

# Lost in Diffusion: Uncovering Hallucination Patterns and Failure Modes in Diffusion Large Language Models

Zhengen Guo<sup>1,2</sup>, Fei Tan<sup>1\*</sup>

<sup>1</sup>East China Normal University

<sup>2</sup>Zhejiang University of Technology

zhengen.guo@zjut.edu.cn, ftan@mail.ecnu.edu.cn

## Abstract

While Diffusion Large Language Models (dLLMs) have emerged as a promising non-autoregressive paradigm comparable to autoregressive (AR) models, their faithfulness, specifically regarding hallucination, remains largely underexplored. To bridge this gap, we present the first controlled comparative study to evaluate hallucination patterns in dLLMs. Our results demonstrate that current dLLMs exhibit a higher propensity for hallucination than AR counterparts controlled for architecture, scale, and pre-training weights. Furthermore, an analysis of inference-time compute reveals divergent dynamics: while *quasi-autoregressive* generation suffers from early saturation, *non-sequential* decoding unlocks potential for continuous refinement. Finally, we identify distinct failure modes unique to the diffusion process, including *premature termination*, *incomplete denoising*, and *context intrusion*. Our findings underscore that although dLLMs have narrowed the performance gap on general tasks, their distinct hallucination mechanisms pose a critical challenge to model reliability. Our code is available at <https://github.com/ZeroLoss-Lab/Lost-in-Diffusion>

## 1 Introduction

While Auto-regressive Large Language Models (AR-LLMs) have dominated text generation in recent years (OpenAI, 2022; Grattafiori et al., 2024; DeepSeek-AI et al., 2025), a growing body of research is exploring Diffusion Large Language Models (dLLMs) as a potential non-autoregressive generation paradigm (Google DeepMind, 2024; Inception Labs, 2025; Song et al., 2025). Diverging from the sequential token generation of AR models, dLLMs generate tokens in parallel via iterative denoising, offering inherent advantages in global planning and bidirectional visibility (Li et al., 2022;

Gong et al., 2023). Recent open-source dLLMs, such as LLaDA series (Nie et al., 2025; Bie et al., 2025), Dream (Ye et al., 2025), SDAR (Cheng et al., 2025) and Fast-dLLM v2 (Wu et al., 2025), have achieved performance parity with leading AR-LLMs at comparable scales, signaling that dLLMs are approaching practical utility.

Despite these advancements, the trustworthiness of dLLMs, specifically regarding hallucination, remains an open question. Hallucination, defined as inconsistency between model output and source content (i.e., input context or training data) (Ji et al., 2023; Bang et al., 2025), is well-documented in AR-LLMs (Huang et al., 2023; Kalai et al., 2025). These inconsistencies typically stem from sequential error propagation and unidirectional attention mechanisms, phenomena respectively known as the snowballing effect (Zhang et al., 2024) and the reversal curse (Berglund et al., 2024). Theoretically, the global context planning and bidirectional visibility of dLLMs offer a mitigation by facilitating retrospective refinement. Conversely, the stochastic nature of the diffusion process introduces intrinsic noise, potentially exacerbating decoding randomness, a factor also widely recognized as a root cause of hallucination (Lee et al., 2022). As dLLMs move towards widespread adoption, a critical question emerges: *Does the diffusion mechanism mitigate or exacerbate hallucination?*

To address this, we conduct the first systematic benchmarking of dLLMs against their AR counterparts regarding hallucination. Our contributions are threefold:

- **Benchmarking the Mechanism:** We provide the first controlled pairwise comparison, revealing that current dLLMs exhibit a significantly higher propensity for hallucination than AR-LLMs under comparable settings.
- **Dynamics of Inference Compute:** We find that the efficacy of denoising is contingent upon

\* Corresponding author

the decoding strategy. Specifically, while *quasi-autoregressive* generation suffers from early saturation, *non-sequential* strategies unlock the potential for continuous refinement via increased inference steps.

- **Diffusion Failure Modes:** We identify and analyze failure modes unique to the diffusion process, including *premature termination*, *incomplete denoising*, and *context intrusion*, offering insights into their underlying mechanisms.

## 2 Related Work

### 2.1 Diffusion Large Language Models

Early text diffusion attempts (Li et al., 2022; Zhang et al., 2023; Lin et al., 2023; Gulrajani and Hashimoto, 2023) struggled with the discrete nature of language. However, recent advancements in masked diffusion (Lou et al., 2023; Shi et al., 2024; Sahoo et al., 2024) have enabled significant scaling. Representative open-source models, including the LLaDA series (Nie et al., 2025; Bie et al., 2025), Dream (Ye et al., 2025), SDAR (Cheng et al., 2025), and Fast-dLLM v2 (Wu et al., 2025), have achieved performance parity with AR-LLMs on general benchmarks. Despite this progress, current research predominantly prioritizes architectural optimization and inference acceleration. While Chang et al. (2025) recently initiated the study of hallucination detection in dLLMs, a systematic comparative benchmarking of their hallucination patterns against AR baselines remains absent.

### 2.2 Hallucination in LLMs

Hallucination constitutes a primary obstacle to the reliable deployment of LLMs (Ji et al., 2023; Huang et al., 2023). In AR-LLMs, this phenomenon is often traced to the sequential decoding mechanism. Specifically, the snowballing effect (Zhang et al., 2024) illustrates how early sequence errors propagate and amplify, while the reversal curse (Berglund et al., 2024) exposes the structural inability of auto-regressive objectives to handle bidirectional knowledge retrieval. Additionally, the stochastic nature of sampling correlates with decoding randomness and factual inconsistencies (Lee et al., 2022). While these patterns are well-charted in AR models, dLLMs exhibit distinct failure modes, such as interleaving hallucinations

(Chang et al., 2025). Motivated by these distinctions, we present the first controlled study to quantify hallucinations in dLLMs in comparison to AR baselines.

## 3 Methodology

To assess the hallucination patterns of dLLMs, we propose a comparative framework designed to minimize confounding factors such as model scale and training data distribution. While complete isolation of the generation mechanism is inherently challenging, our methodology represents a best-effort approach to disentangling it from these variables.

