LongLLaDA: Unlocking Long Context Capabilities in Diffusion LLMs

Xiaoran Liu[1,2], Yuerong Song[1,2], Zhigeng Liu[1,2], Zengfeng Huang[1,2], Qipeng Guo[2,3], Ziwei He[2*], Xipeng Qiu[1,2*]

[1]Fudan University,
[2]Shanghai Innovation Institute,
[3]Shanghai AI Lab
xrliu24@m.fudan.edu.cn, ziwei.he@sii.edu.cn, xpqiu@fudan.edu.cn

Abstract

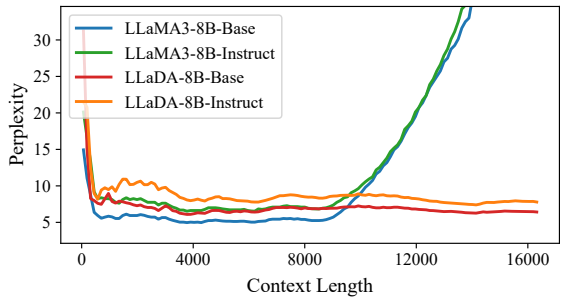
Large Language Diffusion Models, or dLLMs, have emerged as a significant focus in NLP research, with substantial effort directed toward understanding their scalability and downstream task performance. However, their long-context capabilities remain unexplored, lacking systematic analysis or methods for context extension. In this work, we present the first systematic investigation comparing the long-context performance of diffusion LLMs and traditional auto-regressive LLMs. We first identify a unique characteristic of dLLMs, unlike auto-regressive LLMs, they maintain remarkably ***stable perplexity*** during direct context extrapolation. Moreover, where auto-regressive models fail outright during the Needle-In-A-Haystack task with context exceeding their pretrained length, we discover dLLMs exhibit a distinct "***local perception***" phenomenon, enabling successful retrieval from recent context segments. We explain both phenomena through the lens of Rotary Position Embedding (RoPE) scaling theory. Building on these observations, we propose LongLLaDA, a training-free method that integrates LLaDA with the NTK-based RoPE extrapolation. Our results validate that established extrapolation scaling laws remain effective for extending the context windows of dLLMs. Furthermore, we identify long-context tasks where dLLMs outperform auto-regressive LLMs and others where they fall short. Consequently, this study establishes the first length extrapolation method for diffusion LLMs while providing essential theoretical insights and empirical benchmarks critical for advancing future research on long-context diffusion LLMs.

Code — https://github.com/OpenMOSS/LongLLaDA
Extended version — https://arxiv.org/abs/2506.14429

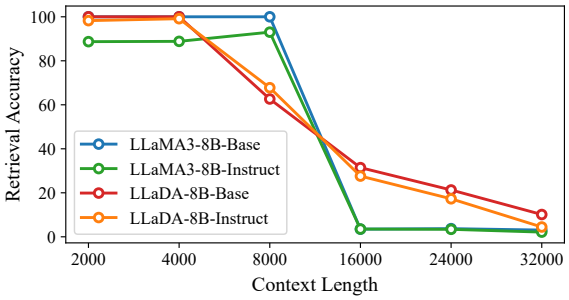
Introduction

Recently, diffusion LLMs, or dLLMs, have become widely discussed in Natural Language Processing research (Nie et al. 2025; Ye et al. 2025). They are regarded as a potential solution to key limitations of traditional auto-regressive LLMs (Touvron et al. 2023; Sun et al. 2024), including the reversal curse (Berglund et al. 2023), complex reasoning (Dziri et al. 2023), and maintaining coherence across extended contexts (Bachmann and Nagarajan 2024; Ye et al. 2024, 2025). Significant research efforts have focused on validating their scalability (Nie et al. 2025; Ye et al. 2025), adapting them for multimodality (Yang et al. 2025; You et al. 2025; Yu, Ma, and Wang 2025), applying them to reasoning tasks (Zhao et al. 2025; Huang et al. 2025; Zhu et al. 2025), and optimizing their efficiency (Ma et al. 2025; Hu et al. 2025; Wu et al. 2025). However, the long-context capabilities of dLLMs, specifically their performance and potential for length extrapolation, remain unexplored.

We begin by systematically evaluating dLLM LLaDA (Nie et al. 2025) against auto-regressive LLM LLaMA3 (Meta 2024a) on perplexity and retrieval tasks, both within and beyond their pretrained context lengths (Figure 1). Notably, dLLMs maintain stable perplexity



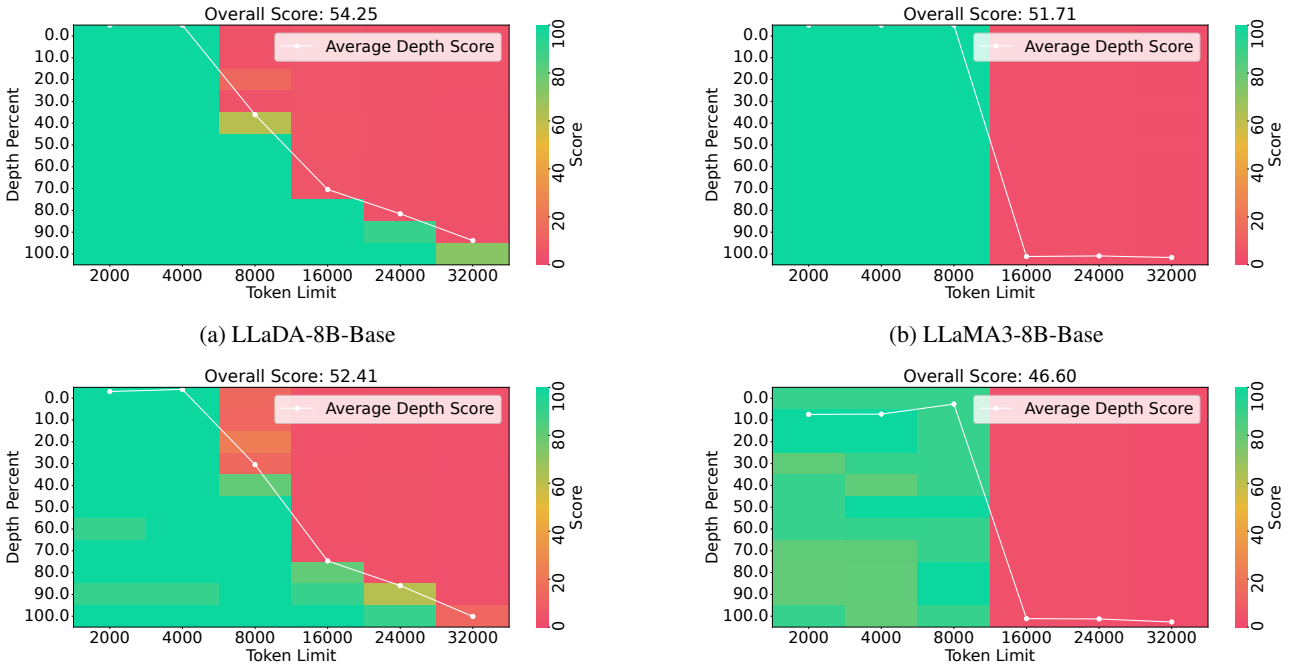
(a) Comparison of Perplexity



(b) Comparison of Retrieval Accuracy

Figure 1: Comparison of perplexity and retrieval accuracy between the diffusion LLM and the auto-regressive LLM, both within and beyond pre-training context length.

