
Reinforcing the Diffusion Chain of Lateral Thought with Diffusion Language Models

Zemin Huang* Zhiyang Chen* Zijun Wang Tiancheng Li Guo-Jun Qi†

MAPLE Lab, Westlake University

{huangzemin, chenzhiyang, wangzijun63, litiancheng}@westlake.edu.cn, guojunq@gmail.com

Abstract

We introduce the *Diffusion Chain of Lateral Thought (DCoLT)*, a reasoning framework for diffusion language models. DCoLT treats each intermediate step in the reverse diffusion process as a latent "thinking" action and optimizes the entire reasoning trajectory to maximize the reward on the correctness of the final answer with outcome-based Reinforcement Learning (RL). Unlike traditional Chain-of-Thought (CoT) methods that follow a causal, linear thinking process, DCoLT allows bidirectional, non-linear reasoning with no strict rule on grammatical correctness amid its intermediate steps of thought. We implement DCoLT on two representative Diffusion Language Models (DLMs). First, we choose SEDD as a representative continuous-time discrete diffusion model, where its concrete score derives a probabilistic policy to maximize the RL reward over the entire sequence of intermediate diffusion steps. We further consider the discrete-time masked diffusion language model – LLaDA, and find that the order to predict and unmask tokens plays an essential role to optimize its RL action resulting from the ranking-based Unmasking Policy Module (UPM) defined by the Plackett-Luce model. Experiments on both math and code generation tasks show that using only public data and 16 H800 GPUs, DCoLT-reinforced DLMs outperform other DLMs trained by SFT or RL or even both. Notably, DCoLT-reinforced LLaDA boosts its reasoning accuracy by +9.8%, +5.7%, +11.4%, +19.5% on GSM8K, MATH, MBPP, and HumanEval.

1 Introduction

To enable complex reasoning, most large language models (LLMs) [15, 20] learn to decompose problems into simpler sub-steps and generate intermediate reasoning in natural language. Chain-of-Thought (CoT) [39] first reveals that step-by-step reasoning facilitates language models, as the outputs from previous steps could be rationales for more accurate next step prediction. Based on that, OpenAI’s PRM [23] supervises these intermediate reasoning steps with progressive rewards, to ensure the correctness of each single step. More recently, DeepSeek-R1 [15] eliminates the need of the reward model and verifies only the correctness of the final answer, relaxing the constraints on the reasoning process. However, due to the causal nature of attention mechanisms, auto-regressive models are still forced to reason in a single, sequential direction.

However, when developing ideas, human cognition does not always proceed through strictly sequential steps. At the beginning of thinking, human does not require an intact linguistic structure. Concepts, words, or ideas emerge spontaneously and independently first, and are gradually refined and organized over time to follow grammar rules. This non-linear and creative mode of reasoning, known as *lateral thinking* [17], contrasts with the structured, step-by-step approach of vertical thinking.

*Equal contribution.

†Corresponding author.

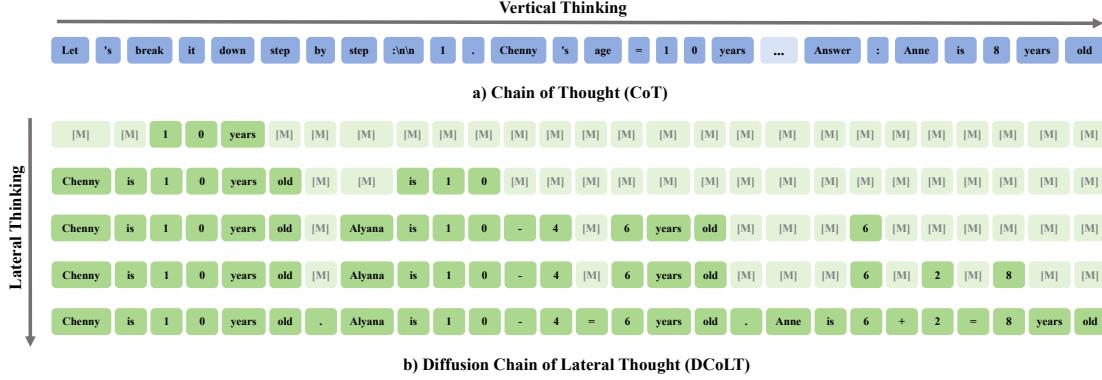


Figure 1: Comparison between CoT and DCoLT. (a) A typical CoT performs vertical thinking by following an auto-regressive convention that generates responses token by token from left to right in a linear way. (b) DCoLT performs lateral thinking that generates the responses in a non-linear way without following the auto-regressive order; moreover, at each step, it can generate multiple tokens at chosen positions. We focus on the lateral thinking in this paper by reinforcing the chain of such lateral thought *as an entirety* in Diffusion Language Models (DLMs).

Contrary to auto-regressive models, Diffusion Language Models (DLMs) [24, 27, 42] have also been adopted for text generation. The intermediate steps of the reverse diffusion process are naturally well-suited to emulate lateral thinking. Unlike auto-regressive models, diffusion models generate all tokens in parallel from a prior distribution. Each token can attend freely to all others under a non-causal mask in self attention, and intermediate reasoning steps are not required to conform to grammatical rules during multi-step generation, thus leading to more divergent thinking. In this paper, we propose the *Diffusion Chain of Lateral Thought (DCoLT)* to reinforce the lateral reasoning in diffusion language models, as shown in Fig. 1. Rather than providing explicit supervision for the thinking process, we employ outcome-based reinforcement learning, offering a rule-based reward that evaluates only the correctness of the final responses. This reward encourages the model to explore diverse, creative, and non-linear thought trajectories that ultimately lead to correct answers.

We study two paradigms of diffusion language models to reinforce the DCoLT, continuous-time diffusion language models [24, 6] and discrete-time diffusion language models [3, 19, 27, 42]. For the continuous-time paradigm, we consider SEDD [24] as a representative DLM. SEDD predicts the concrete score, allowing for a closed-form expression of the predicted diffused distribution of generated tokens at each step. This distribution can be viewed as a probabilistic policy for sampling tokens, which can be trained to optimize the reward on the final answers by reinforcing DCoLT.

For the discrete-time paradigm, we consider a masked DLM – LLaDA [27]. Besides the output distributions over discrete tokens, we note that the unmasking order plays an important role to decide which tokens ought to be kept to form the current step of lateral thought, and thus should be part of the learnable reasoning process. To this end, we introduce a Plackett-Luce model [32, 28] to define a ranking-based unmasking policy, where each masked token is assigned a predicted ranking score, and the unmasking policy selects the top- K ranked tokens to retain in the output sequence at each diffusion step. The unmasking policy is trained together with the token generation policy to optimize the reward on the final answers.

We conduct experiments on both Math and code generation tasks to demonstrate the efficacy of DCoLT. After training the DCoLT on the SEDD 400M model, it achieves 96.2% and 57.0% in accuracy on Sudoku 4×4 and GSM8K-Aug [10] tasks. On the LLaDA 8B model, DCoLT achieves the state-of-the-art performance among existing DLMs trained with SFT or RL or even both. Using only public data and 16 H800 GPUs, the DCoLT-reinforced LLaDA model achieves 88.1% on GSM8K [8], 44.6% on MATH [16], 51.6% on MBPP [4] and 59.1% on HumanEval [7] in the challenging zero-shot setting. Even compared with auto-regressive models that are trained with significantly more proprietary data and fully annotated CoT reasoning processes, it still demonstrates competitive performances.

2 Related Works

Diffusion Language Models Diffusion models have achieved impressive results in image modeling [18, 37, 33, 21, 43], with recent efforts [24, 11, 13, 3, 48, 2, 47, 12, 29, 27, 42] extending to language tasks. Depending on the types of diffused distributions over text tokens, diffusion language models can be categorized into continuous diffusion models [13] and discrete diffusion models [3, 24, 11], with our work focusing primarily on the latter due to the discrete nature of language.

Among discrete diffusion models, masked diffusion models [27, 42, 24] emerge as a promising approach due to their superior performances. Recent advances simplify model training and design by adopting the straight cross-entropy loss [48, 29] and removing the time embeddings [47]. The resultant models [27, 42] achieve the state-of-the-art performance among existing diffusion language models.

Particularly, we consider two representative examples of discrete diffusion language models in this paper: SEDD [24] – a discrete diffusion model with a continuous diffusion time, and LLaDA [27] – a masked diffusion model with discrete diffusion timesteps. We will reinforce the entire sequence of their reverse diffusion processes containing non-linear text generations to release their lateral thinking ability. In contrast, DoT (Diffusion of Thought) [41] uses annotated step-by-step CoT data for *supervised fine-tuning* with existing diffusion losses such as the score entropy loss [24] or the noise prediction loss [13]. Thus, it still encourages a conventional stepwise reasoning process as in *vertical thinking* [17].

Reinforcement Learning for Language Models Reinforcement Learning (RL) helps language models better align with human preference [30] or verifiable knowledge [22] (e.g. rewards on math/code tasks). Recently, there appear various algorithms for RL optimization [35, 1, 36, 25]. These methods enhance model’s ability to produce aligned outputs and show strong potential in inference-time scaling for challenging problems [15, 20].

However, existing RL approaches are primarily applied to auto-regressive language models [30, 15]. For DLMs, previous works explored policy gradient methods within the concrete score-matching framework [45], or estimated the log-probability of the sampled responses to perform reinforcement learning [46]. However, both models merely focused on the generated text response in the final diffusion step as an action to optimize, ignoring the role of intermediate diffusion steps as lateral thought to reinforce amid the reverse diffusion process.

Instead, in this paper, we seek to explicitly reinforce the reverse diffusion process to enable the training of lateral reasoning. We demonstrate in experiments that even though these intermediate diffusion steps perform non-linear rather than step-by-step reasoning, reinforcing them can eventually lead to correct answers at the final step that outperforms many CoT models [39, 10].

3 Methods

3.1 Diffusion Chain of Lateral Thought

To model the probability distribution p_{data} over a finite vocabulary $\mathcal{V} = \{1, \dots, V\}$ for text generation tasks, discrete diffusion processes model how the unknown data distribution p_{data} at $t = 0$ gradually evolves into a prior distribution p_{prior} at $t = T$ [6]. The distribution at an intermediate diffusion time t is denoted by p_t .

Then a generation process is realized by reversing this diffusion process. Practically, it begins by sampling x_0 from the prior p_{prior} and then iteratively denoising through a sequence of diffusion time $t_{0:N}$. At each step n , the model θ estimates the diffused distribution p_{θ, t_n} at time t_n , from which an intermediate sample x_n is drawn, as shown in Eq. 1. The diffusion time t_n decreases progressively as n increases. After completing all denoising steps, reaching $t_N = 0$, the final response x_N is obtained.

$$x_n \sim p_{\theta, t_n}(\cdot | x_{n-1}) \quad (1)$$

This naturally generates a series of intermediate results $x_{0:N-1}$ before arriving at the final output x_N . This behavior is analogous to the Chain-of-Thought (CoT) technique [39]. However, unlike the vertical and causal structure of CoT, the diffusion process enables the model to produce intermediate

content that facilitates reaching final answers, aligning more closely with the concept of lateral thinking—solving problems through indirect and creative approaches. Thus, we define the *entire* sequence of all intermediate steps $x_{0:N}$ as the *Diffusion Chain of Lateral Thought (DCoLT)*, as shown in Fig. 1, and seek to apply reinforcement learning to train it based on the given reward function.

