9 to 0.2 in steps of 0.1 and comparing block lengths of 4 and 6 token positions. For each setting, we vary the superposition

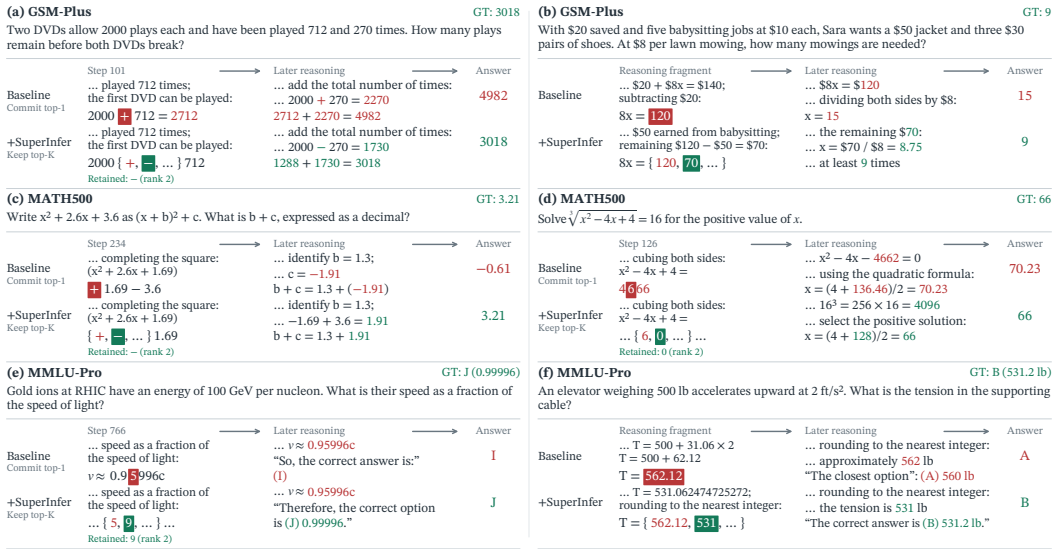


Figure 5: **Reasoning trajectories with discrete decoding and SuperInfer.** Six selected baseline-wrong/SuperInfer-correct cases, two per dataset. Red marks errors; green marks correct alternatives or outcomes. GT denotes the ground-truth answer. Braces show recorded feedback candidates in (a), (c), (d), and (e), and contrasted response values in (b) and (f). Steps mark commitment times; context comes from completed responses. Ellipses abbreviate the display.

Table 4: Temperature sweep on MATH500 (Pass@1). Superposition Inference remains effective across a broad range of sampling temperatures. The gains become more pronounced at higher temperatures, where standard decoding becomes less stable. For LLaDA, SuperInfer reduces to the baseline decoding at $\tau = 0$; for SDAR, $\tau = 0$ is not supported, so we report $\tau = 0.1$ instead.

Temperature τ	**LLaDA-8B**		**SDAR-4B**	
	Base	+SuperInfer	Base	+SuperInfer
0.0 / 0.1	38.4	38.4	46.4	47.2
0.5	37.5	38.8	64.5	65.9
1.0	37.1	39.4	62.4	67.2
1.5	33.2	37.3	60.1	66.5
2.0	12.0	32.1	45.0	63.4

capacity $K \in \{4, 8, 16, 32, 64, 128\}$, denoted by n_{super} in Figure 4, and report Pass@1 accuracy. Red stars identify the best-performing tested capacity for each block length and threshold. The remaining settings follow Section 4.1.

The best-performing capacity generally shifts toward larger values as the confidence threshold decreases. For block length 4, it increases from $K = 4$ at threshold 0.9 to $K = 16$ at thresholds 0.8 and 0.7, reaching $K = 128$ at threshold 0.2. Block length 6 often favors a larger capacity than block length 4. These results suggest that the useful superposition capacity depends on the uncertainty introduced by the decoding schedule. Lower confidence thresholds permit earlier commitments, while larger blocks expose more positions to parallel prediction. Maintaining a broader set of candidate hypotheses can help accommodate this uncertainty, but the best capacity depends on the specific combination of threshold and block length.

5 Conclusion

We introduce a framework for reasoning via superposition to overcome premature discrete commitment in Diffusion LLMs. Our results demonstrate that maintaining a continuous representation of future states significantly improves reasoning performance across diverse benchmarks. Our findings suggest that shifting from sequential generation to holistic, iterative refinement is a promising direction for unlocking non-causal, long-horizon intelligence.