* Corresponding Author.



(a) LLaDA-8B-Base

(b) LLaMA3-8B-Base

(c) LLaDA-8B-Instruct

(d) LLaMA3-8B-Instruct

Figure 2: Results of Needle-In-A-Haystack tests (Gkamradt 2023) on LLaDA-8B Series (Nie et al. 2025) and LLaMA3-8B Series (Meta 2024b) under direct extrapolation.

and exhibit localized perception during direct length extrapolation. In stark contrast, auto-regressive LLMs suffer catastrophic perplexity surges and performance collapse when input length exceeds their maximum supported context window, 8k tokens. This divergence reveals fundamental architectural differences in long-context handling, raising critical questions: (1) What mechanisms enable dLLMs' extrapolation stability? (2) Can established length-extension techniques for auto-regressive LLMs be transferred to diffusion architectures? (3) How do dLLMs perform on long-context benchmarks relative to auto-regressive baselines, and what unique capabilities or limitations emerge?

In this work, we address these questions through comprehensive experiments and analysis. Besides the perplexity and retrieval experiment, we also benchmark Needle-In-A-Haystack (NIAH) performance for dLLMs (LLaDA (Nie et al. 2025), LLaDA-1.5 (Zhu et al. 2025), Dream-v0 (Ye et al. 2025)), quantitatively confirming their local perception bias during length extrapolation. We then analyze this phenomenon through Rotary Position Embedding (RoPE) theory, validating our interpretation with t-SNE visualizations. Building on these insights, we propose LongLLaDA, a training-free method which successfully extends LLaDA's context window using NTK-based RoPE extrapolation (bloc97 2023b), and verify preserved scaling laws (Liu et al. 2023b). Finally, we identify task-dependent capabilities where dLLMs surpass or lag behind auto-regressive counterparts on long-context benchmarks. Our contributions are summarized as follows:

- **First systematic analysis** of dLLMs' long-context behavior, revealing their unique characteristics for stable perplexity and local perception in context extension, with mechanistic explanation via RoPE dynamics.
- **Effective context extension** demonstrating NTK-based RoPE extrapolation and scaling laws transfer seamlessly to dLLMs, achieving 6× context expansion (24k tokens) without further training.
- **Capability benchmarking** revealing dLLMs match auto-regressive models on retrieval tasks, lag in aggregation, but excel at synthetic QA. We provide foundational insights for future long-context diffusion research.

Long-Context Phenomenology of dLLMs

We conduct experiments on the existing dLLM series, including LLaDA-8B (Nie et al. 2025), LLaDA-1.5 (Zhu et al. 2025), and Dream-v0 (Ye et al. 2025). By default, we set the number of sampling steps in dLLM the same as the output length and keep the sampling strategy in the official code of LLaDA and Dream. We use OpenCompass (Contributors 2023) for downstream validation. All experiments are performed on a single H100 GPU with 80 GB memory, a fixed random seed of 2025, FP16 precision for LLaDA series and BF16 precision for Dream series, and accelerated with FlashAttention2 (Dao 2023).

We first evaluate the length extrapolation capabilities of dLLMs, including LLaDA (Nie et al. 2025), LLaDA-1.5 (Zhu et al. 2025), and Dream-v0 (Ye et al. 2025), compared with auto-regressive LLMs such as LLaMA3 (Meta

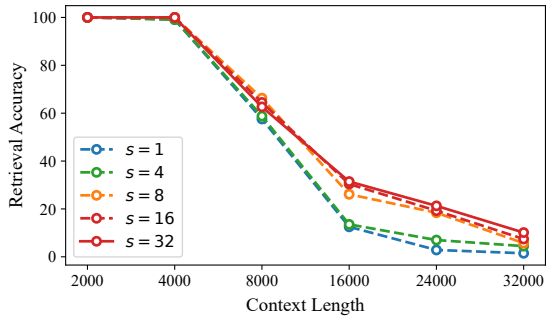


Figure 3: NIAH Results of LLaDA-8B-Base (Nie et al. 2025) with different sampling steps s.

2024a), via Needle-In-A-Haystack (Gkamradt 2023; Li et al. 2024), based on *the experimental setup in the Appendix*. All LLMs are required to generate at most 32 tokens, with dLLMs using a block size and sampling steps of 32. The results are shown in Figure 2. LLaMA3-8B-Base and LLaMA3-8B-Instruct maintain perfect retrieval accuracy within their pretrained 8k length, but suffer catastrophic performance degradation beyond this limit, failing to retrieve information at any depth. In contrast, LLaDA-8B-Base and LLaDA-8B-Instruct achieve 100% retrieval accuracy in 4k. Surprisingly, when exceeding 4k, up to 24k, LLaDA still retrieves information from the nearest 4k window, demonstrating a local perception like a sliding window. This behavior remarkably differs from auto-regressive LLM extrapolation.

Different from auto-regressive LLMs, dLLMs are influenced by sampling steps and strategies. For simplicity, we compare the impact of sampling steps on retrieval depth in NIAH. As shown in Figure 3, using the same input-output settings from previous experiments, we evaluate the LLaDA-8B-Base with sampling step $s = 1, 4, 8, 16, 32$. Results show that at 1 or 4 steps, LLaDA-8B-Base fails to retrieve information beyond 8k length, and increasing s to 8 or higher can achieve retrieval depths of 25% at 16k and almost 10% at 24k context length.

Preliminary: RoPE Extrapolation in Auto-regressive LLM

To understand the long-context phenomenology of dLLMs, we need first to clarify the RoPE-based extrapolation in auto-regressive LLMs. Rotary Position Embedding (RoPE) (Su et al. 2021) employs trigonometric functions to encode absolute positions in Q state $\boldsymbol{q}_t = \left[q_t^{(0)}, \cdots, q_t^{(d-1)}\right]$ and K state $\boldsymbol{k}_s = \left[k_s^{(0)}, \cdots, k_s^{(d-1)}\right]$. By leveraging the properties of rotation matrices, RoPE encodes relative position in the attention matrix $\boldsymbol{A}$,

$$\begin{aligned}\boldsymbol{A}_{t,s} &= (\boldsymbol{q}_t \boldsymbol{R}_t)(\boldsymbol{k}_s \boldsymbol{R}_s)^\top = \boldsymbol{q}_t \boldsymbol{R}_{t-s} \boldsymbol{k}_s^\top \\ &= \sum_{n=0}^{d/2-1} \begin{matrix} \left(q_t^{(2n)} k_s^{(2n)} + q_t^{(2n+1)} k_s^{(2n+1)}\right) \cos\theta_n(t-s), \\ -\left(q_t^{(2n)} k_s^{(2n+1)} - q_t^{(2n+1)} k_s^{(2n)}\right) \sin\theta_n(t-s) \end{matrix}, \end{aligned} \quad (1)$$

and demonstrates superior performance, thus being widely adopted by many auto-regressive LLMs (Sun et al. 2024; Dubey et al. 2024; Yang et al. 2024).