Compared to standard CoT, DCoLT is distinguished with several notable features:

- **Bidirectional Reasoning:** CoT generates tokens sequentially in a causal, irreversible manner; once derived, earlier thought remains fixed in the context. Instead, DCoLT adopts a bidirectional structure: each token is influenced by both preceding and succeeding content with bidirectional self-attention masks, allowing global refinement throughout generations.
- **Format-Free Reasoning:** CoT typically adheres strictly to natural language format with complete linguistic structures. DCoLT, however, relaxes this constraint, allowing early stage of intermediate steps not necessarily to be complete or correct in format, thus enabling more divergent and creative patterns of thought before finally converging to complete responses.
- **Nonlinear Generation:** CoT generates tokens one-by-one linearly in an auto-regressive manner from left to right. In contrast, DCoLT allows nonlinear generation of tokens at various positions. This aligns with how human develops ideas by beginning with keywords or critical points and then refining the details around them.

Fig. 7–Fig. 9 in Appendix B show some examples of intermediate responses from DCoLT-reinforced DLMs, demonstrating the above features with bidirectional and nonlinear generations throughout reverse diffusion processes from incomplete masked phases to complete unmasked responses.

To enhance the lateral thinking, we adopt final-outcome rewarded reinforcement learning to train the model, promoting reasoning trajectories that can lead to correct final responses x_N . For this, we treat the generation of $x_{1:N}$ as a sequence of actions to optimize *as a whole*. At each denoising step n , a distribution $\pi_{\theta,n}(\cdot|x_{n-1})$ over possible outputs is defined, which serves as the policy for sampling x_n to calculate action probabilities during RL training. The reward r is assigned based on verifiable correctness of final results. We do not impose any explicit supervision on the intermediate steps, promoting the model’s lateral thinking ability to explore diverse reasoning strategies to maximize the reward. This could yield useful patterns of thinking processes – for example, as shown in Fig. 3 of Appendix A, an easy-to-hard progressive generation of responses emerges from the DCoLT-trained DLM, in contrast to its counterpart supervisedly trained with the conventional diffusion loss.

We summarize the algorithm in Alg. 1. Details about training specific diffusion language models such as SEDD [24] and LLaDA [27] will be elaborated on in the following two subsections. Considering the demanding computation graph that expands multi-step generations in memory, we back-propagate the gradients at each step and accumulate them after the whole reverse diffusion process (c.f. Line 26 and 28 in Alg. 1). We use the GRPO [36] for fair comparisons with other RL-based models [36, 46], while alternative RL training approaches such as PPO [35] and RLOO [1] can also be adopted.

3.2 A Continuous-Time DLM Case: DCoLT-reinforced SEDD

To define a discrete diffusion process, continuous-time diffusion language models such as SEDD [24] evolve a family of distributions p_t according to a continuous-time Markov process, which can be represented by the following linear ordinary differential equation.

$$\frac{dp_t}{dt} = Q_t p_t, \quad p_0 = p_{\text{data}}, \quad p_T = p_{\text{prior}} \quad (2)$$

Here, Q_t is the transition rate matrix defining the forward process. We may first consider the simple single-token case ($Q_t \in \mathbb{R}^{V \times V}$). Conversely, to generate a sample from p_{prior} , this process has a corresponding reverse process, defined by a reverse transition rate matrix \bar{Q}_t :

$$\frac{dp_{T-t}}{dt} = \bar{Q}_{T-t} p_{T-t}, \quad \bar{Q}_t(y, x) = \frac{p_t(y)}{p_t(x)} Q_t(x, y), \quad \bar{Q}_t(x, x) = - \sum_{y \neq x} \bar{Q}_t(y, x) \quad (3)$$

By Euler’s method, we have the transition probability p_{t_n} at each diffusion step t_n , resulting in an iterative formula for multi-step generations from $t_0 = T$ to $t_N = 0$ with $n = 0, \dots, N$,

$$p_{t_n}(x_n = y | x_{n-1} = x) = \delta_{xy} + (t_{n-1} - t_n) \bar{Q}_{t_{n-1}}(y, x) \quad (4)$$

Algorithm 1 A General Framework for Training DCoLT

Require: Model parameters θ , a dataset \mathcal{D} , and reward_func.

- 1: **while** θ not converged and maximum epochs not reached **do**
- 2: Sample questions $q \sim \mathcal{D}$
- 3: **for** $g = 1$ to G **do** ▷ Generate a group of G trajectories
- 4: Initialize x_0^g with q and mask tokens.
- 5: **for** $n = 1$ to N **do** ▷ N denotes the number of denoising steps
- 6: **if** training SEDD **then**
- 7: Sample $x_n^g \sim p_{\theta, t_n}(\cdot | x_{n-1}^g)$
- 8: **else if** training LLaDA **then**
- 9: Calculate the ranking score $h_{\theta, n}$ for each token
- 10: Sample K tokens to unmask in this step: $\mathcal{U}_n \sim \text{Plackett-Luce}(h_{\theta, n}, K)$
- 11: Sample $x_n^{g,i} \sim p_{\theta, n}^i(\cdot | x_{n-1}^g)$, $\forall i \in \mathcal{U}_n$
- 12: **end if**
- 13: **end for**
- 14: $r^g = \text{reward_func}(q, x_N^g)$ ▷ Compute the rewards
- 15: **end for**
- 16: **for** $g = 1$ to G **do** ▷ Compute the advantages
- 17: $A^g = \frac{r^g - \text{mean}(r^{1:G})}{\text{std}(r^{1:G})}$
- 18: **end for**
- 19: **for** $n = 1$ to N **do** ▷ Compute π_θ and losses for each denoising step
- 20: **if** training SEDD **then**
- 21: $\pi_{\theta, n}(x_n^g | x_{n-1}^g) = \prod_{i=1}^{|x_n^g|} p_{\theta, t_n}(x_n^{g,i} | x_{n-1}^g)$ ▷ see Eq. 5
- 22: **else if** training LLaDA **then**
- 23: $\pi_{\theta, n}(x_n^g | x_{n-1}^g) = \pi_{\theta, n}^{\text{unmask}}(\mathcal{U}_n^g | x_n) \cdot \pi_{\theta, n}^{\text{token}}(x_n | x_{n-1}, \mathcal{U}_n)$ ▷ see Eq. 9
- 24: **end if**
- 25: $\mathcal{L}_{\theta, n} = -\frac{1}{G} \sum_{g=1}^G \frac{\pi_{\theta, n}(x_n^g | x_{n-1}^g)}{\pi_{\text{old}, n}(x_n^g | x_{n-1}^g)} A^g$
- 26: Calculate the gradient $\nabla_{\theta} \mathcal{L}_{\theta, n}$
- 27: **end for**
- 28: Update θ with accumulated gradients $\sum_{n=1}^N \nabla_{\theta} \mathcal{L}_{\theta, n}$ along the descent direction
- 29: **end while**

with x_n denoting x_{t_n} to avoid notational clutter, i.e., the diffused sample x_t at $t = t_n$.

Specifically, we consider a representative discrete diffusion model, SEDD [24]. This model learns to approximate the concrete score, i.e. $s_{\theta}(x, t)_y \approx \frac{p_t(y)}{p_t(x)}$ for any $y \neq x$ to represent the probability to transfer to other tokens. Thus, we may replace $\tilde{Q}_{t_{n-1}}(y, x)$ with the model-estimated $s_{\theta}(x, t_{n-1})_y \cdot Q_{t_{n-1}}(x, y)$ in Eq. 4, making the transition probability tractable.

When scaling to sequences, we may apply τ -leaping to define the probability of action at each timestep as the product of transition probabilities across all tokens:

$$\pi_{\theta, n}(x_n | x_{n-1}) = \prod_{i=1}^{|x_n|} p_{\theta, t_n}(x_n^i | x_{n-1}) \quad (5)$$

$$p_{\theta, t_n}(x_n^i | x_{n-1}) = \begin{cases} s_{\theta}(x_{n-1}, t_{n-1})_{i, x_n^i} \cdot (t_{n-1} - t_n) \cdot Q_{t_{n-1}}(x_{n-1}^i, x_n^i), & x_{n-1}^i \neq x_n^i, \\ 1 - \sum_{y \neq x_n^i} s_{\theta}(x_{n-1}, t_{n-1})_{i, y} \cdot (t_{n-1} - t_n) \cdot Q_{t_{n-1}}(x_{n-1}^i, y), & x_{n-1}^i = x_n^i, \end{cases} \quad (6)$$

where $|x_n|$ denotes the token length of the sequence x_n . Note that the predicted p_{t_n} is dependent on the model parameters θ . During training, we may update θ to reinforce or suppress the probability of all actions along a trajectory of thoughts $[x_0, x_1, \dots, x_N]$, according to the reward associated with each completion. Unlike traditional diffusion model training, which typically optimizes each individual denoising step independently with ground truth responses, DCoLT jointly optimizes the entire reverse process of thought from $t_0 = T$ to $t_N = 0$, promoting the emergence of lateral thought leading to correct answers.

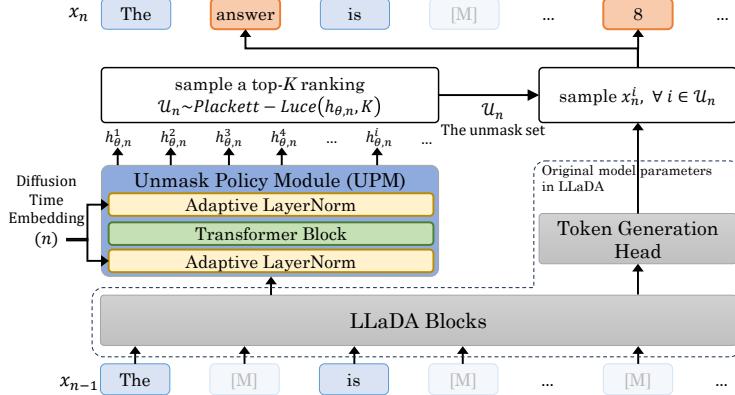


Figure 2: The structure of LLaDOU. It first predicts the token set to unmask \mathcal{U}_n according to the ranking score $h_{\theta,n}$ by the UPM, and then samples those unmasked tokens in \mathcal{U}_n by LLaDA blocks.

3.3 A Discrete-Time DLM Case: DCoLT-reinforced LLaDA

Some other diffusion language models instead operate at discrete timesteps [27, 42]. They can be viewed as Discrete-Time Diffusion Language Models with a multi-step generation process. For these models, we will figure out the probability of actions for each discrete step.