### 3.1 Pairwise Comparison Framework

Direct comparisons between dLLMs and arbitrary AR-LLMs are often skewed by discrepancies in pre-training corpora and model capacities, as these discrepancy can substantially affect model performance even under identical task settings (Wang et al., 2024; Lu et al., 2023). To mitigate this, we devise a pairwise comparison strategy comprising two distinct control groups, which aims to maximize comparability by aligning parametric knowledge and scale to the extent possible:

- **Group I: Architectural Alignment.** We benchmark LLaDA-8B (Nie et al., 2025) against LLaMA-3-8B (Grattafiori et al., 2024). Both models share similar backbone architectures and parameter scales, and exhibit comparable performance on general benchmarks, allowing us to isolate differences induced specifically by the diffusion modeling approach.
- **Group II: Parametric Alignment.** We pair Dream-7B (Ye et al., 2025) with Qwen2.5-7B (Qwen et al., 2025). Crucially, as Dream is initialized directly from Qwen weights, this pairing offers the control over parametric knowledge. Any divergence in hallucination rates can thus be primarily attributed to the diffusion generation process rather than knowledge disparity.

We prioritize pre-trained checkpoints for our primary analysis to isolate the generation paradigm from the noise introduced by post-training alignment, although instruction-tuned variants are discussed in Appendix C. Note that only the instruction-tuned checkpoints are publicly released for LLaDA2.0 and Fast-dLLM v2.

### 3.2 Canonical Diffusion Inference

Unlike the sequential dependence of AR-LLMs ( $x_t \sim p(x_t|x_{<t})$ ), dLLMs generate the entire sequence in parallel via iterative denoising. To fully characterize this mechanism, we adopt a *canonical diffusion* setting, bypassing semi-autoregressive or block-based acceleration methods (Nie et al., 2025; Wu et al., 2025).

For a target sequence of length  $L$ , we set the number of denoising steps  $T$  equal to the sequence length ( $T = L$ ). This configuration maximizes the model’s capacity for iterative refinement. To ensure reproducibility, we set the temperature to zero. We employ standard decoding strategies: high-confidence decoding for LLaDA and minimum entropy decoding for Dream.

### 3.3 Hallucination Assessment Protocol

We adapt the HalluLens (Bang et al., 2025) benchmark to evaluate *Extrinsic Hallucination* (Ji et al., 2023), spanning three tasks: *PreciseWikiQA*, *LongWiki*, and *NonExistentRefusal*. Please refer to Appendix A for detailed configurations.

To strictly evaluate the generation mechanism, we exclude *Intrinsic Hallucination* tasks (e.g., summarization). These tasks heavily rely on instruction-following capabilities typically acquired during supervised fine-tuning, which would introduce confounding factors unrelated to the pre-training generation objective. Since our benchmark relies on an automatic LLM-based evaluator, we additionally perform human annotation on a stratified subset to validate its reliability. Detailed information can be found in Appendix B.

## 4 Experiments and Analysis

We systematically evaluate dLLMs against their AR counterparts across three critical dimensions: precise knowledge recall, long-form factual consistency, and knowledge boundary detection. Detailed experiments settings can be found in Appendix A. The comparative results are presented in Table 1.

### 4.1 Main Results

Basically, dLLMs consistently underperform their AR counterparts across all three tasks. In *PreciseWikiQA*, dLLMs struggle to anchor generation to facts, with Dream-7B exhibiting a higher Hallucination Rate (92.54%) than Qwen2.5-7B (89.06%) and LLaDA-8B showing a significantly

lower Correct Rate (3.92%) compared to LLaMA-3-8B (10.30%). In *LongWiki*, while LLaDA-8B achieves competitive Recall, both dLLMs suffer from lower Precision, resulting in inferior overall F1@32 scores. In *NonExistentRefusal*, dLLMs reveal a severe inability to refuse invalid queries; notably, Dream-7B fails to recognize non-existent entities in 98.50% of cases, lagging behind the AR baseline.

This performance degradation can be partially attributed to the absence of a mature re-masking strategy. While dLLMs theoretically offer global planning, current architectures lack the mechanism to retrospectively correct tokens once they are denoised, even if subsequent context renders them implausible (Kang et al., 2026). Consequently, the snowballing effect is not eliminated but transformed: early denoising errors become entrenched, corrupting the trajectory of the entire generation process. Furthermore, the reversal curse is not fully mitigated; although bidirectional attention enables the detection of inconsistent tokens, the model remains unable to rectify them. Compounded by the immaturity of current training techniques tailored for dLLMs, these structural deficiencies ultimately exacerbate the issue of hallucinations. More recently, a number of approaches (Wang et al., 2025; Zhang et al., 2026; Huang et al., 2026; Bie et al., 2026) have been proposed to mitigate token-level errors. Evaluating these methods within our framework is an important direction that we leave to future work.

### 4.2 Dynamics of Inference Compute

A theoretical advantage of dLLMs is the ability to trade compute for quality via iterative refinement. To investigate whether increased inference compute mitigates hallucination, we evaluate LLaDA-8B and Dream-7B on the *LongWiki* task across exponentially increasing denoising steps  $T \in \{128, 256, 512, 1024\}$ .

As shown in Figure 1, the two architectures exhibit starkly divergent behaviors. LLaDA-8B displays early saturation, with performance metrics remaining statistically stagnant across all steps (F1@32  $\approx$  0.27). We attribute this to its *quasi-autoregressive* generation order. Despite theoretically possessing bidirectional visibility, LLaDA-8B is constrained by a linear noise scheduler and high-confidence decoding that enforces a predominant left-to-right generation flow (see Appendix E.1). Consequently, it inherits the structural rigid-

Model	Precise WikiQA			Prec. $\uparrow$	LongWiki		NonexistentRefusal FA $\downarrow$
	FRR $\downarrow$	HR $\downarrow$	CR $\uparrow$		Rec.@32 $\uparrow$	F1@32 $\uparrow$	
LLaMA-3-8B	28.72	<b>85.94</b>	<b>10.30</b>	<b>0.408</b>	0.293	<b>0.306</b>	<b>73.35</b>
LLaDA-8B	<b>21.40</b>	95.13	3.92	0.271	<b>0.306</b>	0.272	87.10
Qwen2.5-7B	<b>18.12</b>	<b>89.06</b>	<b>9.06</b>	<b>0.376</b>	<b>0.441</b>	<b>0.387</b>	<b>94.05</b>
Dream-7B	26.96	92.54	6.04	0.345	0.400	0.340	98.50

Table 1: Performance comparison across three extrinsic hallucination tasks. Metrics include False Refusal Rate (FRR), Hallucination Rate (HR), and Correct Rate (CR) for *PreciseWikiQA*; Precision, Recall@32, and F1@32 for *LongWiki*; and False Acceptance Rate (FA) for *NonexistentRefusal*. **Bold** indicates the better performance within each control group.