However, RoPE still faces the length extrapolation issue (Press, Smith, and Lewis 2022). When RoPE-based auto-regressive LLMs are tested beyond the pre-trained context length, the perplexity rises significantly, and downstream performance drops sharply. The underlying causes and corresponding solutions come from two key properties of trigonometric functions: ***periodicity*** and ***monotonicity***.

Rule of Periodicity According to the design of RoPE, different dimensions of $\boldsymbol{q}_t, \boldsymbol{k}_s$ use different rotary angles θ_n, with rotary base $\beta_0 = 10000$ by default and the periods T_n for $\sin(\theta_n t)$ and $\cos(\theta_n t)$ increasing from low to high dimensions as shown in Equation 2.

$$\theta_n = \beta_0^{-2n/d}, \quad T_n = 2\pi \cdot \beta_0^{2n/d}, \quad n = 0, \cdots, d/2-1. \quad (2)$$

For lower dimensions, T_n is very short, compared with the pre-trained context length T_{train}, while for higher ones, T_n becomes significantly longer, exceeding T_{train}. Consequently, there exists a ***critical dimension***, d_{extra}, as shown in Equation 3, within which $\sin(\theta_n t)$ or $\cos(\theta_n t)$ complete at least one full period within the pretrained length, whereas those beyond do not.

$$d_{\text{extra}} = 2\left\lceil \frac{d}{2} \log_{\beta_0} \frac{T_{\text{train}}}{2\pi} \right\rceil. \quad (3)$$

Therefore, dimensions beyond d_{extra} will encounter OOD position embedding when processing longer inputs and larger position indices in inference, leading to extrapolation issues (Liu et al. 2023b).

To enable LLM to handle unseen position indices, NTK methods (bloc97 2023b; Xiong et al. 2023) scale the rotary base by a factor λ, reducing the rotary angle to achieve position interpolation. However, since different dimensions undergo different degrees of interpolation, the position embedding at the critical dimension will first become OOD. Thus, based on the scaled period of the critical dimension, the extrapolation upper bound T_{extra} for NTK methods can be derived, as shown in Equation 4.

$$T_{\text{extra}} = 2\pi \cdot (\lambda \cdot \beta_0)^{d_{\text{extra}}/d}. \quad (4)$$

Based on Equation 4, for an input length t, the rotary base scaling factor λ should be set as shown in Equation 5 to ensure no OOD position embeddings occur. Notably, this adjustment coefficient exhibits a sup-linear, power-law increase with inference length (Liu et al. 2023b, 2025a).

$$\lambda_t = \beta_0^{-1} \cdot \left(\frac{t}{2\pi}\right)^{d/d_{\text{extra}}}. \quad (5)$$

It should be noted that while such interpolation could theoretically avoid extrapolation issues, it can only achieve $2\times$ to $6\times$ long-context extension during inference, as longer inputs lead to increased attention entropy, limiting further extrapolation (bloc97 2023b; Han et al. 2023; Wang et al. 2024).

Rule of Monotonicity Since the pre-trained position information of dimensions beyond the critical dimension limits the extrapolation capability of RoPE-based auto-regressive LLMs, if the rotary base is reduced, and each dimension can cover at least half or even a full period, the perplexity curve of auto-regressive LLMs will be flattened (Liu et al. 2023b). However, this does not imply real length extrapolation. Subsequent studies (Men et al. 2024; Hu et al. 2024) find that *LLMs with a smaller rotary base can only perceive local information* in downstream evaluations and fail to retrieve long-context dependencies.

Exposing LLM to periodic position information leads to downstream degeneration, manifesting a sliding-window effect, which reveals another aspect of RoPE-based extrapolation, the impact of monotonicity. Although higher dimensions do not observe complete position information, they provide a relatively complete monotonic interval, reflecting partial ordering in a long context. These dimensions exhibit larger activation values in long-context tasks (Jin et al. 2025), are more sensitive to modeling long-context dependencies (Liu et al. 2024a), and are better suited for capturing sequential information (Wei et al. 2025). Thus, solely optimizing for periodicity at the cost of monotonicity across all dimensions is wrong (Men et al. 2024; Liu et al. 2025a).

Mechanistic Analysis

Based on the length extrapolation theory of RoPE-based auto-regressive LLMs, we attribute the long-context phenomenology of RoPE-based dLLMs, namely stable perplexity and local perception, to dLLMs being trained with richer positional information compared to auto-regressive LLMs. Critically, the bidirectional attention in dLLMs exposes them to a relative position range of $[1-T_{\text{train}}, T_{\text{train}}-1]$ during training, contrasting with the $[0, T_{\text{train}}-1]$ range typical of auto-regressive models. This difference is evident in the RoPE mechanism. As visualized in Figure 4, for LLaDA ($T_{\text{train}} = 4$k) and LLaMA ($T_{\text{train}} = 8$k), we observe how the positional embeddings (sine/cosine) behave within and beyond their maximum trained relative positions.

- **High Frequencies**: Both perceive complete sinusoidal periods within their maximum trained relative distance, yielding comparable positional information encoding.
- **Moderate Frequencies**: LLaMA3's auto-regressive attention observes relative positions $[0, 8191]$ when trained on 8192-token sequences. In contrast, LLaDA's bidirectional attention observes symmetric relative positions $[-4095, 4095]$ despite its shorter 4096-token training length. This symmetric coverage provides a key advantage by fully capturing a complete period of both the cosine and sine, enhancing its tolerance of direct length extrapolation.
- **Low Frequencies**: Both models exhibit limited extrapolation capability beyond their pretrained context windows. However, as visualized in Figure 4, the out-of-distribution (OOD) regions differ remarkably: LLaMA3 struggles to capture all negative position embeddings (gray region), representing half of the potential embedding space, while LLaDA significantly reduces the unlearned OOD spaces, resulting in enhanced robustness in length extrapolation.

This results in a relatively flattened perplexity growth curve, similar to auto-regressive RoPE-based LLMs with a smaller base (Liu et al. 2023b; Men et al. 2024). However, since the cosine function in RoPE, which primarily captures relative distances, is even, negative relative positions do not increase the LLM's maximum perceivable distance in the pre-training stage. Thus, dLLM can only retrieve key information from limited relative positions within the training length, leading to the observed decay pattern in the NIAH evaluation.