Among them is LLaDA [27] which is a typical Discrete-Time Masked Diffusion Language Model. It generates text by progressively unmasking a sequence initially prefilled entirely with mask tokens. At each generation step, the model receives a partially masked sequence as input. Then, it chooses and predicts a subset of masked tokens to reveal as clean text. As it progresses, the number of masked tokens gradually decreases, so that the model ultimately yields a fully generated sequence.

Specifically, we can define the model’s action at each step n in two parts: 1) determining the set \mathcal{U}_n of tokens to unmask, and 2) predicting the values of these tokens to obtain the new sequence x_n over the unmasked part \mathcal{U}_n .

The action of selecting which tokens to unmask at each step can be decided by ranking masked tokens with a score function under the current state. To this end, we introduce an Unmask Policy Module (UPM), which predicts a score value $h_{\theta,n}^i$ for each token i at the current diffusion step n . Based on these scores, we define a policy to sample a top- K ranked list $\mathcal{U}_n = [u_n(1), \dots, u_n(K)]$ from a Plackett–Luce model [32, 28]: a multinomial distribution is formed from the predicted scores, and K tokens are sequentially sampled without replacement, such that the corresponding scores are non-increasingly ordered $h_{\theta,n}^{u_n(1)} \geq \dots \geq h_{\theta,n}^{u_n(K)}$ with high probability.

Formally, let \mathcal{M}_n denote the set of tokens that remain masked after the n -th step, i.e., $\mathcal{M}_{n-1} \setminus \mathcal{U}_n$. Then, the probability of sampling a specific ranking list \mathcal{U}_n is given by:

$$\pi_{\theta,n}^{\text{unmask}}(\mathcal{U}_n | x_{n-1}) = \prod_{k=1}^K \frac{\exp(h_{\theta,n}^{u_n(k)})}{\sum_{j=k}^K \exp(h_{\theta,n}^{u_n(j)}) + \sum_{j \in \mathcal{M}_n} \exp(h_{\theta,n}^{u_n(j)})} \quad (7)$$

Specifically, UPM takes the hidden states from the last hidden layer as inputs at each denoising step n , and predicts a ranking score $h_{\theta,n}^i$ for i -th token. It contains only one transformer block, introducing marginal computation cost. Moreover, the step index n and the mask indicator of each token in x_{n-1} can benefit the model training. Thus, we embed both of them into the Unmask Policy Module (UPM) via adaptive layer normalization. The module structure is shown in Fig. 2. For convenience, we name the DCoLT-trained LLaDA with UPM by **LLaDOU** (LLaDA with Ordered Unmasking).

As shown in Fig. 10 of Appendix B, the learned scores $h_{\theta,n}$ can be viewed as the ranked confidences over the token predictions at the current step n . Higher scores indicate that the UPM predicts lower levels of diffusion noises may be present in the generated tokens, which are less likely to contain errors and thus could be unmasked with higher priorities at the current step. Fig. 7 shows the resulting unmasking orders from the learned ranking scores for some prompts.

Table 1: Experimental Results on SEDD. All models listed below are of medium size, with approximately 400M parameters. On the Sudoku 4×4 dataset, we report the accuracy on a test set of 2,000 samples each with 9 empty cells. On GSM8K-Aug, we use the test set from GSM8K[8] for evaluation following [10].

Models	Post-Training	Sudoku 4×4	GSM8K-Aug
GPT2 + CoT[10]	SFT	71.5	43.9
GPT2 + Implicit CoT[10]	SFT	-	21.9
GPT2 + CoT	RL	74.6	-
SEDD + DoT [41]	SFT	79.4	53.5
SEDD + DCoLT	RL	96.2	57.0

Once the unmask set \mathcal{U}_n is determined, the model predicts their token values based on the output distribution by LLaDA blocks over the vocabulary. Viewing this prediction as a second-stage action, the probability of generating x_n given x_{n-1} and \mathcal{U}_n can be written as

$$\pi_{\theta,n}^{\text{token}}(x_n|x_{n-1}, \mathcal{U}_n) = \prod_{i \in \mathcal{U}_n} p_{\theta,n}(x_n^i|x_{n-1}). \quad (8)$$

Thus, the probability of the complete policy for transitioning from x_{n-1} to x_n is the product of those of the unmask policy and the token prediction policy as used in Line 23 of Alg. 1:

$$\pi_{\theta,n}(x_n|x_{n-1}) = \pi_{\theta,n}^{\text{unmask}}(\mathcal{U}_n|x_n) \cdot \pi_{\theta,n}^{\text{token}}(x_n|x_{n-1}, \mathcal{U}_n). \quad (9)$$

4 Experiments

We conduct DCoLT experiments on two DLMs – SEDD 400M and LLaDA 8B, each representing one of the two categories mentioned in Sec. 3.

4.1 Experiments on SEDD+DCoLT

4.1.1 Settings

We first conduct experiments on SEDD [24] for two different tasks: 1) Sudoku 4×4 task and 2) GSM8K-Aug math problems [10], both assessing math reasoning performance. These two tasks allow us to perform direct comparisons with existing reasoning algorithms, including both supervised fine-tuning [10, 41] and reinforcement learning approaches [36]. For all experiments, we use the rule-based reward function to compute the reward for each sample: a reward of 1 is assigned if the solution is correct, and 0 otherwise. We choose SEDD as the base model to conduct DCoLT training, which is of medium size around 400 M. For detailed settings please refer to Appendix C.1.

4.1.2 Results

From the results in Tab. 1, our method remarkably outperforms both (implicit) CoT [10] and DoT [41] on both tasks. While both CoT and DoT perform vertical thinking, DoT also uses the SEDD as its base model to simulate causal reasoning. In contrast, our approach only uses RL to train the model without any CoT data for supervised training. For CoT, we adopt an auto-regressive language model – GPT2 [31] as the base model, which has the similar model size of about 400M to the SEDD model.

Particularly, SEED + DCoLT achieves an accuracy of 96.2% on Sudoku 4×4 task, the best performance among all the methods. Using the same SEED model, DCoLT greatly exceeds DoT (79.4%), even though the latter introduce step-by-step annotated CoT data for supervised training. For GPT2 model, both CoT and Implicit CoT, no matter if being post-trained with SFT or RL, perform worse than our model clearly. On the GSM8K-Aug dataset, SEED + DCoLT also reaches 57.0% in accuracy, which is notably higher than both DoT and CoT counterparts. We will further analyze the generation process of SEED + DCoLT in Appendix A, demonstrating how tokens are generated in a lateral thinking manner.

Table 2: Model accuracies on math and code generation benchmarks. The "Post-Training" column indicates what kinds of post-training (SFT or RL) phase the model goes through. \dagger indicates the model uses additional proprietary training data for the post-training phase. The numbers in parentheses represent the number of shots for the in-context learning, with “-” indicating unknown cases not mentioned in original papers. The results denoted with * are evaluated with the prompt templates in Appendix C.2 for fair comparison, while others are reported in original papers. The results denoted with \ddagger are evaluated on a subset MATH-500 instead of MATH. We highlight the best-performing model among compared DLMs in **bold**.

Method	Post-Training	Math		Code	
		GSM8K	MATH	HumanEval	MBPP
Diffusion Language Models					
Dream 7B* [42]	baseline	81.1 (0)	42.9 (0)	51.8 (0)	49.6 (0)
LLaDA 8B* [27]	baseline	78.3 (0)	38.9 (0)	39.6 (0)	40.2 (0)
+ SFT [46]	+ SFT	81.1 (0)	34.8 \ddagger (0)	-	-
+ diffu-GRPO [46]	+ RL	81.9 (0)	39.2 \ddagger (0)	-	-
d1-LLaDA (SFT + diffu-GRPO)	+ SFT + RL	82.1 (0)	40.2 \ddagger (0)	-	-
LLaDOU (LLaDA + DCoLT) 8B*	+ RL	88.1 (0)	44.6 (0)	59.1 (0)	51.6 (0)
Auto-regressive Models					
LLaMA2 7B [38]	baseline	14.6 (0)	2.5 (0)	12.8 (0)	20.8 (3)
MetaMath 7B [44]	+ SFT \dagger	66.5 (0)	19.8 (0)	-	-
CodeLLaMA-Instruct 7B [34]	+ SFT \dagger	-	-	34.8 (0)	44.4 (3)
Deepseek 7B [5]	baseline	63.0 (0)	15.8 (0)	48.2 (0)	35.2 (3)
DeepseekMath-Instruct 7B [36]	+ SFT \dagger	82.9 (-)	46.8 (-)	-	-
DeepseekMath-RL 7B [36]	+ SFT \dagger + RL \dagger	88.2 (-)	51.7 (-)	-	-
DeepseekCoder-Instruct 7B [14]	+ SFT \dagger	-	-	66.1 (-)	65.4 (-)

4.2 Experiments on LLaDOU (LLaDA+DCoLT)

4.2.1 Settings

We further apply DCoLT to LLaDA [27], a SOTA discrete-time masked-based DLM with 8B parameters. During generation, it starts with a masked sequence of length 256, and performs a 256-step reverse diffusion process. We adopt a block-wise unmasking strategy as in LLaDA [27], and divide the sequence into blocks of length 8. In Tab. 2, both LLaDA and LLaDOU (LLaDA+DCoLT) are evaluated with this setting for fair comparison. We conduct experiments with 16 H800 GPUs to jointly train both UPM and LLaDA parts. For implementation details, please refer to Appendix C.2.

4.2.2 Results

As shown in Tab. 2, among all compared DLMs, LLaDOU consistently achieves the best performance across all benchmarks. On GSM8K, LLaDOU reaches 88.1%, significantly higher than other methods. On the more challenging MATH dataset, LLaDOU achieves 44.6%, outperforming baseline models such as LLaDA 8B (+5.7%) and Dream 7B (+1.7%), as well as models enhanced with post-training by SFT, RL or both. Notably, d1-LLaDA [46] uses 1K questions paired with detailed reasoning traces [26] in SFT to boost reasoning capabilities. In contrast, LLaDOU achieves superior performance +6.0% on GSM8K and +4.4% on MATH – without any reasoning supervision, relying solely on reward signals based on the final answers.

Even when compared with auto-regressive models trained with a significantly larger amount of proprietary data, LLaDOU remains highly competitive. For example, DeepseekMath [36] gathers 776K questions with reasoning process annotations for SFT and 144K questions for RL, while LLaDOU is trained on just 15K public training samples from GSM8K and MATH. Despite using nearly two orders of magnitude fewer data for RL training, LLaDOU matches DeepseekMath’s performance on GSM8K (88.1%), highlighting its data efficiency and the effectiveness of reward-driven reasoning via diffusion-based lateral reasoning. In Appendix B, we analyze the thinking process generated by LLaDOU.

Similar trends are observed on code generation benchmarks. Unlike models that rely on valuable ground-truth code for supervised training, LLaDOU requires no code for supervised training at all. Instead, it only needs several test cases to provide a simple outcome-based reward for reinforcement

Table 4: Model accuracies with extended generation lengths on MATH without further RL-finetuning the model on these longer generations. The results denoted with * are taken from [46] and evaluated on a subset MATH-500 instead of MATH.