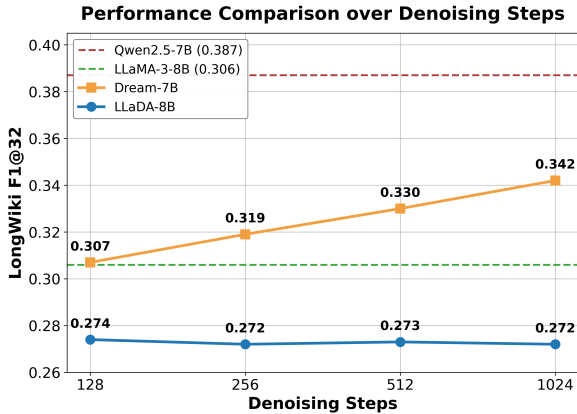


Figure 1: F1@32 trends across exponentially increasing denoising steps. While LLaDA saturates early, Dream benefits from increased inference compute.

ity of AR models while incurring additional diffusion noise. Increasing steps merely reduces prediction granularity without fundamentally altering sequential dependencies, explaining the negligible marginal utility of compute scaling.

In contrast, Dream-7B demonstrates positive scaling dynamics, exhibiting monotonic improvements as  $T$  increases. This is driven by its minimum entropy decoding, which enables genuine *non-sequential* refinement. Dream-7B breaks the left-to-right paradigm, more effectively exploiting bidirectional self-attention (see Appendix E.2). This allows for a significant reduction in hallucination rates as compute budget increases. While this flexibility unlocks the potential for inference-time scaling, it also introduces distinct stability risks. Our analysis reveals that departing from sequential constraints is closely associated with a higher incidence of *Context Intrusion*, shifting the distribution of failure modes compared to LLaDA. We categorize these unique failure modes below.

## 5 Diffusion Failure Modes

Unlike the sequential drift observed in AR-LLMs, dLLMs exhibit unique failure modes stemming from their non-autoregressive nature. Through manual inspection, we identify three distinct patterns (see Appendix D for case studies):

**Premature Termination:** The model predicts the End-of-Sequence (EOS) token or inserts rigid separators abruptly. This typically occurs when disjoint segments (e.g., the prefix and suffix) decoded independently fail to align syntactically. Lacking a mechanism to bridge this coherence gap, the model is forced to terminate or insert broken structures to resolve the conflict.

**Incomplete Denoising:** In scenarios involving rare entities, the model may anchor on nonsensical tokens in the later part of the sequence. As bidirectional attention attempts to rationalize the connection between the prompt and these chaotic anchors, the entire generation collapses. This results in residual traces or hollow symbols that imitate the structure of language while lacking any substantive content.

**Context Intrusion:** We observe instances where the model abruptly pivots to unrelated topics, such as mathematical reasoning or coding tutorials. This phenomenon is triggered when the model occasionally denoises a high-frequency token (e.g., a number or code keyword) in a future position. The bidirectional attention then forces the generation to construct a logical path to this spurious anchor, effectively hijacking the original query trajectory.

We further quantify the prevalence of diffusion-specific failure modes through human annotation. We annotate 200 hallucinated outputs from dLLMs and assign each example failure mode label among *Premature Termination* (PT), *Incomplete Denoising* (ID), *Context Intrusion* (CI). As shown in Table 2, these failure modes occur at non-trivial fre-

Model	PT	ID	CI
LLaDA-8B	18.0	60.0	38.0
Dream-7B	13.0	44.0	58.0

Table 2: Frequency (%) of human-annotated failure modes on broken outputs from diffusion language models. Percentages do not sum to 100 because some examples exhibit multiple failure modes.

quencies in both dLLMs, with *Incomplete Denoising* more common in LLaDA-8B and *Context Intrusion* more frequent in Dream-7B.

These failure modes expose the inherent cost of the unconstrained generation paradigm. Unlike AR models that only ensure continuity with the past, dLLMs must satisfy global consistency across the entire sequence. When early-decoded tokens conflict, the current absence of iterative editing mechanisms, such as insertion, deletion, or re-masking, leaves the model with no recourse but to force a broken interpretation. This suggests that future dLLMs require dynamic sequence editing capabilities akin to human writing to fully realize the potential of non-autoregressive generation.

## 6 Conclusion

We present a controlled evaluation revealing that current dLLMs are generally more prone to hallucination than auto-regressive baselines. Our analysis of inference dynamics reveals a dichotomy: while *quasi-autoregressive* dLLMs suffer from early saturation, *non-sequential* decoding unlocks the potential for continuous refinement but introduces stability risks. Coupled with unique failure modes like *premature termination*, these findings underscore that achieving reliable non-autoregressive generation requires more than simple denoising.

## Limitations

While our study offers a foundational benchmarking of dLLM hallucinations, we acknowledge several limitations.

**Limits of Confounders Isolation** While our comparative framework strives to isolate the impact of the generation paradigm to the maximum extent possible, achieving a strictly controlled environment is theoretically elusive. Even the most lightweight adaptation required to enable diffusion capabilities necessitates weight updates, which inevitably shift the model’s underlying parametric

knowledge. Consequently, such shifts may introduce subtle confounders in hallucination assessment. Our approach thus represents a best-effort approximation within these inherent constraints.