REFERENCES

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in neural information processing systems*, 34:17981–17993, 2021.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. *Advances in neural information processing systems*, 28, 2015.

Tiwei Bie, Maosong Cao, Kun Chen, Lun Du, Mingliang Gong, Zhuochen Gong, Yanmei Gu, Jiaqi Hu, Zenan Huang, Zhenzhong Lan, et al. Llada2. 0: Scaling up diffusion language models to 100b. *arXiv preprint arXiv:2512.15745*, 2025.

Eden Biran, Daniela Gottesman, Sohee Yang, Mor Geva, and Amir Globerson. Hopping too late: Exploring the limitations of large language models on multi-hop queries. *arXiv preprint arXiv:2406.12775*, 2024.

Matthew Botvinick and Marc Toussaint. Planning as inference. *Trends in cognitive sciences*, 16(10): 485–488, 2012.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11315–11325, 2022.

Haolin Chen, Yihao Feng, Zuxin Liu, Weiran Yao, Akshara Prabhakar, Shelby Heinecke, Ricky Ho, Phil Mui, Silvio Savarese, Caiming Xiong, et al. Language models are hidden reasoners: Unlocking latent reasoning capabilities via self-rewarding. *arXiv preprint arXiv:2411.04282*, 2024.

Shuang Cheng, Yihan Bian, Dawei Liu, Linfeng Zhang, Qian Yao, Zhongbo Tian, Wenhai Wang, Qipeng Guo, Kai Chen, Biqing Qi, et al. Sdar: A synergistic diffusion-autoregression paradigm for scalable sequence generation. *arXiv preprint arXiv:2510.06303*, 2025.

Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and LingPeng Kong. Diffuseq: Sequence to sequence text generation with diffusion models. *arXiv preprint arXiv:2210.08933*, 2022.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor OK Li, and Richard Socher. Non-autoregressive neural machine translation. *arXiv preprint arXiv:1711.02281*, 2017.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*, 2024.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.

Qintong Li, Leyang Cui, Xueliang Zhao, Lingpeng Kong, and Wei Bi. Gsm-plus: A comprehensive benchmark for evaluating the robustness of llms as mathematical problem solvers. *arXiv preprint arXiv:2402.19255*, 2024.

Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. *Advances in neural information processing systems*, 35: 4328–4343, 2022.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. *arXiv preprint arXiv:2310.16834*, 2023.

Rabeeh Karimi Mahabadi, Hamish Ivison, Jaesung Tae, James Henderson, Iz Beltagy, Matthew E Peters, and Arman Cohan. Tess: Text-to-text self-conditioned simplex diffusion. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2347–2361, 2024.

Zanlin Ni, Shenzhi Wang, Yang Yue, Tianyu Yu, Weilin Zhao, Yeguo Hua, Tianyi Chen, Jun Song, Cheng Yu, Bo Zheng, et al. The flexibility trap: Rethinking the value of arbitrary order in diffusion language models. *arXiv preprint arXiv:2601.15165*, 2026.

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

Jingyang Ou, Shen Nie, Kaiwen Xue, Fengqi Zhu, Jiacheng Sun, Zhenguo Li, and Chongxuan Li. Your absorbing discrete diffusion secretly models the conditional distributions of clean data. *arXiv preprint arXiv:2406.03736*, 2024.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024.

Yangjun Ruan, Neil Band, Chris J Maddison, and Tatsunori Hashimoto. Reasoning to learn from latent thoughts. *arXiv preprint arXiv:2503.18866*, 2025.

Subham Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. *Advances in Neural Information Processing Systems*, 37:130136–130184, 2024.

Daniel L Schacter, Donna Rose Addis, Demis Hassabis, Victoria C Martin, R Nathan Spreng, and Karl K Szpunar. The future of memory: remembering, imagining, and the brain. *Neuron*, 76(4): 677–694, 2012.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Yuval Shalev, Amir Feder, and Ariel Goldstein. Distributional reasoning in llms: Parallel reasoning processes in multi-hop reasoning. *arXiv preprint arXiv:2406.13858*, 2024.

Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. Codi: Compressing chain-of-thought into continuous space via self-distillation. *arXiv preprint arXiv:2502.21074*, 2025.

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis Titsias. Simplified and generalized masked diffusion for discrete data. *Advances in neural information processing systems*, 37: 103131–103167, 2024.

Kimberly L Stachenfeld, Matthew M Botvinick, and Samuel J Gershman. The hippocampus as a predictive map. *Nature neuroscience*, 20(11):1643–1653, 2017.