Model	Generation Length				
	256	320	384	448	512
LLaDA 8B [27]	38.9	40.1	41.5	42.3	42.5
+ SFT* [46]	38.6	-	-	-	40.2
+ diffu-GRPO* [46]	37.2	-	-	-	39.2
d1-LLaDA (SFT+diffu-GRPO)* [46]	38.6	-	-	-	40.2
LLaDA + DCoLT (LLaDOU) 8B	44.6	45.7	46.2	45.7	45.9

learning: the model is rewarded only if the generated code passes all unit tests. Despite this code-free reward signal, LLaDOU achieves 51.6% on MBPP and 59.1% on HumanEval, outperforming other DLMs and auto-regressive models, except for DeepseekCoder [14] that benefits from 2B tokens of high-quality instruction-tuning data—approximately 150× more than we used (about 12M tokens). Details of post-training datasets used by all compared methods are provided in Appendix C.3.

4.2.3 Ablation Studies

We provide some ablation studies to reveal the role of some model design and hyper-parameter settings in LLaDOU. Unless otherwise specified, all experiments are conducted on LLaDOU 8B with $N = 64$ diffusion steps, and the model is trained for 150 iterations with a batch size of 32.

Better performance with the UPM In Tab. 3, we ablate the model design in LLaDOU. Even if we freeze the original model parameters in LLaDA part and only train the UPM part, it achieves a significant improvement from 47.27% to 69.24%. This result indicates that the unmasking policy plays a crucial role in our model. By training the LLaDA part with RL together, the accuracy further improves to 81.06%.

We use the adaptive layernorm in the UPM. Removing it degrades accuracy to 80.53%. This demonstrates that incorporating the embeddings of diffusion step n and the mask indicators also benefits the training of the unmasking policy.

Extension to longer generation length Though the models in Tab. 2 is trained to generate completions with a fixed generation length 256 (i.e., the length of initialized mask tokens), it can benefit from generating longer sequences without further RL training on longer generations. This is especially effective when answering difficult questions on MATH dataset, in which solving problems sometimes require longer reasoning. Tab. 4 reports the accuracies with different generation lengths. It suggests that longer generations in inference lead to better accuracies. For example, a generation length of 384 achieves an accuracy of 46.22%, which is +1.62% higher than the default length of 256. This improvement vanishes as we further increase the generation length. However, if we further tune the model on a longer generation length, for example $L = 512$, the accuracy further improves to 47.3%. This suggests the potential of how scaled generation lengths may improve the performance of DCoLT-trained DLMs, as shown in Fig. 12 of Appendix D.

Table 3: Ablation on the Unmasking Policy Module

Trained parameters		GSM8K Acc.
UPM	LLaDA	
×	×	47.27 (Baseline)
w/ AdaLN	×	69.24
w/o AdaLN	✓	80.53
w/ AdaLN	✓	81.06

5 Conclusion

In this paper, we propose a new paradigm of model reasoning ability, Diffusion Chain of Lateral Thought (DCoLT), for diffusion language models. DCoLT considers the intermediate steps in the reverse diffusion process as the thinking actions, and optimizes this process with final-outcome rewarded Reinforcement Learning (RL). We implement DCoLT on two representative diffusion language models, SEDD and LLaDA. On SEDD, we derive the corresponding RL optimization over the distribution chain of thinking actions via the predicted concrete scores at denoising steps. On the masked-based diffusion language model LLaDA, we regard the order in which tokens are unmasked at each diffusion step as its action, and propose an Unmask Policy Module (UPM) to

optimize over the order distribution with the Plackett-Luce model. The experiments indicate that DCoLT outperforms other SFT- or RL-based training algorithms, and demonstrate its effectiveness on a wide range of tasks and benchmarks, including math and code generation.

Limitations First, due to limited training data and compute, our model’s performances on Math and code generation tasks still have much rooms to improve. Prior works [36, 44, 14, 34] demonstrate that proprietary data often significantly improves the model performances. Also, scaling the token length of sequences during training can also enhance reasoning. We will seek to boost the model ability along these directions when computing resources become available. Second, currently DCoLT is only validated on tasks having a verifiable reward function. We may need a reward model to cover more general tasks. We will develop these further in our future research.

References

- [1] Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms. *arXiv preprint arXiv:2402.14740*, 2024.
- [2] Marianne Arriola, Aaron Gokaslan, Justin T Chiu, Zhihan Yang, Zhixuan Qi, Jiaqi Han, Subham Sekhar Sahoo, and Volodymyr Kuleshov. Block diffusion: Interpolating between autoregressive and diffusion language models. *arXiv preprint arXiv:2503.09573*, 2025.
- [3] 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.
- [4] Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.
- [5] Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024.
- [6] Andrew Campbell, Joe Benton, Valentin De Bortoli, Thomas Rainforth, George Deligiannidis, and Arnaud Doucet. A continuous time framework for discrete denoising models. *Advances in Neural Information Processing Systems*, 35:28266–28279, 2022.
- [7] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.
- [8] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems, 2021. URL <https://arxiv.org/abs/2110.14168>, 9, 2021.
- [9] DeepSeek-AI. Deepseek-v3 technical report, 2024. URL <https://arxiv.org/abs/2412.19437>.
- [10] Yuntian Deng, Kiran Prasad, Roland Fernandez, Paul Smolensky, Vishrav Chaudhary, and Stuart Shieber. Implicit chain of thought reasoning via knowledge distillation. *arXiv preprint arXiv:2311.01460*, 2023.
- [11] Itai Gat, Tal Remez, Neta Shaul, Felix Kreuk, Ricky TQ Chen, Gabriel Synnaeve, Yossi Adi, and Yaron Lipman. Discrete flow matching. *Advances in Neural Information Processing Systems*, 37:133345–133385, 2024.
- [12] Shansan Gong, Shivam Agarwal, Yizhe Zhang, Jiacheng Ye, Lin Zheng, Mukai Li, Chenxin An, Peilin Zhao, Wei Bi, Jiawei Han, et al. Scaling diffusion language models via adaptation from autoregressive models. *arXiv preprint arXiv:2410.17891*, 2024.
- [13] Ishaan Gulrajani and Tatsunori B Hashimoto. Likelihood-based diffusion language models. *Advances in Neural Information Processing Systems*, 36:16693–16715, 2023.

- [14] Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Yu Wu, YK Li, et al. Deepseek-coder: When the large language model meets programming—the rise of code intelligence. *arXiv preprint arXiv:2401.14196*, 2024.
- [15] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- [16] 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.
- [17] James S Hernandez and MBBS Prathibha Varkey. Vertical versus lateral thinking. *Physician executive*, 34(3):26, 2008.
- [18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- [19] Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion: Learning categorical distributions. *Advances in neural information processing systems*, 34:12454–12465, 2021.
- [20] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.
- [21] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.
- [22] Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, et al. Training language models to self-correct via reinforcement learning. *arXiv preprint arXiv:2409.12917*, 2024.
- [23] Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let’s verify step by step. In *The Twelfth International Conference on Learning Representations*, 2023.
- [24] Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In *Forty-first International Conference on Machine Learning*, 2024.
- [25] Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37:124198–124235, 2024.
- [26] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsumori Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025.
- [27] 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.
- [28] Shuzi Niu, Yanyan Lan, Jiafeng Guo, and Xueqi Cheng. A new probabilistic model for top-k ranking problem. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 2519–2522, 2012.
- [29] 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.

- [30] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.
- [31] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.
- [32] Stephen Ragain and Johan Ugander. Choosing to rank. *arXiv preprint arXiv:1809.05139*, 2018.
- [33] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
- [34] Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, et al. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*, 2023.
- [35] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- [36] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
- [37] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.
- [38] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwala Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- [39] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
- [40] Zhangchen Xu, Yang Liu, Yueqin Yin, Mingyuan Zhou, and Radha Poovendran. Kodcode: A diverse, challenging, and verifiable synthetic dataset for coding. *arXiv preprint arXiv:2503.02951*, 2025.
- [41] Jiacheng Ye, Shansan Gong, Liheng Chen, Lin Zheng, Jiahui Gao, Han Shi, Chuan Wu, Xin Jiang, Zhenguo Li, Wei Bi, et al. Diffusion of thought: Chain-of-thought reasoning in diffusion language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
- [42] Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7b, 2025. URL <https://hkunlp.github.io/blog/2025/dream>.
- [43] Zilyu Ye, Zhiyang Chen, Tiancheng Li, Zemin Huang, Weijian Luo, and Guo-Jun Qi. Schedule on the fly: Diffusion time prediction for faster and better image generation, June 2025. URL <https://arxiv.org/abs/2412.01243>.
- [44] Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*, 2023.
- [45] Oussama Zekri and Nicolas Boullé. Fine-tuning discrete diffusion models with policy gradient methods. *arXiv preprint arXiv:2502.01384*, 2025.
- [46] Siyan Zhao, Devaansh Gupta, Qinling Zheng, and Aditya Grover. d1: Scaling reasoning in diffusion large language models via reinforcement learning. *arXiv preprint arXiv:2504.12216*, 2025.

- [47] Kaiwen Zheng, Yongxin Chen, Hanzi Mao, Ming-Yu Liu, Jun Zhu, and Qinsheng Zhang. Masked diffusion models are secretly time-agnostic masked models and exploit inaccurate categorical sampling. *arXiv preprint arXiv:2409.02908*, 2024.
- [48] Lin Zheng, Jianbo Yuan, Lei Yu, and Lingpeng Kong. A reparameterized discrete diffusion model for text generation. *arXiv preprint arXiv:2302.05737*, 2023.

A Analysis of Thinking Process in SEDD+DCoLT

For the Sudoku dataset, we aim to visualize the thinking process by analyzing the generation step of different cells. We categorize all 4×4 cells into three groups: (1) *given cells* – those initially provided by the puzzle; (2) *easy cells* – those that can be directly inferred using basic Sudoku rules (each number 1–4 must appear exactly once per row, column, and 2×2 subgrid); and (3) *hard cells* – the remaining positions that require more complex reasoning, often involving the resolution of dependencies among other non-given cells.

In Fig. 3, we plot the distribution of generation step for these three cell types, the left side displays statistics for the SEDD + SFT model, while the right side shows our further RL-trained SEDD + DCoLT model. The right figure reveals that the our SEDD + DCoLT significantly prioritizes generating *easy cells* in earlier steps than *hard cells*, as these cells can be directly determined from the initial *given cells*, indicating that our model learns a progressive generation strategy from easy to hard. Notably, in the left figure, we observe almost no difference between generation orders across various difficulty levels for SEDD + SFT model. This comparison reveals that the DCoLT changes its SFT-trained counterpart’s reasoning behavior to more progressive generation aligned naturally with how humankind handles problems in a real world.