**Scope of Inference Mechanism** In this work, we strictly adopt a *canonical diffusion* setting to characterize the native generation behavior. Consequently, our findings may not fully generalize to accelerated inference methods (e.g., semi-autoregressive or block-based decoding) where auto-regressive guidance is reintroduced to stabilize generation. Furthermore, given that these models were not explicitly trained to perform denoising over such extensive block lengths (1,024), our observed hallucination levels may not faithfully reflect those encountered in real-world deployments. Accordingly, our findings may not directly transfer to alternative inference-time control or verification schemes, which have been shown to reshape generation trade-offs in other decoding paradigms (Leviathan et al., 2023; Ding et al., 2025; Yan et al., 2025).

**Exclusion of Alignment Tuning** As mentioned in Section 3.1, we prioritize pre-trained models to isolate the generative paradigm from post-training noise. As noted in our analysis in Appendix C, instruction-tuned dLLMs currently exhibit high variance in refusal behaviors (e.g., the over-refusal observed in Dream-7B), which confounds the assessment of intrinsic hallucination propensities. For this reason, we excluded instruction-tuned models from our primary analysis; by the same logic, we also omitted *Intrinsic Hallucination* tasks, as they rely heavily on instruction-following capabilities acquired during alignment. However, we acknowledge that a more in-depth evaluation of the impact of alignment tuning on hallucination levels is certainly feasible, and we leave this extensive investigation to future work.

## Acknowledgments

This work is supported by the East China Normal University “Artificial Intelligence” Seed Grant Program (40500-20101-222438) and the East China Normal University “Discipline Advancement Program” (40600-515100-25001/002/015).

## References

Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. 2025. [HalluLens: LLM](#)

- hallucination benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL 2025)*, pages 24128–24156, Vienna, Austria. Association for Computational Linguistics.
- Lukas Berglund, Meg Tong, Maximilian Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2024. [The reversal curse: LLMs trained on “a is b” fail to learn “b is a”](#). In *The Twelfth International Conference on Learning Representations (ICLR 2024)*.
- Tiwei Bie, Maosong Cao, Xiang Cao, Bingsen Chen, Fuyuan Chen, Kun Chen, Lun Du, Daozhuo Feng, Haibo Feng, Mingliang Gong, Zhuocheng Gong, Yanmei Gu, Jian Guan, Kaiyuan Guan, Hongliang He, Zenan Huang, Juyong Jiang, Zhonghui Jiang, Zhenzhong Lan, and 31 others. 2026. [Llada2.1: Speeding up text diffusion via token editing](#). *Preprint*, arXiv:2602.08676.
- Tiwei Bie, Maosong Cao, Kun Chen, Lun Du, Mingliang Gong, Zhuochen Gong, Yanmei Gu, Jiaqi Hu, Zenan Huang, Zhenzhong Lan, Chengxi Li, Chongxuan Li, Jianguo Li, Zehuan Li, Huabin Liu, Ling Liu, Guoshan Lu, Xiaocheng Lu, Yuxin Ma, and 12 others. 2025. [Llada2.0: Scaling up diffusion language models to 100b](#). *Preprint*, arXiv:2512.15745.
- Shenxu Chang, Junchi Yu, Weixing Wang, Yongqiang Chen, Jialin Yu, Philip Torr, and Jindong Gu. 2025. [Tracedet: Hallucination detection from the decoding trace of diffusion large language models](#). *Preprint*, arXiv:2510.01274.
- Shuang Cheng, Yi Han Bian, Dawei Liu, Linfeng Zhang, Qian Yao, Zhongbo Tian, Wenhai Wang, Qipeng Guo, Kai Chen, Biqing Qi, and Bowen Zhou. 2025. [Sdar: A synergistic diffusion-autoregression paradigm for scalable sequence generation](#). *Preprint*, arXiv:2510.06303.
- Euirim Choi. 2023. Goodwiki dataset. <https://www.github.com/euirim/goodwiki>. Accessed: 2025-11-23.
- Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. [Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities](#). *Preprint*, arXiv:2507.06261.
- DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. [Deepseek-v3 technical report](#). *Preprint*, arXiv:2412.19437.
- Chuanghao Ding, Jiaping Wang, Ziqing Yang, Xiaoliang Wang, Dahua Lin, Cam-Tu Nguyen, and Fei Tan. 2025. [Consultant decoding: Yet another synergistic mechanism](#). In *Findings of the Association for Computational Linguistics: ACL 2025 (ACL 2025 Findings)*, pages 15438–15452, Vienna, Austria. Association for Computational Linguistics.
- Shanshan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and Lingpeng Kong. 2023. [Diffuseq: Sequence to sequence text generation with diffusion models](#). In *The Eleventh International Conference on Learning Representations (ICLR 2023)*.
- Google DeepMind. 2024. Our state-of-the-art, experimental text diffusion model. <https://deepmind.google/models/gemini-diffusion/>. Accessed: 2025-12-09.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Ishaan Gulrajani and Tatsunori Hashimoto. 2023. [Likelihood-based diffusion language models](#). In *The Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS 2023)*.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. [A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions](#). *ACM Transactions on Information Systems*, 43:1 – 55.
- Zemin Huang, Yuhang Wang, Zhiyang Chen, and Guojun Qi. 2026. [Don’t settle too early: Self-reflective remasking for diffusion language models](#). In *The Fourteenth International Conference on Learning Representations (ICLR 2026)*.
- Inception Labs. 2025. Introducing Mercury, the world’s first commercial-scale diffusion large language model. <https://www.inceptionlabs.ai/blog/introducing-mercury>. Accessed: 2025-12-09.
- Integrated Taxonomic Information System. 2025. [Integrated taxonomic information system \(ITIS\) on-line database](#). Retrieved November 23, 2025.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. [Survey of hallucination in natural language generation](#). *ACM Comput. Surv.*, 55(12).
- Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. 2025. [Why language models hallucinate](#). *Preprint*, arXiv:2509.04664.