Robin Strudel, Corentin Tallec, Florent Altché, Yilun Du, Yaroslav Ganin, Arthur Mensch, Will Grathwohl, Nikolay Savinov, Sander Dieleman, Laurent Sifre, et al. Self-conditioned embedding diffusion for text generation. *arXiv preprint arXiv:2211.04236*, 2022.

DiJia Su, Hanlin Zhu, Yingchen Xu, Jiantao Jiao, Yuandong Tian, and Qinqing Zheng. Token assorted: Mixing latent and text tokens for improved language model reasoning. *arXiv preprint arXiv:2502.03275*, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Yinjie Wang, Ling Yang, Bowen Li, Ye Tian, Ke Shen, and Mengdi Wang. Revolutionizing reinforcement learning framework for diffusion large language models. *arXiv preprint arXiv:2509.06949*, 2025.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37: 95266–95290, 2024.

Chengyue Wu, Hao Zhang, Shuchen Xue, Zhijian Liu, Shizhe Diao, Ligeng Zhu, Ping Luo, Song Han, and Enze Xie. Fast-dllm: Training-free acceleration of diffusion llm by enabling kv cache and parallel decoding. *arXiv preprint arXiv:2505.22618*, 2025a.

Junhong Wu, Jinliang Lu, Zixuan Ren, Gangqiang Hu, Zhi Wu, Dai Dai, and Hua Wu. Llms are single-threaded reasoners: Demystifying the working mechanism of soft thinking. *arXiv preprint arXiv:2508.03440*, 2025b.

Yige Xu, Xu Guo, Zhiwei Zeng, and Chunyan Miao. Softcot: Soft chain-of-thought for efficient reasoning with llms. *arXiv preprint arXiv:2502.12134*, 2025.

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. Mmada: Multimodal large diffusion language models. *arXiv preprint arXiv:2505.15809*, 2025.

Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. Do large language models latently perform multi-hop reasoning? *arXiv preprint arXiv:2402.16837*, 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 preprint arXiv:2508.15487*, 2025.

Zebin You, Shen Nie, Xiaolu Zhang, Jun Hu, Jun Zhou, Zhiwu Lu, Ji-Rong Wen, and Chongxuan Li. Llada-v: Large language diffusion models with visual instruction tuning. *arXiv preprint arXiv:2505.16933*, 2025.

Peiyu Yu, Sirui Xie, Xiaojian Ma, Baoxiong Jia, Bo Pang, Ruiqi Gao, Yixin Zhu, Song-Chun Zhu, and Ying Nian Wu. Latent diffusion energy-based model for interpretable text modeling. *arXiv preprint arXiv:2206.05895*, 2022.

Zhen Zhang, Xuehai He, Weixiang Yan, Ao Shen, Chenyang Zhao, Shuohang Wang, Yelong Shen, and Xin Eric Wang. Soft thinking: Unlocking the reasoning potential of llms in continuous concept space. *arXiv preprint arXiv:2505.15778*, 2025.

Lin Zheng, Jianbo Yuan, Lei Yu, and Lingpeng Kong. A reparameterized discrete diffusion model for text generation. *arXiv preprint arXiv:2302.05737*, 2023.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*, 2022.

AI USE STATEMENT

Generative AI tools were used to polish the writing of this paper, assist in checking the correctness of experimental implementations, and help search for relevant literature. They were not used to develop the research ideas or methods, or draw conclusions from the results.

A APPENDIX

A.1 LIMITATIONS

Our work focuses on improving reasoning in diffusion large language models, and the empirical evaluation is therefore limited to existing diffusion LLM architectures and reasoning-oriented benchmarks. While Superposition Inference is model-agnostic at the algorithmic level, its practical benefit depends on the quality and calibration of the model's token distribution during denoising; weaker or poorly calibrated models may require larger superposition capacity or more conservative decoding schedules.

Besides, our experiments primarily evaluate mathematical, scientific, and multi-domain reasoning benchmarks. Extending the approach to broader open-ended generation, long-context tasks, multilingual settings, and safety-critical applications remains an important direction for future work.

A.2 BROADER IMPACT

This work aims to improve the reasoning reliability of diffusion large language models by reducing premature discrete commitment during denoising. Potential positive impacts include more robust reasoning in non-autoregressive language models, improved efficiency–accuracy trade-offs for parallel decoding, and better understanding of how continuous intermediate representations can support language-model reasoning.