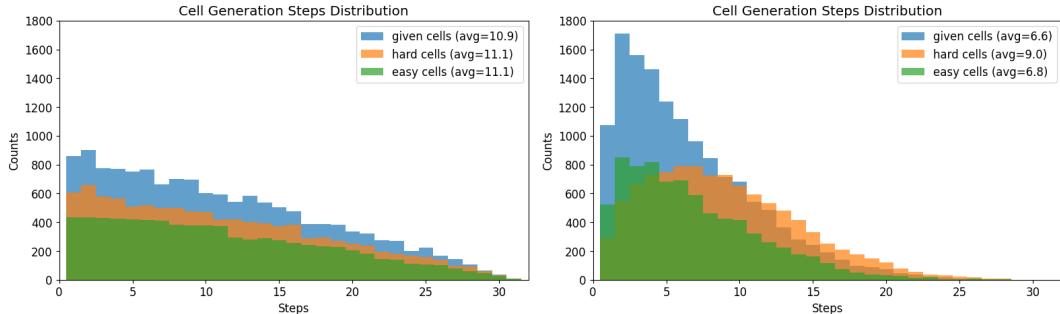


Figure 3: This figure shows the model’s inference patterns on Sudoku dataset, with the left side displaying the pattern of SEDD + SFT model and the right side showing that of our SEDD + DCoLT model. We plot the total number of generated tokens for these three cell types over diffusion steps on Sudoku 4×4 dataset, with the number of sampling steps set to 32.

To further demonstrates the progressive generation behavior on sample level, we visualize the generation process of our SEDD + DCoLT on the Sudoku dataset in Fig. 4. We show the predicted \hat{x}_0 at each timestep, obtained by selecting the token with the highest probability (excluding the mask token) at each token position. As shown in the top row, positions (0, 0) and (0, 1) can be regarded as *hard cells*, because the value can’t be directly inferred from *given cells* denoted by black borders. Initially, our model assigns the highest probabilities to incorrect answers in these red cells due to insufficient contextual information. However, as the denoising process progresses, the model gradually refines its predictions by leveraging information from newly unmasked cells (shown in dark blue), eventually converging to the correct values. This demonstrates the importance of nonlinear generation for Sudoku 4×4 tasks.

For GSM8K-Aug dataset, since it’s hard to determine which token position is more difficult to predict, we only show the generation process and the generation step statistics for each position. Fig. 5 demonstrates the contrasting generation behaviors between SEDD + DCoLT and GPT2 + CoT on the GSM8K-Aug dataset. While GPT2 + CoT consistently follows a left-to-right generation order (where earlier generation steps correspond to lower token positions), SEDD + DCoLT shows minimal positional bias in generation order across token positions. Furthermore, our results in Fig. 6 reveal

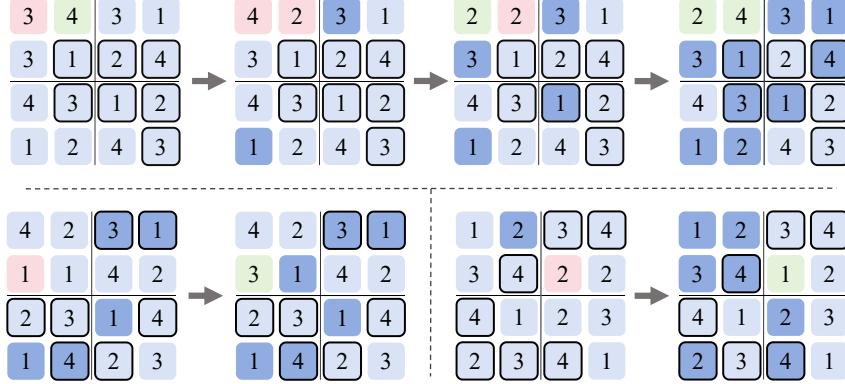


Figure 4: Visualization of predicted tokens \hat{x}_0 by SEDD + DCoLT on Sudoku 4×4 : those that are still masked appear in light blue, with unmasked ones in dark blue, incorrect predictions in red, and corrected ones in green. Black borders indicate the given cells.

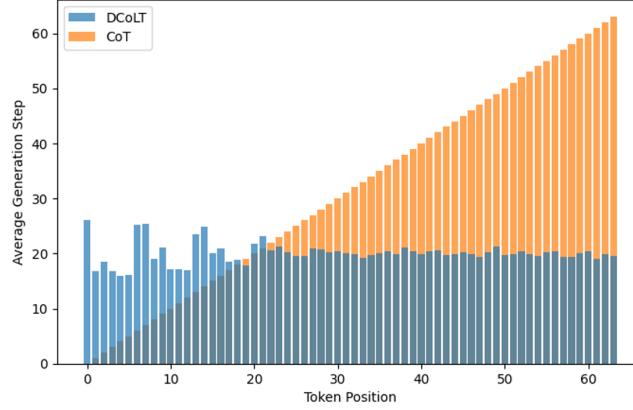


Figure 5: Average generation step for each token position in GSM8K-Aug. Different from CoT, SEDD + DCoLT generates in a non-linear way.

that SEDD + DCoLT’s generation order is sample-dependent, maintaining a nonlinear generation throughout the process.

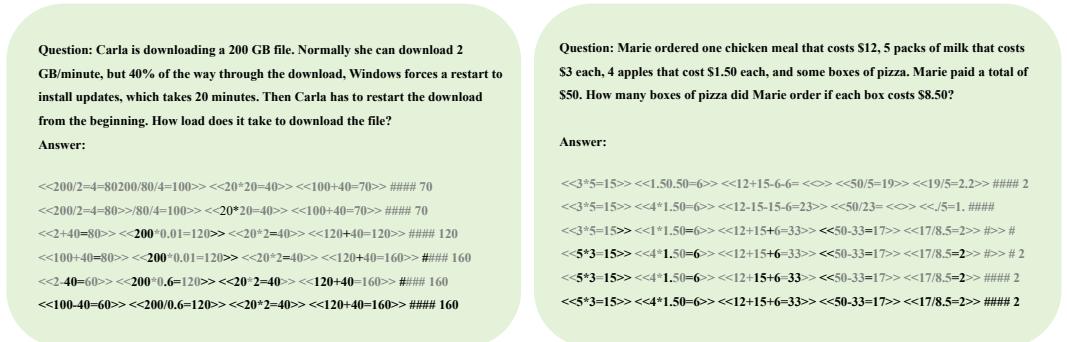


Figure 6: Our model’s generation process on GSM8K-Aug, each line displaying the predicted- \hat{x}_0 tokens across different steps, with masked tokens shown in gray and unmasked ones shown in black.

Table 5: Evaluation Hyperparameters for Sudoku and GSM8K-Aug

Models	Sudoku			GSM8K-Aug				
	Generation	Tokens	NFEs	Temp.	Generation	Tokens	NFEs	Temp.
GPT2 + CoT	256	256	0.0		256	256	-	
SEDD + DoT	256	512	0.5		256	64	0.5	
SEDD + DCoLT	19	32	0.5		64	64	0.0	

Table 6: Training Hyperparameters for Sudoku

Models	Post-Training	Post-Training Steps	Learning Rate	Batch Size
GPT2 + CoT	SFT	1,500	3e-4	1024
SEDD + DoT	SFT	1,500	1e-4	1024
GPT2 + CoT	RL	500	1e-5	64
SEDD + DCoLT	RL	500	1e-5	64

B Analysis of Thinking Process in LLaDOU (LLaDA+DCoLT)

We illustrate how LLaDOU generates a 64-token answer to a GSM8K question in Fig. 7. Tokens shown in darker shades are generated in later diffusion steps. As the figure shows, key numbers and symbols tend to emerge early in the generation process, while surrounding textual elements are filled in later to ensure grammatical correctness and fluency.

Fig. 8 and Fig. 9 further show intermediate steps of the output when generating responses of length 256 with 64 steps, in which key numbers and symbols are generated to gradually form the structure of thinking processes before grammatically correct sentences are completed.

Fig. 10 visualizes the predicted ranking scores $h_{\theta,n}$ to unmask tokens during the generation process. The examples show that at each step, the unmasking score evaluates how likely the predicted token is correct at each position. For example, in Step 20 of case (a), the token “3” prior to the selected token “0” has a smaller value of the unmasking score, suggesting that it is more likely than “0” to be wrong. This token is corrected to “0” before being unmasked at a later step.

C Experiment Settings

C.1 SEDD+DCoLT

Sudoku 4×4 We first generated 50,000 sudoku puzzles of size 4×4 by uniformly replacing 1-9 cells with zeros. Using DeepSeek V3-0324[9], we produced CoT reasoning steps and filtered the results to remove repetitive responses and incorrect solutions. All training datasets were created from this cleaned dataset to ensure fair comparison. Dataset samples can be found in the Fig. 11.

We trained both DoT and CoT (with SFT) for 1,500 steps. For CoT (with RL), we first performed 1,000 steps of SFT to pretrain the CoT response, reaching 70.5% in accuracy, followed by another 500 steps of RL training using GRPO[36]. For a fair comparison, DCoLT was also pre-trained to reach 72.3% accuracy before applying 500 steps of RL training. Tab. 6 presents training hyperparameters for the Sudoku 4×4 dataset. For both SFT and RL post-training, we employ Adam optimizer with $(\beta_1, \beta_2) = (0.9, 0.999)$. During RL post-training, we set the group size of GRPO to 32, and the kl coefficient to 0.

GSM8K-Aug Our experiments are conducted on the GSM8K-Aug dataset (sourced from [10]), which contains 384,623 training samples with simplified CoT processes. After filtering out samples with too long answers (exceeding 64 tokens) and those with invalid formats, we retained 382,553 examples. For the pretraining of SEDD on GSM8K-Aug, we train it for 120K steps using the Adam optimizer with a learning rate of 3×10^{-4} , a batch size of 512, and a generation length of 64 tokens. In the RL training phase, we employ the Adam optimizer with a learning rate of 5×10^{-5} and a batch size of 32, and set the group size for GRPO to 64. This RL phase is trained for 1,000 steps.

M	ig	uel	uses	2	x	3	0	=	6	0	sheets	of	paper	per
week	.	If	there	are	4	weeks	in	a	month	,	then	Miguel		
will	use	6	0	x	4	=	2	4	0	sheets	of	paper	every	
month	.	Here	is	the	solution	:	2	4	0	< eot_id >				

(a) Miguel uses 2 pads of paper a week for his drawing. If there are 30 sheets of paper on a pad of paper, how many sheets of paper does he use every month?

If	each	train	travels	8	0	miles	west	and	then	1	5	0	
miles	north	,	each	train	travels	a	total	of	8	0	+	1	5
0	=	2	3	0	miles	.	Therefore	,	each	train	covers	2	3
0	miles	in	the	two	days	.	Result	:	2	3	0	< eot_id >	

(b) Two trains leave San Rafael at the same time. They begin traveling westward, both traveling for 80 miles. The next day, they travel northwards, covering 150 miles. What's the distance covered by each train in the two days?