We validate this interpretation with the t-SNE visualization (Van der Maaten and Hinton 2008; Zandieh et al. 2024) of QK states from the final layer of LLaMA3-8B-Base (Meta 2024a) and LLaDA-8B-Base (Nie et al. 2025), as shown in Figure 5. As shown in Figure 5a, for auto-regressive LLMs such as LLaMA3-8B-Base, the QK states within and beyond the maximum supported context length, 8k, present two different distribution clusters, and the manifold for QK states with RoPE also shows a different trend when position embedding becomes OOD. Comparatively, regarding the clustering feature for dLLMs such as LLaDA-8B-Base, there is no distribution shift between QK states within and beyond 4k, and a uniform manifold for QK states with RoPE. This demonstrates that dLLM is more robust for the OOD position embeddings in length extrapolation. Therefore, unlike traditional auto-regressive LLMs that exhibit catastrophic performance degradation when exceeding their maximum supported context length, dLLMs ***maintain stable outputs*** and ***demonstrate local perception*** in extended context.

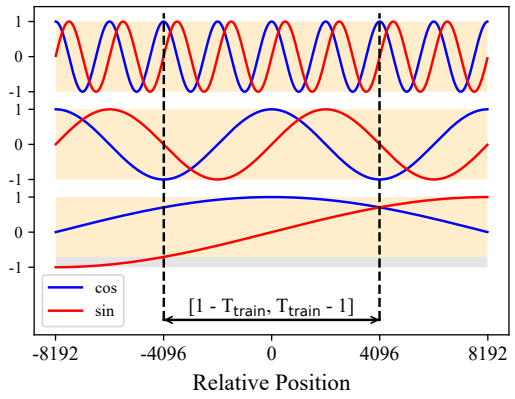
Context Extension For dLLMs

Since the reason for the surprising phenomenon has been clarified, we now move on to the extrapolation methods for dLLMs. Since the retrievable depth of dLLMs remains constrained by the range of cosine values encountered during pre-training, we transfer the NTK-based extrapolation (bloc97 2023b) and its scaling laws (Liu et al. 2023b) to dLLMs, thus proposing the length extrapolation method for dLLMs, LongLLaDA. The scaling factor λ in training-free NTK scaling (bloc97 2023b) for RoPE-based auto-regressive LLMs is decided by the extrapolation context length t and critical dimension d_{extra} calculated by rotary base β_0 and pretrained context length T_{train}, as shown in Equation 6.

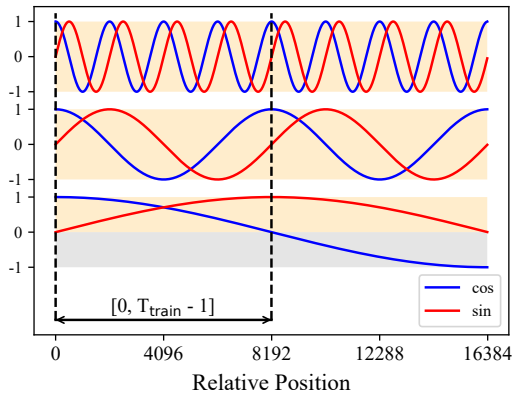
$$\lambda = 10^{-4} \cdot \left(\frac{t}{2\pi}\right)^{d/d_{\text{extra}}}, \quad d_{\text{extra}} = 2\left\lceil \frac{d}{2} \log_{\beta_0} \frac{T_{\text{train}}}{2\pi} \right\rceil. \tag{6}$$

Similarly, in LongLLaDA, based on Nie et al. (2025), the pretrained rotary base $\beta_0 = 500000$, and the pre-training context length T_{train} is 4k. This yields a critical dimension $d_{\text{extra}} = 64$. Accordingly, the required scaling factor λ for extrapolation to 8k, 16k, 24k, and 32k is calculated as 4, 14, 31, and 55, respectively. The extrapolation results are illustrated in Figure 6.

When $\lambda = 4, 14$, LongLLaDA can effectively extrapolate dLLMs to the corresponding context lengths, achieving near

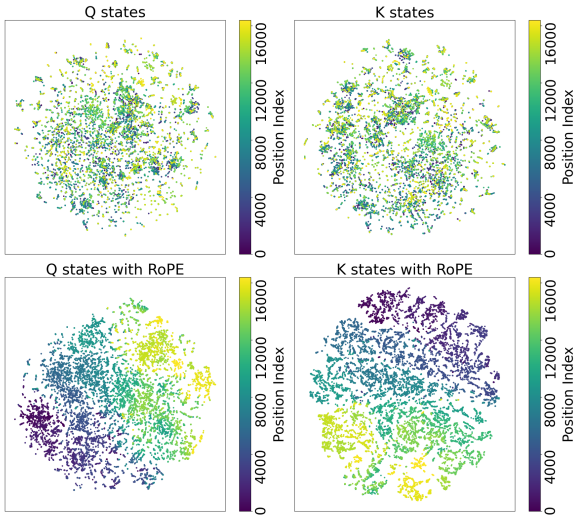


(a) LLaDA with $T_{\text{train}} = 4\text{k}$

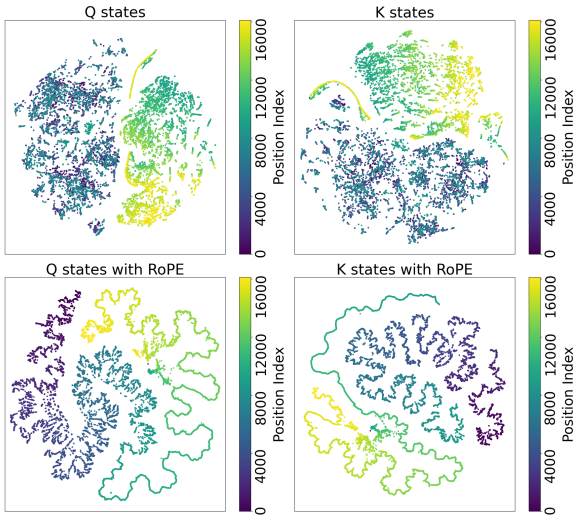


(b) LLaMA3 with $T_{\text{train}} = 8\text{k}$

Figure 4: Comparison of trained position embedding interval between LLaDA-8B and LlaMA3-8B. The area within the dashed line represents trained relative position, while that beyond represents the relative position in length extrapolation, with unlearned position embedding values colored in gray.