At the same time, stronger reasoning capabilities in language models may also amplify existing risks associated with generative AI systems, including the generation of misleading content, overreliance on model outputs, or misuse in automated decision-making pipelines. Our method does not introduce a new dataset, collect private data, or target any specific high-risk application, but it may improve the capabilities of models that require careful deployment.

We recommend that future applications of this approach follow standard responsible deployment practices, including task-specific safety evaluation, monitoring for harmful or misleading outputs, clear communication of model limitations, and appropriate human oversight in high-stakes settings.

A.3 Visualization of Superposition Inference

Figure 6 shows a reasoning excerpt in its original token order, exposing how parallel denoising updates preserve alternative candidates at different positions.

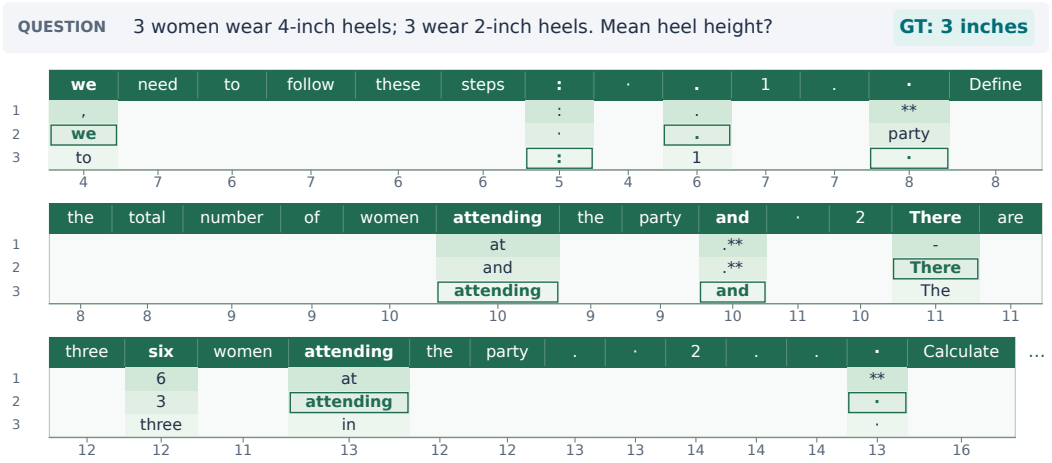


Figure 6: **Visualization of superposition inference.** Dark green rows preserve the annotated excerpt in sequence order; bold tokens indicate annotations below top-1. At these positions, the grid displays candidate ranks 1–3, with outlined cells matching the reading row. Bottom ticks indicate denoising steps; repeated values mark parallel updates. Shading encodes rank, not probability. The top-3 display does not change the underlying top-4 mixtures, and annotations are not discrete selections. GT is computed from the displayed question.

A.4 Failure Case of Standard Discrete Decoding

Figure 7 shows a standard discrete-decoding trace for the same problem as Figure 6. The trace contains inconsistent counting and algebra, followed by an incorrect final answer. The original wording and equations are preserved to expose these failures.

QUESTION 3 women wear 4-inch heels; 3 wear 2-inch heels. Mean heel height? **GT: 3 inches**

To determine how many of the three women at the party wear heels in inches , let follow these logical steps :

1. Define the variables

Let (x) be the number of inches heels worn by three of the women .

The remaining three women wearing are 2 - inch heels .

2. Calculate the total height of heels

The total height of heels for the 2 women can is (x + x + 2 x) = 3 x

The total height of heels for the 3 women 2 - inch heels is (3 times 2 = 6) inches .

Therefore , the total height of heels for all women is (x + x + 6) = 3 x + 6 inches .

3. Set up the equation for the average

The average height of heels is all women is

(3 x + 6) / 3 = 3

So , the equation becomes :

3 x + 6 = 3

4. Solve for (x)

Subtract from both sides :

6 = 6 x

Divide both sides by 6 :

x = 1

Therefore , the of the women wear boxed { 1 } inches heels .

Figure 7: **A failure case of standard discrete diffusion decoding.** The output is organized by its four original reasoning stages. Amber spans retain the source annotations of premature commitments; red spans retain its annotations of erroneous or incoherent text. Light gray cells represent blank padding, and dark red marks the final prediction of 1 inch, which differs from the ground truth of 3 inches computed from the displayed question. Incompatible expressions, including $x + x + 2x = 3x$ and the transition from $(3x + 6)/3 = 3$ to $3x + 6 = 3$, reveal the loss of consistency within the trace.