She	pays	5	0	0	+	8	0	0	=	\$	1	3	0	0
0	0	*	1	0	=	\$	1	3	0	for	insurance	.	That	means
she	paid	1	3	0	0	+	1	3	0	=	\$	1	4	3
is	1	4	3	0	< eot_id >	< endoftext >				The	answer			

(c) Janet buys a brooch for her daughter. She pays \$500 for the material to make it and then another \$800 for the jeweler to construct it. After that, she pays 10% of that to get it insured. How much did she pay?

The	child	already	has	\$	4	savings	,	so	he	needs	to	save	an
additional	\$	1	2	-	\$	4	=	\$	8	.	If	he	promises
\$	2	daily	,	then	it	will	take	him	\$	8	/	\$	2
to	save	the	remaining	amount	of	money	.	Solution	:	4	=	4	days

(d) An 8-year old child wants to buy a toy car which costs \$12. He already has \$4 savings. How many days will it take him to save the remaining amount of money if he promises to save \$2 daily from his allowance?

Figure 7: Examples of 64-token responses generated by LLaDOU on GSM8K. Token color reflects the generation order, progressing from light (early steps) to dark (later steps). The corresponding questions are provided in the sub-captions. We observe from the results that many tokens bearing key information to final answers such as the numbers, units and factual entities tend to be unmasked at earlier steps, while those tokens for function words and symbols such as “of” and “the” tend to be unmasked at later steps.

Step 10

The robe takes 2 bolts blue fiber.
** Step 2. Determine mask amount:

Step 20

To determine total bolts make robe we follow follow follow steps:

Step 1: Identify the amount<mask> blue fiber used.
The robe takes 2 bolts <mask> blue fiber.
<Step 2: Determine<mask> amount<mask> white<mask><mask><mask>
The robe takes half<mask> white<mask> white<mask> white<mask> blue fiber<mask>. Since<mask> there are 2 bolts<mask> blue fiber<mask><mask><mask> amount<mask> white<mask> is:

Step 36

To determine the total bolts used make robe we follow follow steps:

**Step 1: Identify the amount<mask> blue fiber used **
The robot takes 2 bolts<mask> blue fiber.

**Step 2: Determine the amount<mask> white fiber used **
The robe takes half as much white fiber<mask> blue fiber. Since there are 2 bolts<mask> blue fiber<mask><mask> amount<mask> white fiber is:
$$\frac{1}{2} \times 2 = 1$$
 bolt<mask> white fiber.

**Step 3: Calculate the total number<mask> bolts<mask> **
Add the total number of bolts from the blue fiber<mask> white fiber.
$$2 + 1 = 3$$
 bolts<mask>

Step 50

To determine the total number<mask> bolts used<mask> make<mask> robe<mask> we follow follow follow steps.

Step 1: Identify the amount blue fiber used.
The robe takes 2 bolts blue fiber.

Step 2: Determine the amount white fiber used.
The robe takes half as much white fiber as blue fiber. Since there are 2 bolts blue fiber amount white fiber is:
 $\frac{1}{2} \times 2 = 1$ bolt white fiber.

Step 3: Calculate the total number bolts used.
Add the total number of bolts from the blue fiber and white fiber:
 $2 + 1 = 3$ bolts

><eos><eos><eo

To determine the total number of bolts used to make the robe, we follow the steps:
Step 1: Identify the amount of blue fiber used.
The robe takes 2 bolts of blue fiber.

Step 2: Determine the amount of white fiber used.
The robe takes half as much white fiber as blue fiber. Since there are 2 bolts of blue fiber, the amount of white fiber is:
$$\lfloor \frac{1}{2} \rfloor \times 2 = 1$$

Step 3: Calculate the total number of bolts used.
Add the total number of bolts from the blue fiber and the white fiber:
$$2 + 1 = 3$$

Figure 8: Intermediate step outputs from LLaDOU during the generation of a 256-token response. The question is from GSM8K: *A robe takes 2 bolts of blue fiber and half that much white fiber. How many bolts in total does it take?*

All training was conducted on 8 H800 GPUs. For GSM8K-Aug pre-training, we spent 160 GPU hours on pretraining for 120K steps, and additional 168 GPU hours for the RL post-training. Evaluation hyperparameters for both datasets are provided in Tab. 5.

C.2 LLaPOU (LLaPA \pm DCoLT)

C.2.1 Implementation Details for Math Problems

Datasets & Benchmarks We consider two popular benchmarks, GSM8K [8] and MATH [16]. In our experiments, we follow the common train-test split on these datasets. For GSM8K, there are 7.5K questions for training and 1.32K questions for testing. For MATH, there are 7.5K questions for training and 5K questions for testing. We report the accuracy on their test set. During training, we extract the final answers from generated responses to compute the rewards.

Reward function We assign a hard reward for each completion, that it is 1 only if the final answer is equivalent to the ground truth [16].

Step 12

To determine `mask><mask>` cars drove`<mask><mask><mask><mask><mask><mask><mask><mask><mask><mask><mask><mask><mask><mask><mask><mask><mask><mask><mask>` can follow follow steps:`<mask>`
1. Total the number`<mask>` cars`<mask>` the motor`<mask>:`”
There originally 33 cars on`<mask>` motor`<mask>`.

Step 23

To determine `<mask><mask>` cars drove `<mask><mask>` traffic`</mask><mask><mask><mask><mask><mask><mask><mask>`, we can follow follow steps:`<mask>`

1. "Total the total number`<mask>` cars on the motor`<mask>`."

There originally 33 cars on`<mask>` motor`<mask>`.

2. "Subtract" **number** **cars** **took an exit** **cars** **took an exit** **cars** **took an exit**
5 cars took an exit so total cars drove through traffic is:
 $30 - 5 = 25$

Step 45

To determine `<mask><mask>` cars drove `<mask><mask>` traffic jam `<mask><mask>` first `<mask>1<mask>` minutes, we can follow follow steps

1. **Total the total number of cars on the motorway:**
There originally 22 cars on the motorway.

There originally 33 cars on the motorway.

2. **Subtract the number<mask> cars that
5 cars took an exit<mask> so<mask> total

$$30 - 5 = 25$$

30 - 5 = 25.
<mask>]

3. **Determine the number<mask> cars<mask> drove<mask><mask><mask> re

20 cars drove<mask><mask><mask> remaining 1<mask> minutes.
4<mask> **Calculate<mask> number<mask> cars<mask> drove through<mask><mask> first 1<mask> minutes<mask>
subtract the number<mask> cars<mask> drove<mask><mask><mask> remaining 1<mask> minutes<mask> total<mask><mask> cars<mask> drove<mask><mask> traffic jam:
V
25 - 20 = 5.
<mask><mask>

To determine how many cars drove through

1. ***Total the
There original

3. **Subtract the

5 cars took an exit, so the total of cars that drove through the town is 10.

$$\begin{array}{r} \underline{\quad} \\ - 5 = 25 \end{array}$$

3. **Determine the number of cars that drove through in the remaining 15 minutes:**
20 cars drove through in the remaining 15 minutes.

4. **Calculate the number of cars that drove through in the first 15 minutes:**
Subtract the number of cars that drove through in the remaining 15 minutes from the total number of cars that drove through the traffic jam:
 └
 └ 25 - 20 = 5.
 └

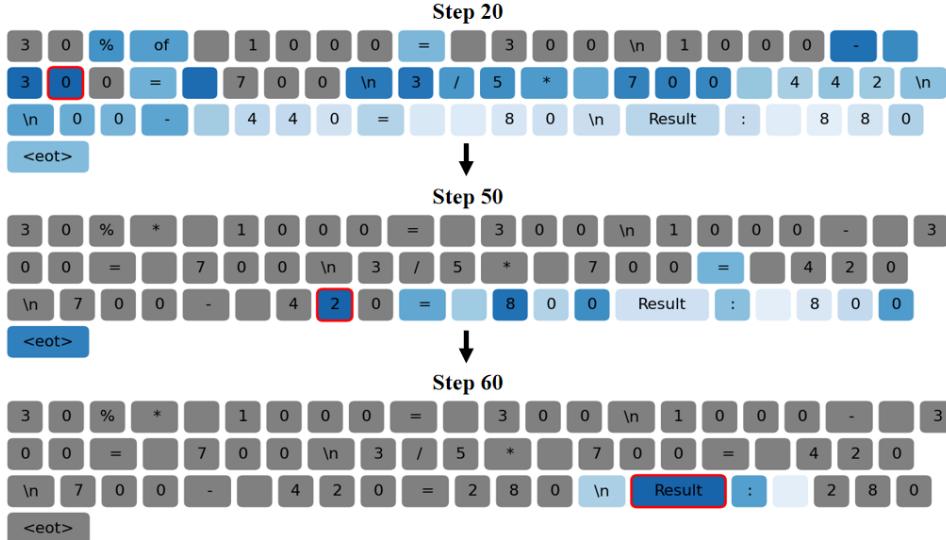
Figure 9: Intermediate step outputs from LLaDOU during the generation of a 256-token response. The question is from GSM8K: *Cars have lined up on the motorway. Some of the cars drive through in the first 15 minutes of the traffic jam, then 20 more cars drive through in the remaining 15 minutes of the jam. 5 cars from the line take an exit so they don't have to drive through the traffic jam. If there were originally 30 cars on the motorway, how many cars drove through the traffic jam in the first 15 minutes?*

Training configurations The model is trained with 64 prompts in a batch, each generating 16 completions to form a group for advantage calculation. We take an AdamW optimizer with a learning rate of 5×10^{-6} , and $(\beta_1, \beta_2) = (0.9, 0.999)$. We do not apply the KL penalty by default, as it provides marginal benefits in our experiments. The whole training lasts for 140 iterations on 16 H800 GPUs, which takes about 63 GPU days (i.e., about 4 days on wall clock with 16 GPUs).

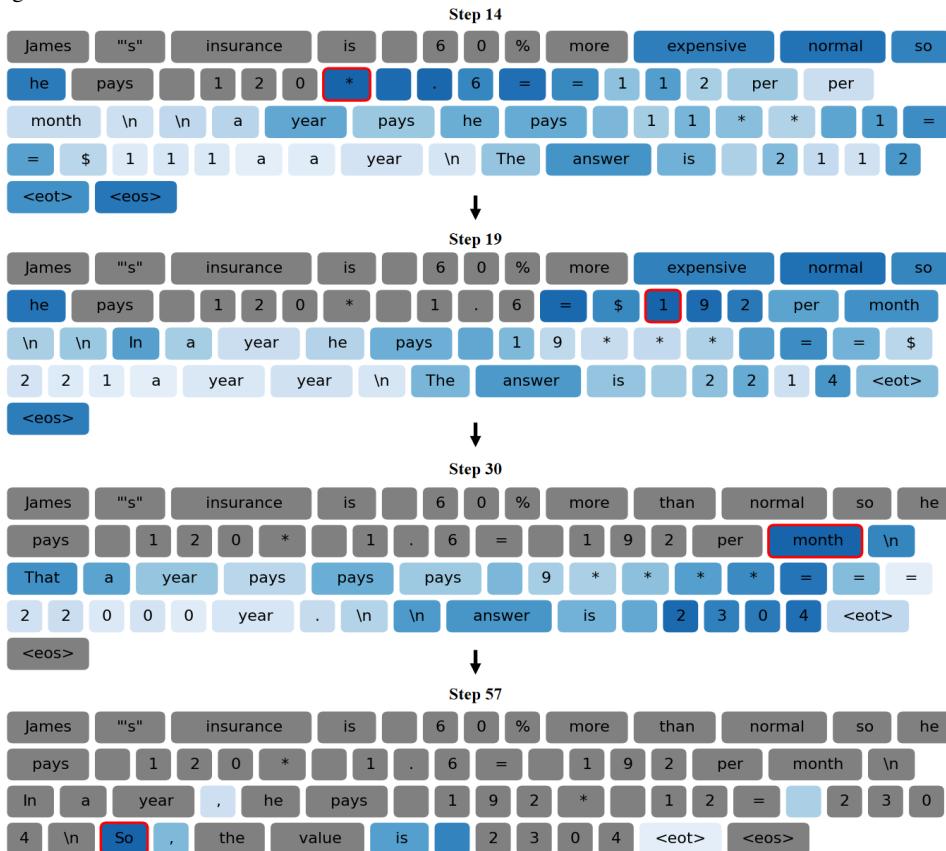
C.2.2 Implementation Details for Code Generation

Dataset We filter the KodCode-V1-SFT-R1 [40] dataset to construct our training dataset. KodCode is a large fully-synthetic open-source dataset providing verifiable code solutions and test cases for coding tasks.