(a) LLaDA-8B-Base



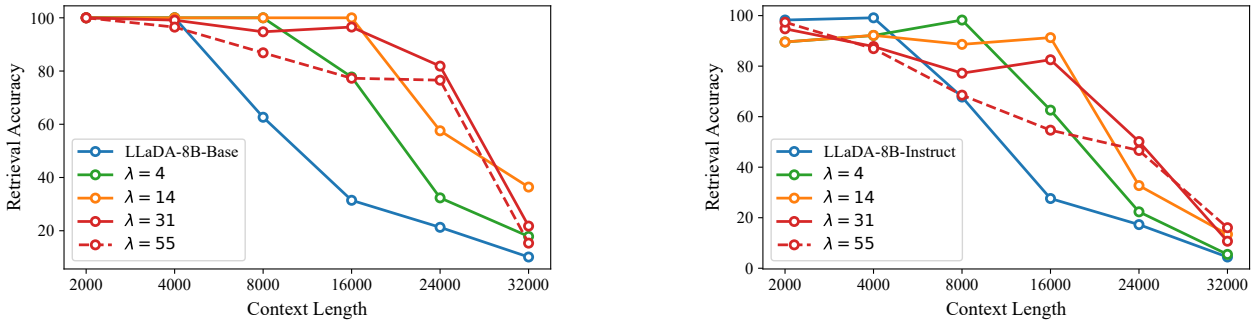
(b) LLaMA3-8B-Base

Figure 5: Visualization of the QK states from the final layer of LLaMA3-8B-Base and LLaDA-8B-Base for sample from the GovReport subsets in LongBench (Bai et al. 2023). The visualization uses a 2D t-SNE projection (Van der Maaten and Hinton 2008), with each token represented as a point and the position index shown via color changing.

100% recall across all depths within these ranges. As the context length increases beyond the extrapolation limit, the retrievable depth proportionally expands while maintaining the local-perception effect. The average depth score curves exhibit a right shift across different context lengths. When $\lambda = 31$, a lost-in-the-middle phenomenon (Liu et al. 2023a) similar to auto-regressive models emerges in intermediate depths, indicating that LongLLaDA approaches its practical extrapolation limit (bloc97 2023b). When $\lambda = 55$, further extrapolation is unachievable. Consequently, for RoPE-based dLLMs, ***NTK extrapolation and its scaling law remain applicable*** during inference.

Task-Driven Long-Context Capability Analysis

Regarding the downstream long-context performance of dLLMs and their difference from traditional auto-regressive LLMs, apart from the NIAH retrieval evaluation, we conduct comparative analyses across more benchmarks using LLaDA and LLaMA as examples. We first evaluate LLaDA-8B (Nie et al. 2025), LLaDA-1.5 (Zhu et al. 2025), and LLaMA3-8B (Meta 2024a), including pre-trained models and those employing NTK-based extrapolation during inference, with LongBench (Bai et al. 2023), in 4k and 8k context length, with the exceeding part being truncated from



(a) LLaDA-8B-Base

(b) LLaDA-8B-Instruct

Figure 6: NIAH Results of LLaDA-8B Series (Nie et al. 2025) with different RoPE scaling factor.

	4k							8k						
	SD	MD	Sum	ICL	Syn	Code	Avg	SD	MD	Sum	ICL	Syn	Code	Avg
LLaDA-8B-Base	15.1	18.4	32.0	42.0	54.7	59.6	34.1	13.9	13.1	30.8	40.7	56.0	57.4	32.4
$+\lambda = 4$	14.7	19.1	31.5	40.9	52.4	63.0	34.0	15.2	18.6	31.0	41.4	53.8	59.2	33.6
LLaDA-8B-Instruct	25.1	19.4	30.6	36.4	62.8	62.7	37.2	22.7	14.0	33.4	33.2	66.7	66.4	36.8
$+\lambda = 4$	22.1	19.8	33.0	38.0	63.3	65.0	**37.8**	23.4	19.8	**35.3**	39.8	**72.9**	67.3	40.6
LLaDA-1.5	24.4	19.4	31.6	33.5	**63.6**	66.7	37.6	22.6	14.5	33.4	33.0	67.6	67.6	37.1
$+\lambda = 4$	21.8	19.7	33.1	35.3	63.4	**67.3**	**37.8**	23.0	20.6	34.9	39.3	**72.9**	67.9	40.7
LLaMA3-8B-Base	17.2	18.7	25.0	41.7	47.6	66.5	33.6	18.2	18.3	26.1	**44.5**	49.6	**69.4**	35.1
LLaMA3-8B-Instruct	**31.9**	**26.1**	**33.6**	39.6	46.6	55.9	37.0	**37.5**	**28.3**	34.7	40.7	62.8	56.1	**41.9**

Table 1: Results of LLaDA-8B (Nie et al. 2025), LLaDA-1.5 (Zhu et al. 2025) and LLaMA3-8B (Meta 2024b) on LongBench (Bai et al. 2023) in 4k and 8k context length. SD, MD, Sum, and Syn stand for Single-Doc QA, Multi-Doc QA, Summarization, and Synthetic tasks, while Avg is the average score of all subtasks weighted by the evaluation data number.

the middle. For the summary tasks, the output length is 512, while for the others, the output length is 64. We still keep the sampling steps the same as the output length, and the block size to 64 for dLLMs. The results are shown in Table 1. Still, LLaDA can give a stable output and get a decent performance beyond the maximum supported context length. Moreover, we find that in all task domains besides synthetic tasks, the difference between LLaDA Series and LLaMA3 Series is relatively limited compared with the difference within LLaMA3 Series. Only in the synthetic domain does LLaDA Series outperform LLaMA3 Series consistently. This inspires us to conduct an in-depth discussion of dLLMs on the performance of the synthesis tasks.

We further the discussion with RULER benchmark (Hsieh et al. 2024), we compare LLaDA-8B (Nie et al. 2025), LLaDA-1.5 (Zhu et al. 2025), and LLaMA3-8B (Meta 2024a), at context lengths of 4k, 8k, and 16k. We set the block size and sampling steps to 64 for dLLMs. The results are shown in Table 2. First, consistent with the NIAH results, dLLMs can still produce valid outputs beyond their effective context length, while auto-regressive LLMs fails. Regarding task types, dLLMs achieve comparable results to auto-regressive LLMs on NIAH tasks, including Single-Key and its variants. However, dLLMs show significantly inferior performance in aggregation tasks, including Variable Tracing and Frequent or Common Word Extraction, where auto-regressive LLMs typically perform well. Surprisingly, on QA tasks, SQuAD and Hotpot, that challenge auto-regressive LLMs (Hsieh et al. 2024), dLLMs demonstrate superior capability. These observations reveal the distinctive characteristics of dLLMs in long-context tasks, that current dLLMs, like LLaDA, show comparable performance to the auto-regressive LLMs, like LLaMA3, in most task types, but ***underperform in aggregation tasks***, and ***outperform in synthetic QA tasks*** consistently.