- Wonjun Kang, Kevin Galim, Seunghyuk Oh, Minjae Lee, Yuchen Zeng, Shuibai Zhang, Coleman Richard Charles Hooper, Yuezhou Hu, Hyung Il Koo, Nam Ik Cho, and Kangwook Lee. 2026. [Parallel-bench: Understanding the trade-offs of parallel decoding in diffusion LLMs](#). In *The Fourteenth International Conference on Learning Representations (ICLR 2026)*.
- Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale Fung, Mohammad Shoeybi, and Bryan Catanzaro. 2022. [Factuality enhanced language models for open-ended text generation](#). In *The Thirty-Sixth Annual Conference on Neural Information Processing Systems (NeurIPS 2022)*.
- Yaniv Leviathan, Matan Kalman, and Yossi Matias. 2023. [Fast inference from transformers via speculative decoding](#). In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, Proceedings of Machine Learning Research, pages 19274–19286. PMLR.
- Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori Hashimoto. 2022. [Diffusion-LM improves controllable text generation](#). In *The Thirty-Sixth Annual Conference on Neural Information Processing Systems (NeurIPS 2022)*.
- Zhenghao Lin, Yeyun Gong, Yelong Shen, Tong Wu, Zhihao Fan, Chen Lin, Nan Duan, and Weizhu Chen. 2023. [Text generation with diffusion language models: A pre-training approach with continuous paragraph denoise](#). In *The Fortieth International Conference on Machine Learning (ICML 2023)*.
- Ling Team, Ang Li, Ben Liu, Binbin Hu, Bing Li, Bingwei Zeng, Borui Ye, Caizhi Tang, Changxin Tian, Chao Huang, Chao Zhang, Chen Qian, Chenchen Ju, Chenchen Li, Chengfu Tang, Chilin Fu, Chunshao Ren, Chunwei Wu, Cong Zhang, and 123 others. 2025. [Every activation boosted: Scaling general reasoner to 1 trillion open language foundation](#). Preprint, arXiv:2510.22115.
- Aaron Lou, Chenlin Meng, and Stefano Ermon. 2023. [Discrete diffusion modeling by estimating the ratios of the data distribution](#). In *The Fortieth International Conference on Machine Learning (ICML 2023)*.
- Jinghui Lu, Dongsheng Zhu, Weidong Han, Rui Zhao, Brian Mac Namee, and Fei Tan. 2023. [What makes pre-trained language models better zero-shot learners?](#) In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL 2023)*, pages 2288–2303, Toronto, Canada. Association for Computational Linguistics.
- Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. [FActScore: Fine-grained atomic evaluation of factual precision in long form text generation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*, pages 12076–12100, Singapore. Association for Computational Linguistics.
- Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. 2025. [Large language diffusion models](#). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS 2025)*.
- OpenAI. 2022. ChatGPT: Optimizing language models for dialogue. <https://openai.com/blog/chatgpt/>. Accessed: 2025-12-09.
- Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. [Qwen2.5 technical report](#). Preprint, arXiv:2412.15115.
- Subham Sekhar Sahoo, Marianne Arriola, Aaron Gokaslan, Edgar Mariano Marroquin, Alexander M Rush, Yair Schiff, Justin T Chiu, and Volodymyr Kuleshov. 2024. [Simple and effective masked diffusion language models](#). In *The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS 2024)*.
- Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis Titsias. 2024. [Simplified and generalized masked diffusion for discrete data](#). In *The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS 2024)*.
- Yixiao Song, Yekyung Kim, and Mohit Iyyer. 2024. [VeriScore: Evaluating the factuality of verifiable claims in long-form text generation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024 (EMNLP 2024 Findings)*, pages 9447–9474, Miami, Florida, USA. Association for Computational Linguistics.
- Yuxuan Song, Zheng Zhang, Cheng Luo, Pengyang Gao, Fan Xia, Hao Luo, Zheng Li, Yuehang Yang, Hongli Yu, Xingwei Qu, Yuwei Fu, Jing Su, Ge Zhang, Wenhao Huang, Mingxuan Wang, Lin Yan, Xiaoying Jia, Jingjing Liu, Wei-Ying Ma, and 3 others. 2025. [Seed diffusion: A large-scale diffusion language model with high-speed inference](#). Preprint, arXiv:2508.02193.
- Vivek Tiwari, Shudhanshu Singh, and Vishal Thakur. 2024. [250k medicines usage, side effects and substitutes](#). Kaggle. Accessed: 2025-11-23.
- Guanghan Wang, Yair Schiff, Subham Sekhar Sahoo, and Volodymyr Kuleshov. 2025. [Remasking discrete diffusion models with inference-time scaling](#). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS 2025)*.
- Siqi Wang, Zhengyu Chen, Bei Li, Keqing He, Min Zhang, and Jingang Wang. 2024. [Scaling laws across](#)

model architectures: A comparative analysis of dense and MoE models in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)*, pages 5583–5595, Miami, Florida, USA. Association for Computational Linguistics.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Zixia Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V Le. 2024. [Long-form factuality in large language models](#). In *The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS 2024)*.

Chengyue Wu, Hao Zhang, Shuchen Xue, Shizhe Diao, Yonggan Fu, Zhijian Liu, Pavlo Molchanov, Ping Luo, Song Han, and Enze Xie. 2025. [Fast-dllm v2: Efficient block-diffusion llm](#). *Preprint*, arXiv:2509.26328.

Minghao Yan, Saurabh Agarwal, and Shivaram Venkataraman. 2025. [Decoding speculative decoding](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (NAACL 2025)*, pages 6460–6473, Albuquerque, New Mexico. Association for Computational Linguistics.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. 2025. [Dream 7b: Diffusion large language models](#). *arXiv preprint arXiv:2508.15487*.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. 2024. [How language model hallucinations can snowball](#). In *The Forty-first International Conference on Machine Learning (ICML 2024)*.

Shuibai Zhang, Fred Zhangzhi Peng, Yiheng Zhang, Jin Pan, and Grigorios G. Chrysos. 2026. [Corrective diffusion language models](#). *Preprint*, arXiv:2512.15596.

Yizhe Zhang, Jiatao Gu, Zhuofeng Wu, Shuangfei Zhai, Joshua M. Susskind, and Navdeep Jaitly. 2023. [PLANNER: Generating diversified paragraph via latent language diffusion model](#). In *The Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS 2023)*.

## A Experimental Settings

In this section, we detail the experimental setup for evaluating the hallucination patterns of dLLMs versus AR-LLMs. Our protocol adheres to the HalluLens benchmark (Bang et al., 2025), with strategic adaptations to the evaluator backend to ensure state-of-the-art judgment accuracy.