To ensure the quality and consistency of the training data, we retain only samples labeled with the "instruct" style and containing exactly one entry function for test with a non-empty docstring. We further exclude any samples where the provided solutions include class definitions or constructor methods, and ensure that the solutions contain exactly one function definition. Additionally, we only use samples from a group of subsets, including Prefill, Taco, Leetcode, Codeforces, Code Contests, and Filter, for their appropriate difficulty level.



- (a) The number of students in a school hall was 1000. The hall had 3 entrances A, B, and C which also served as the exits. If after a meeting 30% of the students went out of the building through exit A, $\frac{3}{5}$ of the remaining went out through exit B, and the rest went out through exit C, calculate the number of students who went out through exit C.



- (b) James has to buy insurance. Since he had an accident it was 60% more than normal. The normal cost is \$120 a month. How much does he pay a year?

Figure 10: Illustration of intermediate steps of generated responses by LLaDOU on GSM8K. Gray tokens represent positions that have already been unmasked prior to the current step. For each masked position, we show the predicted token with the highest generation probability. Token color indicates the unmasking score, with darker blue indicating higher probability of being unmasked. The token eventually selected to unmask at each step is bounded in a red box. The corresponding prompts are given in the sub-captions.

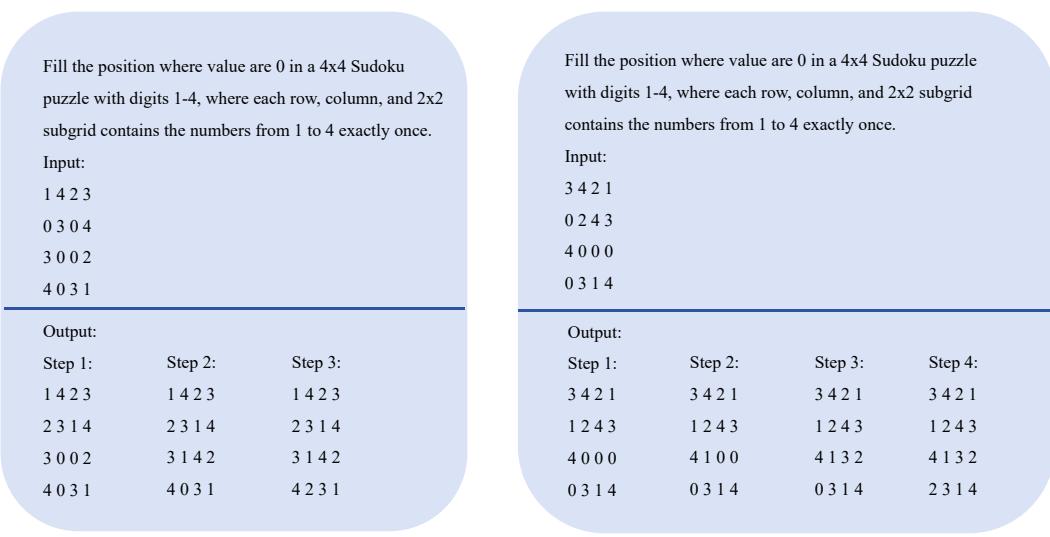


Figure 11: Examples of CoT data used for training GPT2+CoT and SEDD+DoT.

After filtering, we obtain a training set with 48.9K samples (around 12M tokens). We then format the samples into the prompt template below:

You are an expert Python programmer. Your task is to complete the implementation of a function named `<function_name>`.

```
** TARGET FUNCTION **
<docstring description>

** UNIT TESTS **
Your code should pass unit tests like:
<assert statement 1>
<assert statement 2>
...

Here is the function to complete:
```python
def <function_name>(<parameters>):
 """<docstring description>"""
```

```

Reward function We assign a hard 0/1 reward for each completion, that it is 1 only if the generated code block passes all provided test cases. The code blocks are extracted from the responses based on the markdown format.

Training configurations The training configurations are kept the same as for math. The training lasts for 240 iterations, and it takes about 127 GPU days on 16 H800 GPUs.

Evaluation details We evaluate the model in a zero-shot setting on MBPP [4] and HumanEval [7], and report the pass@1 metric. We show some examples of their input prompts below.

- **MBPP:**

You are an expert Python programmer. Your task is to complete the implementation of a function named `remove_Occ`.

```
** TARGET FUNCTION **
```

Write a python function to remove first and last occurrence of a given character from the string.

** UNIT TESTS **

Your code should pass unit tests like:

```
assert remove_Occ("hello", "l") == "heo"
assert remove_Occ("abcda", "a") == "bcd"
assert remove_Occ("PHP", "P") == "H"
```

Here is the function to complete:

```
```python
def remove_Occ(input_param_1, input_param_2):
 """Write a python function to remove first and last occurrence of a
 given character from the string."""
```

```

- **Humaneval:**

You are an expert Python programmer. Your task is to complete the implementation of a function named `has_close_elements`.

Here is the function to complete:

```
```python
from typing import List
def has_close_elements(numbers: List[float], threshold: float) -> bool:
 """Check if in given list of numbers, any two numbers are closer to each
 other than the given threshold.

Examples:
>>> has_close_elements([1.0, 2.0, 3.0], 0.5)
False
>>> has_close_elements([1.0, 2.8, 3.0, 4.0, 5.0, 2.0], 0.3)
True
"""
```

```

C.3 Post-Training Data Used in Tab. 2

In Tab. 7, we list the data used in post-training for math and code generation tasks by different models. Note that some methods [34, 14, 36] involve a considerable large amount of proprietary math/code data for both pretraining and post-training of models, which makes crucial contributions to high accuracies.

D Potential Scaling Law for Longer Generations

Fig. 12 illustrates some examples of generated responses of various lengths to the same prompt for the LLaDOU. In DLMs, the generation length tends to increase as more mask tokens are initialized at the beginning of the reverse diffusion process. This differs from auto-regressive language models, where the length of a generated response is determined by when the End of Text (EoT) token appears.

Although DLMs also use the EoT token to mark the end of the output, the token often appears later in the output sequence to generate a longer thinking process when more mask tokens are initialized. This offers a flexible way for DLMs to control various lengths of generated responses. As shown in Tab. 4, longer generations could improve performance, particularly when the model is fine-tuned by DCoLT with increasing generation length. This suggests the existence of a potential scaling law for longer generations.

Fig. 12b shows that some mistakes generated in the thinking process can be corrected in longer generations, thus achieving better results over shorter ones. We will study such a scaling phenomenon further in the future.

Answer (Generation Length=256):

- Set the equations equal to each other: $x^4 = 7x^2 - 10$.
- Rearrange the equation: $x^4 - 7x^2 + 10 = 0$.
- Let $u = x^2$, then $u^2 - 7u + 10 = 0$.
- Solve the quadratic equation: $(u = \frac{7 \pm \sqrt{49 - 40}}{2})$
 $\Rightarrow u = 5 \text{ or } u = 2$.
- Therefore, $x^2 = 5$ or $x^2 = 2$.
- The (x) -coordinates are $\pm\sqrt{5}$ and $\pm\sqrt{2}$.
- Identify $(m = 5)$ and $(n = 2)$.
- Calculate $(m - n)$: $m - n = 5 - 2 = 3$.

Answer (Generation Length=384):

- Set the equations equal to each other: $x^4 = 7x^2 - 10$
- Rearrange the equation: $x^4 - 7x^2 + 10 = 0$
- Let $u = x^2$:
- Solve the quadratic equation:

$$u = \frac{-(-7) \pm \sqrt{(-7)^2 - 4 \cdot 1 \cdot 10}}{2 \cdot 1}$$

$$u = \frac{7 \pm \sqrt{49 - 40}}{2}$$

$$u = \frac{7 \pm \sqrt{9}}{2}$$

$$u = \frac{7 \pm 3}{2}$$

$$u = 5 \quad \text{or} \quad u = 2$$
- Substitute back to find x :

$$x^2 = 5 \quad \text{or} \quad x^2 = 2$$

$$x = \pm\sqrt{5} \quad \text{or} \quad x = \pm\sqrt{2}$$
- Determine (m) and (n) :

$$m = 5 \quad \text{and} \quad n = 2$$
- Calculate $(m - n)$:

$$m - n = 5 - 2 = 3$$

Answer (Generation Length=512):

- Step 1: Set the equations equal to each other.
 $x^4 = 7x^2 - 10$
- Step 2: Rearrange the equation to set it to zero.
 $x^4 - 7x^2 + 10 = 0$
- Step 3: Let $u = x^2$. The equation becomes:
 $u^2 - 7u + 10 = 0$
- Step 4: Solve the quadratic equation for u .
Using the quadratic formula $u = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$, where $(a = 1)$, $(b = -7)$, and $(c = 10)$:

$$u = \frac{-(-7) \pm \sqrt{(-7)^2 - 4 \cdot 1 \cdot 10}}{2 \cdot 1}$$

$$u = \frac{7 \pm \sqrt{49 - 40}}{2}$$

$$u = \frac{7 \pm \sqrt{9}}{2}$$

$$u = \frac{7 \pm 3}{2}$$
- Step 5: Calculate the two solutions for u .
 $u = \frac{10}{2} = 5$
 $u = \frac{4}{2} = 2$

- Step 6: Find the corresponding x values.
For $(u = 5)$:
 $x^2 = 5 \Rightarrow x = \pm\sqrt{5}$
For $(u = 2)$:
 $x^2 = 2 \Rightarrow x = \pm\sqrt{2}$

- Step 7: Determine (m) and (n) .
From $(x^2 = 5)$, $(m = 5)$.
From $(x^2 = 2)$, $(n = 2)$.