Related Work

Large Language Diffusion Models Recently, Large Language Diffusion Models, or dLLMs, have become a widely discussed topic in NLP research. After the theoretical simplification (Sahoo et al. 2024; Ou et al. 2024; Shi et al. 2024) and empirical verification (He et al. 2022; Gong et al. 2024), researchers scale the size of dLLMs to billions of parameters (Nie et al. 2024, 2025; Ye et al. 2025) and demonstrate that dLLMs can achieve comparable results with more promising performance in the reversal

	4k				8k				16k			
	NIAH	AGG	QA	Avg	NIAH	AGG	QA	Avg	NIAH	AGG	QA	Avg
LLaDA-8B-Base	99.7	65.2	82.5	89.1	53.8	45.2	41.0	49.8	22.0	1.9	36.0	19.5
+ $\lambda = 4$	99.5	82.3	80.5	92.6	96.4	61.0	73.0	84.7	51.8	17.1	53.5	44.1
+ $\lambda = 14$	99.8	83.5	77.0	92.5	99.3	68.9	64.0	86.8	85.4	48.1	54.0	72.0
+ $\lambda = 31$	**100.0**	83.8	77.0	92.7	97.8	75.2	62.5	87.1	**97.2**	51.8	41.0	78.0
LLaDA-8B-Instruct	99.3	57.8	90.5	88.4	52.3	44.3	48.0	49.8	18.9	12.0	47.0	21.6
+ $\lambda = 4$	99.8	65.5	89.5	90.3	95.9	56.6	89.0	85.8	41.6	31.0	73.0	44.0
+ $\lambda = 14$	**100.0**	76.4	89.0	92.9	97.3	66.2	**89.5**	88.9	67.1	53.8	**84.5**	66.7
+ $\lambda = 31$	**100.0**	77.7	86.5	92.8	98.8	73.5	88.5	91.3	88.0	62.2	82.0	81.1
LLaDA-1.5	98.7	66.0	90.0	89.8	53.9	45.1	48.5	51.0	19.0	14.2	46.0	22.1
+ $\lambda = 4$	99.8	73.9	**91.0**	92.5	96.3	59.3	88.0	86.5	43.3	32.2	73.0	45.3
+ $\lambda = 14$	**100.0**	79.8	88.5	93.6	**99.9**	67.8	89.0	90.8	67.4	51.6	84.0	66.3
+ $\lambda = 31$	**100.0**	81.6	87.5	93.8	98.9	75.1	86.5	91.5	85.8	58.2	81.5	78.7
LLaMA3-8B-Base	99.8	98.1	67.5	94.4	99.6	93.5	63.0	92.5	0.0	0.0	0.0	0.0
+ $\lambda = 4$	99.9	**98.7**	65.0	94.2	99.8	**94.1**	59.0	92.2	97.0	86.6	54.5	88.1
+ $\lambda = 13$	99.5	98.6	66.0	94.1	99.1	94.0	59.0	91.8	93.8	**90.3**	56.0	87.2
LLaMA3-8B-Instruct	99.6	97.2	68.5	94.3	98.2	92.6	54.0	90.1	0.0	0.0	0.0	0.0
+ $\lambda = 4$	99.8	96.9	72.0	**94.9**	99.6	93.5	65.0	**92.8**	95.0	89.5	63.0	**88.8**
+ $\lambda = 13$	99.5	96.7	68.0	94.0	99.3	92.4	63.5	92.2	95.3	78.6	62.5	86.4

Table 2: Results of LLaDA-8B, LLaDA-1.5 and LLaMA3-8B on RULER (Hsieh et al. 2024) in 4k, 8k and 16k context length.

curse (Berglund et al. 2023). These immediately attract the attention of many more researchers. Significant research efforts have focused on adapting dLLMs for multimodality, such as MMaDA (Yang et al. 2025), LLaDA-V (You et al. 2025), and LaViDa (Li et al. 2025), applying them to reasoning tasks, such as d1 (Zhao et al. 2025), DCoLTHuang et al. (2025), and LLaDA-1.5 (Zhu et al. 2025), and optimizing their efficiency (Ma et al. 2025; Hu et al. 2025; Wu et al. 2025), including dKV-Cache (Ma et al. 2025), Dimple (Yu, Ma, and Wang 2025), dLLM-Cache (Liu et al. 2025b), FreeCache (Hu et al. 2025), Fast-dLLM (Wu et al. 2025), and so on. However, there is still no discussion on the long-context capability of diffusion LLMs.

Length Extrapolation in LLM Length extrapolation, or length generalization, or context extension, is an important issue for LLMs (Press, Smith, and Lewis 2022; Liu et al. 2025a). The mainstream extrapolation research mainly focuses on adjusting position embedding, especially the widely used RoPE (Su et al. 2021). For example, Linear PI (Chen et al. 2023) first achieves LLMs' length extrapolation by scaling position indices to the pre-training range with little fine-tuning. The NTK method (bloc97 2023b,a; Peng et al. 2023) then scales the rotary base in RoPE (Su et al. 2021) to achieve plug-and-play length extrapolation. Subsequently, amplifying the rotary base and training on longer lengths has become the dominant approach for length extrapolation (Rozière et al. 2023; Xiong et al. 2023; Liu et al. 2023b; Ding et al. 2024). In addition, ReRoPE (Su 2023), ReAttention (Liu et al. 2024b), and DCA (An et al. 2024a,b) also achieve plug-and-play extrapolation by limiting the relative position. In this paper, we still focus on the length extrapolation via NTK scaling (bloc97 2023b; Liu et al. 2023b) in the inference stage, and try to reveal and explain the similarities and differences in length extrapolation between diffusion-based and auto-regressive LLM.

Conclusion

In this work, we provide the first systematic analysis of long-context capabilities in diffusion LLMs. We demonstrate and analyze their characteristics for stable perplexity and local perception in direct context extrapolation from the perspective of the RoPE dynamic. Then, we propose LongLLaDA, which extends the context length in NTK scaling effectively without further training, and validate that the scaling laws still work for diffusion LLMs. Besides, we also show that diffusion LLMs match auto-regressive models on the average score of LongBench as well as the retrieval tasks, lag in aggregation tasks, but excel at QA in RULER evaluation. We hope our work can pave the foundation for future long-context research in diffusion LLMs.

Acknowledgments

This work was supported by the Shanghai Pilot Program for Basic Research - Fudan University 21TQ1400100 (22TQ018).

References

An, C.; Huang, F.; Zhang, J.; Gong, S.; Qiu, X.; Zhou, C.; and Kong, L. 2024a. Training-Free Long-Context Scaling of Large Language Models. *arXiv preprint arXiv:2402.17463*.

An, C.; Zhang, J.; Zhong, M.; Li, L.; Gong, S.; Luo, Y.; Xu, J.; and Kong, L. 2024b. Why Does the Effec-

tive Context Length of LLMs Fall Short? *arXiv preprint arXiv:2410.18745*.

Bachmann, G.; and Nagarajan, V. 2024. The pitfalls of next-token prediction. *arXiv preprint arXiv:2403.06963*.

Bai, Y.; Lv, X.; Zhang, J.; Lyu, H.; Tang, J.; Huang, Z.; Du, Z.; Liu, X.; Zeng, A.; Hou, L.; et al. 2023. Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*.