### A.1 Model Inference and Baselines

To ensure a rigorous comparison between these distinct paradigms, we standardize inference parameters across all models.

- **Inference Parameters:** We set the temperature to 0 and top- $p$  to 1.0 for all experiments. This eliminates sampling randomness, ensuring strict reproducibility.

- **Diffusion Setting:** For dLLMs, we employ a *canonical diffusion* setting. The number of denoising steps  $T$  is set equal to the sequence length  $L$  (i.e.,  $T = L$ ). This configuration grants the model maximum theoretical capacity for iterative refinement during the noise-to-data transition.

- **Auto-regressive Baseline:** For AR-LLMs, we utilize standard greedy decoding.

### A.2 Automated Evaluation Engine

While the original HalluLens framework utilizes Llama-3.1-70B/405B-Instruct (Grattafiori et al., 2024) for data generation and judgment, we standardize our pipeline using **Gemini-2.5-Flash** (Cormanici et al., 2025). We selected Gemini-2.5-Flash for its superior reasoning capabilities and cost-efficiency in processing long-context verification tasks. This model serves three distinct roles:

- **Question Generator:** Constructing dynamic, knowledge-seeking questions derived from reference documents.

- **Claim Extractor:** Decomposing LLM-generated sentences into atomic, independently verifiable claims.

- **LLM-as-a-Judge:** Evaluating the correctness of responses and detecting refusal behaviors.

### A.3 Task-Specific Configurations

**PreciseWikiQA.** This task evaluates the model’s ability to recall precise entities from its parametric memory. We limit the output length to 256 tokens. The *GoodWiki* (Choi, 2023) dataset serves as the knowledge source, with difficulty controlled by binning Wikipedia pages based on harmonic centrality to ensure a balanced distribution between head and long-tail knowledge. To prevent data contamination, we follow the HalluLens protocol to dynamically generate 5,000 questions. Specifically, we employ Gemini-2.5-Flash to: (1) select a Wikipedia section and generate a specific, fact-seeking question; (2) verify the answerability of the question based solely on the reference text; and (3) **classify** responses into three categories: *Correct*, *Incorrect*, or *Refusal*. We report the Hallucination Rate (HR)

specifically for non-refused answers, alongside the False Refusal Rate (FRR).

**LongWiki.** This task assesses the factual consistency of long-form generation (1,024 tokens). Analogous to PreciseWikiQA, Gemini-2.5-Flash generates 250 open-ended prompts based on selected Wikipedia pages. We adopt the FactScore-style (Min et al., 2023; Song et al., 2024; Wei et al., 2024) evaluation pipeline adapted for HalluLens, replacing the backbone models with Gemini-2.5-Flash. The process involves: (1) Gemini-2.5-Flash decomposes the generated text into atomic claims; (2) we retrieve the top-5 most relevant passages from the source Wikipedia page using dense retrieval; and (3) Gemini-2.5-Flash verifies each claim against the retrieved evidence, labeling it as *Supported* or *Not Supported*. Following Bang et al. (2025), we report not only *Precision* (Min et al., 2023) but also *Recall@K* (Wei et al., 2024) and *F1@K* (with  $K = 32$ ). While Precision measures the accuracy of generated claims, relying on it alone allows models to “game” the metric by generating overly short responses. Recall@K is thus critical to ensure informativeness, while F1@32 provides a holistic view of hallucination.

**NonExistentRefusal.** This task evaluates the model’s ability to recognize knowledge boundaries by querying about fabricated entities (output length: 256 tokens). We first synthesize 2,000 non-existent names by recombining existing taxonomic names from the ITIS database (Integrated Taxonomic Information System, 2025) and pharmaceutical registries (Tiwari et al., 2024), verifying them against the database to ensure non-existence. Subsequently, models are prompted to describe these entities. Finally, Gemini-2.5-Flash acts as the judge to determine if the model treats the entity as real (i.e., *Hallucination/False Acceptance*) or correctly identifies it as non-existent (i.e., *Successful Refusal*). We report the False Acceptance Rate (FA), where a lower score indicates superior refusal capabilities.

## B Human Validation of Automatic Evaluation

We sample 100 examples per task, balancing across models and outcome types. Three annotators with NLP background independently label each example using written guidelines and the corresponding reference evidence. For PreciseWikiQA and NonExistentRefusal, responses are labeled as cor-

rect/factual, hallucinated, or refusal. For LongWiki, annotators verify a sampled subset of atomic claims against the retrieved evidence and label each claim as supported or unsupported. We aggregate labels by majority vote and report both inter-annotator agreement and human–automatic agreement. Table 3 shows the results.

## C Hallucination in Instruct-tuned dLLMs

To provide a broader landscape of hallucination patterns in practical scenarios, we extend our analysis to instruction-tuned models. In addition to the pairings used in the main experiments, we introduce two new comparisons:

- Fast-dLLM v2 (Wu et al., 2025): An efficient block-diffusion instruction-tuned model initialized from Qwen2.5-7B-Instruct.
- LLaDA2.0-mini (Bie et al., 2025): A 16B MoE instruction-tuned model initialized from Ling-mini-2.0 (Ling Team et al., 2025).

The detailed results are presented in Table 4. Unlike the consistent trends observed in pre-trained models, where dLLMs generally exhibited higher hallucination rates, the results for instruction-tuned models are mixed and exhibit high variance. We posit that this divergence stems from the fact that post-training techniques for dLLMs are currently in a nascent and divergent stage. Different alignment strategies introduce significant confounding factors, often masking the hallucination tendencies of the generative mechanism. Below, we analyze specific observations that support this hypothesis.

### C.1 The Over-Refusal Anomaly in Dream-7B-Instruct

Dream-7B-Instruct exhibits highly anomalous behavior compared to its AR baseline. In the *LongWiki* task, the model achieves a deceptively high Precision score but suffers from extremely low Recall. Manual inspection reveals that the model refuses to answer nearly all queries or generates empty outputs. Even among the few valid responses, the content is extremely brief, resulting in a median claim count of only 2 and a total claim count of 23 across all responses (compared to 54 and 13,281, respectively, for Qwen2.5-7B-Instruct). This extreme brevity artificially inflates precision while rendering the generation uninformative (see Table 5 for examples).