- Step 8: Calculate $(m - n)$.
 $m - n = 5 - 2 = 3$

Final answer:
 $\boxed{3}$

Answer (Generation Length=512 finetuned):

To find the points of intersection between the graphs of $y = x^4$ and $y = 7x^2 - 10$, we set the equations equal to each other:
 $x^4 = 7x^2 - 10$

Rearrange the equation to form a quartic equation:
 $x^4 - 7x^2 + 10 = 0$

Let $u = x^2$. Then the equation becomes:
 $u^2 - 7u + 10 = 0$

We solve this quadratic equation using the quadratic formula $u = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$:

$$u = \frac{-(-7) \pm \sqrt{(-7)^2 - 4 \cdot 1 \cdot 10}}{2 \cdot 1}$$

$$u = \frac{7 \pm \sqrt{49 - 40}}{2}$$

$$u = \frac{7 \pm 3}{2}$$

This gives us two solutions:
 $u = \frac{10}{2} = 5$
 $u = \frac{4}{2} = 2$

Since $u = x^2$, we have:
 $x^2 = 5$
 $x^2 = 2$

Thus, the x -coordinates are:
 $x = \pm\sqrt{5}$
 $x = \pm\sqrt{2}$

We are given that $(m > n)$, so $(m = 5)$ and $(n = 2)$. Therefore,
 $m - n = 5 - 2 = 3$

Thus, the answer is:
 $\boxed{3}$

- (a) The graphs of $y = x^4$ and $y = 7x^2 - 10$ intersect at four points with x -coordinates $\pm\sqrt{m}$ and $\pm\sqrt{n}$, where $m > n$. What is $m - n$?

Answer (Generation Length=256):

First, we convert the numbers in base 5 to base 10 for easier calculation:
 $(4321_5 = 4 \cdot 5^3 + 3 \cdot 5^2 + 2 \cdot 5 + 1 = 276)$
 $(1234_5 = 1 \cdot 5^3 + 2 \cdot 5^2 + 3 \cdot 5 + 4 = 139)$
- Perform the subtraction: $(276 - 139 = 137)$

Next, we convert 137 back to base 5:
- $(137 \text{ div } 5 = 27)$ remainder (2)
- $(27 \text{ div } 5 = 5)$ remainder (2)
- $(5 \text{ div } 5 = 1)$ remainder (0)
- $(1 \text{ div } 5 = 0)$ remainder (1)

Reading the remainders from bottom to top, we get (1042_5) .

Therefore, the answer in base 5 is $\boxed{3042_5}$

Answer (Generation Length=384):

Step 1: Convert both numbers to base 10.
 $(4321_5):$
 $(4 \cdot 5^3 + 3 \cdot 5^2 + 2 \cdot 5 + 1 \cdot 5^0 = 1000 + 75 + 10 + 1 = 1086)$
 $(1234_5):$
 $(1 \cdot 5^3 + 2 \cdot 5^2 + 3 \cdot 5^1 + 4 \cdot 5^0 = 125 + 50 + 15 + 4 = 194)$

Step 2: Perform the subtraction in base 10.
 $[1086 - 194 = 892]$

Step 3: Convert the result back to base 5.
 $[892 \text{ div } 5 = 178 \text{ remainder } 2]$
 $[178 \text{ div } 5 = 35 \text{ remainder } 3]$
 $[35 \text{ div } 5 = 7 \text{ remainder } 0]$
 $[7 \text{ div } 5 = 1 \text{ remainder } 2]$
 $[1 \text{ div } 5 = 0 \text{ remainder } 1]$

Reading the remainders from bottom to top, we get (1202_5) .

$\boxed{3202_5}$

Answer (Generation Length=512 finetuned):

Step 1: Convert (4321_5) to base 10.

```
\[
4321_5 = 4 \cdot 5^3 + 3 \cdot 5^2 + 2 \cdot 5 + 1 = 4 \cdot 125 + 3 \cdot 25 + 2 \cdot 5 + 1 = 500 + 75 + 10 + 1 = 586
]
```

Step 2: Convert (1234_5) to base 10.

```
\[
1234_5 = 1 \cdot 5^3 + 2 \cdot 5^2 + 3 \cdot 5 + 4 = 1 \cdot 125 + 2 \cdot 25 + 3 \cdot 5 + 4 = 125 + 50 + 15 + 4 = 194
]
```

Step 3: Subtract the base 10 equivalents.

```
\[
586 - 194 = 392
]
```

Step 4: Convert 392 back to base 5.

```
\[
392 \text{ div } 5 = 78 \quad \text{remainder} \quad 2
\]
\[
78 \text{ div } 5 = 15 \quad \text{remainder} \quad 3
\]
\[
15 \text{ div } 5 = 3 \quad \text{remainder} \quad 0
\]
\[
3 \text{ div } 5 = 0 \quad \text{remainder} \quad 3
\]

```

Reading the remainders from bottom to top, we get (3032_5) .

$\boxed{3032_5}$

Answer (Generation Length=512 finetuned):

First, we convert both numbers to base 10 for easier calculation.

```
\[
(4321_5):
\[
4 \cdot 5^3 + 3 \cdot 5^2 + 2 \cdot 5 + 1 = 4 \cdot 125 + 3 \cdot 25 + 2 \cdot 5 + 1 = 500 + 75 + 10 + 1 = 586
\]
]
```

```
\[
(1234_5):
\[
1 \cdot 5^3 + 2 \cdot 5^2 + 3 \cdot 5 + 4 = 1 \cdot 125 + 2 \cdot 25 + 3 \cdot 5 + 4 = 125 + 50 + 15 + 4 = 194
\]
]
```

Next, we perform the subtraction in base 10:

```
\[
586 - 194 = 392
\]
```

Now, we convert 392 back to base 5.

First, repeatedly divide by 5:

```
\[
392 \text{ div } 5 = 78 \quad \text{remainder} \quad 2
\]
\[
78 \text{ div } 5 = 15 \quad \text{remainder} \quad 3
\]
\[
15 \text{ div } 5 = 3 \quad \text{remainder} \quad 0
\]
\[
3 \text{ div } 5 = 0 \quad \text{remainder} \quad 3
\]
```

Reading the remainders from bottom to top, we get (3032_5) .

Thus, the final answer is:

```
\[
\boxed{3032_5}
\]
```

(b) Find $4321_5 - 1234_5$. Express your answer in base 5.

Figure 12: Examples of how increasing generation lengths enable the DCoLT-reinforced model to unfold more complex reasoning processes. Incorrect steps are marked in red, while the corrected steps are highlighted in green. The generation length means the number of mask tokens initialized in the input sequence to the reverse diffusion process. Except for the model finetuned with 512 generation length, other models are only trained with 256 generation length as shown in experiments.

Table 7: Post-training data for math / code generation used in Tab. 2. “-” means this type of post-training was not used for the model.

| Method | SFT | RL |
|--------------------------------|---|---|
| Auto-regressive Models | | |
| LLaMA2 7B [38] | - | - |
| MetaMath 7B [44] | 395K samples (MetaMathQA) | - |
| CodeLLaMA-Instruct 7B [34] | thousands of SFT
+ millions of RS examples | - |
| Deepseek 7B [5] | - | - |
| DeepseekMath-Instruct 7B [36] | 776K (problem, solution) pairs | - |
| DeepseekMath-RL 7B [36] | 776K (problem, solution) pairs | 144K questions |
| DeepseekCoder-Instruct 7B [14] | 2B tokens | - |
| Diffusion Language Models | | |
| Dream 7B [42] | - | - |
| LLaDA 8B [27] | - | - |
| + SFT [46] | 1K samples from s1K | - |
| + diffu-GRPO [46] | - | GSM8K: 7.5K / MATH: 7.5K |
| d1-LLaDA (SFT + diffu-GRPO) | 1K samples from s1K | GSM8K: 7.5K / MATH: 7.5K |
| LLaDOU (LLaDA + DCoLT) 8B | - | Math: 7.5K GSM8K + 7.5K MATH /
Code: 48K filtered from KodCode |

Table 8: Model accuracies with different block length on GSM8K dataset. LLaDOU (LLaDA + DCoLT) is trained with a generation length of 256 and 64 steps of reverse diffusion under respective block lengths.

| Model | Block Length | | |
|-------------------------|--------------|-------|-------|
| | 8 | 32 | 256 |
| LLaDA 8B | 63.26 | 62.12 | 49.47 |
| LLaDOU (LLaDA+DCoLT) 8B | 83.17 | 83.11 | 82.27 |

E Ablation on Various Block Lengths

LLaDA [27] divides the sequence into several blocks and generates responses blockwise from left to right in a semi-autoregressive manner. This policy introduces additional priors on the block-wise order of text generation. In Tab. 8, for the baseline LLaDA 8B model, setting block length to 8 achieves an accuracy of 63.26% on GSM8K, significantly higher than 49.47% without dividing blocks.

However, when we apply DCoLT to LLaDA with different block lengths, all LLaDOU models achieve similar performance, with the one with a block length 8 achieving slightly higher accuracy (83.17%) than the one with block length 256 without dividing blocks (82.27%). These results suggest that LLaDOU does not rely on the blocking prior as in the baseline model.

F Licenses for existing assets

The code and models associated with this paper will be released to the public later. For all code and data assets, we refer to their licenses in Tab. 9.

G Broader Impacts

Our work introduces a novel perspective of lateral reasoning in language models, in contrast to the widely adopted CoT paradigms used in LLMs. By enabling non-linear, format-free reasoning through a reverse diffusion process, our approach offers new insights into the mechanisms of reasoning in

Table 9: Reference assets and their licenses.

| Asset | License | Utility |
|----------------|--------------|--------------|
| SEDD [24] | MIT | Code & Model |
| GSM8K-Aug [10] | - | Data |
| LLaDA [27] | MIT | Code & Model |
| MATH [16] | MIT | Data |
| GSM8K [8] | MIT | Data |
| KodCode [40] | CC BY-NC 4.0 | Data |

generative models. This perspective has the potential to benefit a broad range of reasoning-intensive tasks, including embodied AI, autonomous agents, and complex decision-making systems, providing transparent thinking processes to minimize the hallucination risks that could emerge from inadequate reasoning of language models.

However, we acknowledge that DCoLT could pose potential unexpected societal impacts if misused, especially when handling misleading or adversarial prompts. In this work, we focus on math and code generation tasks, in which outputs are objectively verifiable. With such a reward design, we can reduce the influence of dataset biases and encourage the development of reasoning behaviors that are aligned with the factual and logical consistency. Handling subjective preferences in rewarding the models to train the DCoLT could be more challenging, and we leave it to our future works.

H Safeguards

DCoLT is designed for math and code generation tasks that involve objectively verifiable outputs and well-defined correctness criteria. These domains present a relatively low risk of societal misuse compared to open-ended language generation tasks. To further reduce the potential for unintended use, we recommend deploying DCoLT alongside an input checker that ensures that the model only processes input relevant to its target domains. This approach helps mitigate the risks related to adversarial prompting or misuse beyond the intended scope.