Berglund, L.; Tong, M.; Kaufmann, M.; Balesni, M.; Stickland, A. C.; Korbak, T.; and Evans, O. 2023. The Reversal Curse: LLMs trained on" A is B" fail to learn" B is A". *arXiv preprint arXiv:2309.12288*.

bloc97. 2023a. Dynamically Scaled RoPE further increases performance of long context LLaMA with zero fine-tuning.

bloc97. 2023b. NTK-Aware Scaled RoPE allows LLaMA models to have extended (8k+) context size without any fine-tuning and minimal perplexity degradation.

Chen, S.; Wong, S.; Chen, L.; and Tian, Y. 2023. Extending Context Window of Large Language Models via Positional Interpolation. *CoRR*, abs/2306.15595.

Contributors, O. 2023. OpenCompass: A Universal Evaluation Platform for Foundation Models.

Dao, T. 2023. FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning. *CoRR*, abs/2307.08691.

Ding, Y.; Zhang, L. L.; Zhang, C.; Xu, Y.; Shang, N.; Xu, J.; Yang, F.; and Yang, M. 2024. LongRoPE: Extending LLM Context Window Beyond 2 Million Tokens. *arXiv preprint arXiv:2402.13753*.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Dziri, N.; Lu, X.; Sclar, M.; Li, X. L.; Jiang, L.; Lin, B. Y.; Welleck, S.; West, P.; Bhagavatula, C.; Le Bras, R.; et al. 2023. Faith and fate: Limits of transformers on compositionality. *Advances in Neural Information Processing Systems*, 36: 70293–70332.

Gkamradt. 2023. Needle In A Haystack - Pressure Testing LLMs. https://github.com/gkamradt/LLMTest_NeedleInAHaystack.

Gong, S.; Agarwal, S.; Zhang, Y.; Ye, J.; Zheng, L.; Li, M.; An, C.; Zhao, P.; Bi, W.; Han, J.; et al. 2024. Scaling Diffusion Language Models via Adaptation from Autoregressive Models. *arXiv preprint arXiv:2410.17891*.

Han, C.; Wang, Q.; Xiong, W.; Chen, Y.; Ji, H.; and Wang, S. 2023. LM-Infinite: Simple On-the-Fly Length Generalization for Large Language Models. *CoRR*, abs/2308.16137.

He, Z.; Sun, T.; Wang, K.; Huang, X.; and Qiu, X. 2022. Diffusionbert: Improving generative masked language models with diffusion models. *arXiv preprint arXiv:2211.15029*.

Hsieh, C.-P.; Sun, S.; Kriman, S.; Acharya, S.; Rekesh, D.; Jia, F.; and Ginsburg, B. 2024. RULER: What's the Real Context Size of Your Long-Context Language Models? *arXiv preprint arXiv:2404.06654*.

Hu, Y.; Huang, Q.; Tao, M.; Zhang, C.; and Feng, Y. 2024. Can Perplexity Reflect Large Language Model's Ability in Long Text Understanding? *arXiv preprint arXiv:2405.06105*.

Hu, Z.; Meng, J.; Akhauri, Y.; Abdelfattah, M. S.; Seo, J.-s.; Zhang, Z.; and Gupta, U. 2025. Accelerating Diffusion Language Model Inference via Efficient KV Caching and Guided Diffusion. *arXiv preprint arXiv:2505.21467*.

Huang, Z.; Chen, Z.; Wang, Z.; Li, T.; and Qi, G.-J. 2025. Reinforcing the diffusion chain of lateral thought with diffusion language models. *arXiv preprint arXiv:2505.10446*.

Jin, M.; Mei, K.; Xu, W.; Sun, M.; Tang, R.; Du, M.; Liu, Z.; and Zhang, Y. 2025. Massive Values in Self-Attention Modules are the Key to Contextual Knowledge Understanding. *arXiv preprint arXiv:2502.01563*.

Li, M.; Zhang, S.; Liu, Y.; and Chen, K. 2024. Needlebench: Can llms do retrieval and reasoning in 1 million context window? *arXiv preprint arXiv:2407.11963*.

Li, S.; Kallidromitis, K.; Bansal, H.; Gokul, A.; Kato, Y.; Kozuka, K.; Kuen, J.; Lin, Z.; Chang, K.-W.; and Grover, A. 2025. LaViDa: A Large Diffusion Language Model for Multimodal Understanding. *arXiv preprint arXiv:2505.16839*.

Liu, N. F.; Lin, K.; Hewitt, J.; Paranjape, A.; Bevilacqua, M.; Petroni, F.; and Liang, P. 2023a. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.

Liu, X.; He, S.; Wang, Q.; Li, R.; Song, Y.; Liu, Z.; Huang, M.; Li, L.; Liu, Q.; Huang, Z.; Guo, Q.; He, Z.; and Qiu, X. 2024a. Beyond Homogeneous Attention: Memory-Efficient LLMs via Fourier-Approximated KV Cache. *arXiv preprint arXiv:2506.11886*.

Liu, X.; Li, R.; Guo, Q.; Liu, Z.; Song, Y.; Lv, K.; Yan, H.; Li, L.; Liu, Q.; and Qiu, X. 2024b. ReAttention: Training-Free Infinite Context with Finite Attention Scope. *arXiv preprint arXiv:2407.15176*.

Liu, X.; Li, R.; Huang, M.; Liu, Z.; Song, Y.; Guo, Q.; He, S.; Wang, Q.; Li, L.; Liu, Q.; et al. 2025a. Thus spake long-context large language model. *arXiv preprint arXiv:2502.17129*.

Liu, X.; Yan, H.; Zhang, S.; An, C.; Qiu, X.; and Lin, D. 2023b. Scaling Laws of RoPE-based Extrapolation. *CoRR*, abs/2310.05209.

Liu, Z.; Yang, Y.; Zhang, Y.; Chen, J.; Zou, C.; Wei, Q.; Wang, S.; and Zhang, L. 2025b. dLLM-Cache: Accelerating Diffusion Large Language Models with Adaptive Caching.

Ma, X.; Yu, R.; Fang, G.; and Wang, X. 2025. dkv-cache: The cache for diffusion language models. *arXiv preprint arXiv:2505.15781*.

Men, X.; Xu, M.; Wang, B.; Zhang, Q.; Lin, H.; Han, X.; and Chen, W. 2024. Base of rope bounds context length. *arXiv preprint arXiv:2405.14591*.

Meta, A. 2024a. Introducing meta llama 3: The most capable openly available llm to date. *Meta AI*.

Meta, A. 2024b. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models. *Meta AI*.

Nie, S.; Zhu, F.; Du, C.; Pang, T.; Liu, Q.; Zeng, G.; Lin, M.; and Li, C. 2024. Scaling up Masked Diffusion Models on Text. *arXiv preprint arXiv:2410.18514*.

Nie, S.; Zhu, F.; You, Z.; Zhang, X.; Ou, J.; Hu, J.; Zhou, J.; Lin, Y.; Wen, J.-R.; and Li, C. 2025. Large language diffusion models. *arXiv preprint arXiv:2502.09992*.