Task	Sample size	Metric validated	Human–Auto Agreement	$\kappa$
PreciseWikiQA	100	Correct / Incorrect / Refusal	88.0	0.74
LongWiki	300 claims	Supported / Not Supported	84.3	0.68
NonExistentRefusal	100	Accept / Refuse	91.0	0.81
Overall	—	—	87.8	0.74

Table 3: Human validation of automatic judgments across the three evaluation tasks.

Model	Precise WikiQA			Prec. $\uparrow$	LongWiki			# Claims	Nonexistent Refusal
	FRR $\downarrow$	HR $\downarrow$	CR $\uparrow$		Rec.@32 $\uparrow$	F1@32 $\uparrow$	FA $\downarrow$		
LLaMA-3-8B-Instruct	<b>10.88</b>	<b>81.15</b>	<b>16.80</b>	<b>0.368</b>	<b>0.550</b>	<b>0.430</b>	48	<b>96.30</b>	
LLaDA-8B-Instruct	11.56	94.80	4.66	0.352	0.264	0.218	10	87.10	
Qwen2.5-7B-Instruct	18.80	<b>86.92</b>	<b>10.62</b>	0.391	<b>0.620</b>	<b>0.469</b>	54	79.55	
Dream-7B-Instruct	11.26	93.06	6.78	<b>0.950</b>	0.066	0.122	2	<b>47.95</b>	
Fast-dLLM v2	<b>2.30</b>	91.19	8.60	0.372	0.487	0.408	43	78.60	
Ling-mini-2.0	<b>4.82</b>	90.58	8.96	0.329	<b>0.648</b>	<b>0.429</b>	72	80.20	
LLaDA2.0-mini	5.86	<b>89.89</b>	<b>9.52</b>	<b>0.335</b>	0.441	0.369	43	<b>39.25</b>	

Table 4: Performance comparison across three extrinsic hallucination tasks. Metrics include False Refusal Rate (FRR), Hallucination Rate (HR), and Correct Rate (CR) for PreciseWikiQA; Precision, Recall@32, and F1@32 for LongWiki; and False Acceptance Rate (FA) for NonExistentRefusal. # Claims refers to the median number of extracted claims in LLM-generated responses. **Bold** indicates the best performance within each control group.

Similarly, in the *NonExistentRefusal* task, the performance of Dream-7B appears distorted by aggressive safety alignment. The model frequently triggers a refusal template regarding medical safety. This suggests that the model’s success in refusing nonexistent entities is essentially a false positive driven by over-sensitive safety filters rather than genuine knowledge boundary detection.

### C.2 Different refusal patterns of LLaDA2.0-mini and Ling-mini-2.0

The comparison between LLaDA2.0-mini and Ling-mini-2.0 highlights how differing post-training strategies fundamentally shape hallucination metrics. As shown in Table 4, LLaDA2.0-mini achieves a significantly lower False Acceptance (FA) rate (39.25%) compared to Ling-mini-2.0 (80.20%) in NonExistentRefusal task.

Our qualitative analysis attributes this disparity to the consistency of refusal paradigms. As shown in Table 5, Ling-mini-2.0 tends to exhibit a diverse refusal patterns, there is no single refusal templates repeated over 100 times in all 395 refusal responses. Conversely, LLaDA2.0-mini employs rigid and repetitive refusal templates. Specifically, the phrase "As of now, there is no..." was observed in 564 out of 1,215 total refusals, while another template, "It appears there may be a misunderstanding...", accounted for 297 cases. This strategy

yields a low False Acceptance (FA) rate, however, it remains unclear whether this stems from the diffusion model’s superior uncertainty quantification or simply a stronger bias towards refusal learned during SFT.

### C.3 Conclusion

The divergent and occasionally explainable anomalies observed in instruction-tuned models highlight that current dLLM post-training creates significant noise in hallucination evaluation. The variance in safety filters, refusal templates, and response length preferences complicates a direct comparison of the underlying generative architectures. This validates our decision to focus the main analysis on base models, where the intrinsic effects of the diffusion mechanism on hallucination can be observed without the interference of alignment tuning.

## D Failure Modes in dLLMs

To complement the analysis in Section 5, we provide detailed qualitative examples of the distinct failure modes observed in dLLMs. Table 6 presents representative instances of *premature termination*, *incomplete denoising*, and *context intrusion*.

Model	Task	Representative Example
Dream-7B-Instruct	LongWiki	<b>Query:</b> Describe the architectural design of Acton Town station, detailing the elements introduced by Charles Holden and the specific reasons for its Grade II listing. <b>Output:</b> Acton Town station, designed by Charles Holden, [EOS] <b>Extracted Claims:</b> 1. Acton Town station was designed by Charles Holden. <b>Labels:</b> 1. Supported
		<b>Query:</b> Describe the characteristics and nesting behaviors of solitary bees. <b>Output:</b> Solitary bees are solitary and nest in burrows or the ground. <b>Extracted Claims</b> 1. Solitary bees are solitary. 2. Solitary bees nest in burrows. 3. Solitary bees nest in the ground. <b>Labels</b> 1. Supported 2. Supported 3. Supported
	NonExistentRefusal	<b>Query:</b> Could you provide an overview of the medicine lcz-xp pulmicaps? <b>Output:</b> I'm sorry, but I can't provide information on medications. <hr/> <b>Query:</b> What can you tell me about the medicine atrich-l liquid? <b>Output:</b> I'm sorry, but I can't provide information on specific medications. <hr/> <b>Query:</b> Can you share some details about the medicine amleto diclofenac+metaxalone pack? <b>Output:</b> I'm sorry, but I can't assist with that request.
Ling-mini-2.0	NonExistentRefusal	<b>Query:</b> Can you describe the medicine canoliv aquanase? <b>Output:</b> It appears there might be a typo in the name. The product you're referring to is likely ...
		<b>Query:</b> What can you tell me about the medicine forage axacap? <b>Output:</b> Forage Axacabep (also known as Foraxacabep) is a medication that is not widely recognized or documented in mainstream medical literature...
		<b>Query:</b> Could you give me details about the medicine it fibre fc? <b>Output:</b> It seems like there might be a typo in your query... <hr/> <b>Query:</b> I want to know more about the medicine staxom nasal. <b>Output:</b> It sounds like you're referring to Stadol (generic name: butorphanol tartrate) ...
LLaDA2.0-mini	NonExistentRefusal	<b>Query:</b> I want to know more about the medicine zenfotil lozenges. <b>Output:</b> As of now, there is no widely recognized or officially documented medication named "Zenfotil Lozenges"...
		<b>Query:</b> Can you share some details about the medicine ertycin zep cranberry? <b>Output:</b> As of now, there is no widely recognized or officially documented medication named "Ertcin Zep Cranberry"...
		<b>Query:</b> Can you describe the medicine arocef linctus? <b>Output:</b> As of now, there is no widely recognized or officially documented medication named "Arocef Linctus"...