Ou, J.; Nie, S.; Xue, K.; Zhu, F.; Sun, J.; Li, Z.; and Li, C. 2024. Your absorbing discrete diffusion secretly models the conditional distributions of clean data. *arXiv preprint arXiv:2406.03736*.

Peng, B.; Quesnelle, J.; Fan, H.; and Shippole, E. 2023. YaRN: Efficient Context Window Extension of Large Language Models. *CoRR*, abs/2309.00071.

Press, O.; Smith, N. A.; and Lewis, M. 2022. Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Rozière, B.; Gehring, J.; Gloeckle, F.; Sootla, S.; Gat, I.; Tan, X. E.; Adi, Y.; Liu, J.; Remez, T.; Rapin, J.; Kozhevnikov, A.; Evtimov, I.; Bitton, J.; Bhatt, M.; Canton-Ferrer, C.; Grattafiori, A.; Xiong, W.; Défossez, A.; Copet, J.; Azhar, F.; Touvron, H.; Martin, L.; Usunier, N.; Scialom, T.; and Synnaeve, G. 2023. Code Llama: Open Foundation Models for Code. *CoRR*, abs/2308.12950.

Sahoo, S.; Arriola, M.; Schiff, Y.; Gokaslan, A.; Marroquin, E.; Chiu, J.; Rush, A.; and Kuleshov, V. 2024. Simple and effective masked diffusion language models. *Advances in Neural Information Processing Systems*, 37: 130136–130184.

Shi, J.; Han, K.; Wang, Z.; Doucet, A.; and Titsias, M. 2024. Simplified and generalized masked diffusion for discrete data. *Advances in neural information processing systems*, 37: 103131–103167.

Su, J. 2023. ReRoPE: Rectified Rotary Position Embeddings.

Su, J.; Lu, Y.; Pan, S.; Wen, B.; and Liu, Y. 2021. RoFormer: Enhanced Transformer with Rotary Position Embedding. *CoRR*, abs/2104.09864.

Sun, T.; Zhang, X.; He, Z.; Li, P.; Cheng, Q.; Liu, X.; Yan, H.; Shao, Y.; Tang, Q.; Zhang, S.; Zhao, X.; Chen, K.; Zheng, Y.; Zhou, Z.; Li, R.; Zhan, J.; Zhou, Y.; Li, L.; Yang, X.; Wu, L.; Yin, Z.; Huang, X.; Jiang, Y.-G.; and Qiu, X. 2024. MOSS: An Open Conversational Large Language Model. *Machine Intelligence Research*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. *CoRR*, abs/2302.13971.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Wang, J.; Ji, T.; Wu, Y.; Yan, H.; Gui, T.; Zhang, Q.; Huang, X.; and Wang, X. 2024. Length Generalization of Causal Transformers without Position Encoding. *arXiv preprint arXiv:2404.12224*.

Wei, X.; Liu, X.; Zang, Y.; Dong, X.; Zhang, P.; Cao, Y.; Tong, J.; Duan, H.; Guo, Q.; Wang, J.; Qiu, X.; and Lin, D. 2025. VideoRoPE: What Makes for Good Video Rotary Position Embedding? *arXiv preprint arXiv:2502.05173*.

Wu, C.; Zhang, H.; Xue, S.; Liu, Z.; Diao, S.; Zhu, L.; Luo, P.; Han, S.; and Xie, E. 2025. Fast-dLLM: Training-free Acceleration of Diffusion LLM by Enabling KV Cache and Parallel Decoding. *arXiv preprint arXiv:2505.22618*.

Xiong, W.; Liu, J.; Molybog, I.; Zhang, H.; Bhargava, P.; Hou, R.; Martin, L.; Rungta, R.; Sankararaman, K. A.; Oguz, B.; et al. 2023. Effective long-context scaling of foundation models. *arXiv preprint arXiv:2309.16039*.

Yang, A.; Yang, B.; Hui, B.; Zheng, B.; Yu, B.; Zhou, C.; Li, C.; Li, C.; Liu, D.; Huang, F.; Dong, G.; Wei, H.; Lin, H.; Tang, J.; Wang, J.; Yang, J.; Tu, J.; Zhang, J.; Ma, J.; Xu, J.; Zhou, J.; Bai, J.; He, J.; Lin, J.; Dang, K.; Lu, K.; Chen, K.; Yang, K.; Li, M.; Xue, M.; Ni, N.; Zhang, P.; Wang, P.; Peng, R.; Men, R.; Gao, R.; Lin, R.; Wang, S.; Bai, S.; Tan, S.; Zhu, T.; Li, T.; Liu, T.; Ge, W.; Deng, X.; Zhou, X.; Ren, X.; Zhang, X.; Wei, X.; Ren, X.; Fan, Y.; Yao, Y.; Zhang, Y.; Wan, Y.; Chu, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Fan, Z. 2024. Qwen2 Technical Report. *arXiv preprint arXiv:2407.10671*.

Yang, L.; Tian, Y.; Li, B.; Zhang, X.; Shen, K.; Tong, Y.; and Wang, M. 2025. Mmada: Multimodal large diffusion language models. *arXiv preprint arXiv:2505.15809*.

Ye, J.; Gao, J.; Gong, S.; Zheng, L.; Jiang, X.; Li, Z.; and Kong, L. 2024. Beyond autoregression: Discrete diffusion for complex reasoning and planning. *arXiv preprint arXiv:2410.14157*.

Ye, J.; Xie, Z.; Zheng, L.; Gao, J.; Wu, Z.; Jiang, X.; Li, Z.; and Kong, L. 2025. Dream 7B.

You, Z.; Nie, S.; Zhang, X.; Hu, J.; Zhou, J.; Lu, Z.; Wen, J.-R.; and Li, C. 2025. LLaDA-V: Large Language Diffusion Models with Visual Instruction Tuning. *arXiv preprint arXiv:2505.16933*.

Yu, R.; Ma, X.; and Wang, X. 2025. Dimple: Discrete Diffusion Multimodal Large Language Model with Parallel Decoding. *arXiv preprint arXiv:2505.16990*.

Zandieh, A.; Han, I.; Mirrokni, V.; and Karbasi, A. 2024. SubGen: Token Generation in Sublinear Time and Memory. *arXiv preprint arXiv:2402.06082*.

Zhao, S.; Gupta, D.; Zheng, Q.; and Grover, A. 2025. d1: Scaling reasoning in diffusion large language models via reinforcement learning. *arXiv preprint arXiv:2504.12216*.

Zhu, F.; Wang, R.; Nie, S.; Zhang, X.; Wu, C.; Hu, J.; Zhou, J.; Chen, J.; Lin, Y.; Wen, J.-R.; et al. 2025. LLaDA 1.5: Variance-Reduced Preference Optimization for Large Language Diffusion Models. *arXiv preprint arXiv:2505.19223*.