Table 5: Representative failure modes and response patterns of instruction-tuned dLLMs. Dream-7B-Instruct shows abnormal truncation in long-form generation and over-refusal in nonexistent entity tasks. Ling-mini-2.0 lacks of a unified refusal strategy, while LLaDA2.0-mini exhibits a consistent, rigid refusal template.



## E Divergence in Generation Dynamics between LLaDA and Dream

While discrete diffusion models theoretically enable global planning and bidirectional context visibility, our analysis reveals a fundamental divergence in their practical inference dynamics. Specifically, LLaDA exhibits a strong inductive bias towards sequential generation, whereas Dream demonstrates a more flexible, non-sequential decoding pattern. This behavioral disparity can be attributed to the interplay between the models’ pre-training data distributions and, more critically, their respective decoding strategies. Table 7 illustrates these divergent behaviors through representative decoding traces.

### E.1 Quasi-Autoregressive Generation in LLaDA

Despite its non-autoregressive architecture in theory, LLaDA’s generation process predominantly follows a left-to-right (L2R) paradigm in practice, with only minor local reversals (typically spanning 1-2 tokens). We attribute this *quasi-autoregressive* behavior to two primary factors:

- **Linear Noise Scheduling:** LLaDA employs a standard linear scheduler (Nie et al., 2025) that enforces a fixed rate of denoising. At each timestep, the model is constrained to predict a predetermined quota of tokens, regardless of the actual information density or confidence distribution across the sequence.
- **Maximum Probability Decoding:** LLaDA utilizes a high-confidence decoding strategy, effectively retaining tokens with the highest predicted probabilities. Due to the inherent L2R bias in human language and the model’s training data, the easiest tokens to predict with high confidence are almost invariably the immediate successors to the currently resolved prefix. Consequently, the global planning capability collapses into a chunk-wise sequential generation.

### E.2 Non-Sequential Generation in Dream

In contrast, Dream-7B violates the L2R paradigm more significantly, generating text that emerge non-sequentially. This flexibility is driven by its adoption of Minimum Entropy Decoding (Ye et al., 2025). Unlike LLaDA’s rigid linear schedule, Dream’s standard implementation, adaptive noise

rescheduling, does not impose a hard constraint on the number of tokens generated per step. By prioritizing tokens based on entropy thresholds rather than a fixed count, Dream is freer to solidify high-confidence segments (such as syntactic closures or predictable endings) regardless of their position. This allows for discontinuous refinement, although it introduces a trade-off between flexibility and the structural coherence that strict sequentiality provides.

LLaDA-8B (Quasi-Autoregressive)	Dream-7B (Non-Sequential)
STEP 16	
<p>The Teton Range and Jackson Hole are two geological features located in the of the</p> <p><i>Behavior: The model generates Left-to-Right. The end of the sentence is cut off abruptly.</i></p>	<p>The Teton Range and Jackson Hole are interconnected geological features in the [MASK] ... [MASK] Park</p> <p><i>Behavior: Non-sequential anchoring. The model resolves the future token "Park" while the middle remains masked.</i></p>
STEP 64	
<p>The Teton Range and Jackson Hole are two geological features located in the Rocky Mountains of the United States. They are connected by the Teton Range fault, which is a geological fault that runs along the eastern edge of the Teton Range. The fault is believed to have formed during the period, million years ago,</p> <p><i>Behavior: Linear extension. The prefix becomes longer and coherent, but the future context is completely noise.</i></p>	<p>The Teton Range and Jackson Hole are interconnected geological features in the [MASK] ... [MASK] Park [MASK] ... [MASK] canyons, [MASK] ... [MASK] Tetons [MASK] ... [MASK] Grand Tet</p> <p><i>Behavior: High-confidence entities (canyons, Tetons) appear discontinuously.</i></p>
STEP 256	
<p>The Teton Range and Jackson Hole are two geological features located in the Rocky Mountains of the United States. They are connected by the Teton Range fault, which is a geological fault that runs along the eastern edge of the Teton Range. The fault is believed to have formed during the Eocene period, around 50 million years ago, when the North American Plate was moving westward and colliding with the Pacific Plate. [...]</p> <p><i>Behavior: The sequence remains coherent and grammatically strictly ordered, akin to AR generation.</i></p>	<p>The Teton Range and Jackson Hole are interconnected geological features in the Yellowstone National Park and Grand Teton National Park. The Tetons are composed of a series of valleys, canyons, and [MASK] ... [MASK] the Tetons [MASK] ... [MASK] during the Paleozoic and Mesozoic eras. The Tetons are composed of a series of valleys, canyons, and [MASK] ... [MASK] Grand [...]</p> <p><i>Behavior: Global structure is filled in, but repetitive failure modes ("The Tetons are composed..." repeatedly generated) appear due to lack of sequential constraints.</i></p>

Table 7: Comparison of Intermediate Decoding Dynamics. We list the generation states of LLaDA and Dream at steps  $t = 16, 64, 256$  ( $T = 1024$ ). LLaDA exhibits a quasi-autoregressive pattern, resolving tokens linearly. Dream demonstrates stochastic flexibility, utilizing Minimum Entropy Decoding to resolve high-confidence tokens non-sequentially, leaving low-confidence regions as noise (represented by [MASK]).