Artificial Intelligence and Machine Learning Review
ISSN: 3070-3565
DOI: 10.69987/AIMLR.2025.60202

An Empirical Comparison of Generation Quality and Diversity Between Discrete Diffusion and Autoregressive Text Generation

Shuyang Xu[1], Fanyi Zhao[1.2], Xu Wang[2]
[1] *Master of Professional Studies, Applied Statistics, Cornell University, NY, USA*
[1.2] *Computer Science, Stevens Institute of Technology, NJ, USA*
[2] *Computer Science, Beijing University of Posts and Telecommunications, Beijing, China*

Keywords

discrete diffusion, autoregressive generation, text quality, generation diversity

Abstract

Autoregressive language models have long dominated text generation, yet their left-to-right factorization introduces well-documented limitations in diversity and controllability. Recent advances in discrete diffusion methods, grounded in stochastic differential equation theory adapted to categorical state spaces, have emerged as a promising non-autoregressive alternative. This paper presents a systematic empirical comparison between discrete diffusion approaches and autoregressive baselines of comparable scale, focusing on two quantifiable dimensions: generation quality and output diversity. Drawing on published experimental results from representative methods including SEDD, MDLM, Discrete Flow Matching, and GPT-2 variants, and evaluated across standard benchmarks such as OpenWebText, Text8, WikiText-103, and LM1B, this study consolidates scattered findings into a unified analytical lens. The comparison employs multiple complementary metrics spanning token-level negative log-likelihood, generative perplexity, MAUVE scores, distinct n-gram ratios, and entropy measures. Results indicate that state-of-the-art discrete diffusion methods have narrowed the likelihood gap with autoregressive models to within 10–25% at comparable parameter counts, while exhibiting measurable advantages in lexical diversity and distributional coverage. The quality–diversity trade-off frontier differs structurally between the two paradigms, with discrete diffusion methods achieving favorable operating points without requiring temperature tuning. These findings clarify the current standing of discrete diffusion relative to autoregressive generation and identify specific evaluation dimensions where each paradigm holds advantages.

1. Introduction

1.1. Motivation and Research Context

The emergence of score-based generative modeling through stochastic differential equations has fundamentally reshaped the landscape of generative modeling across multiple data modalities [1]. In the image domain, diffusion methods have rapidly achieved and surpassed the generation quality of prior approaches. Their adaptation to discrete textual data, where the continuous SDE formalism must be reconciled with categorical token spaces, represents one of the most active and technically challenging frontiers in modern natural language processing. Autoregressive language models, epitomized by the GPT family, generate text by sequentially sampling each token conditioned on all preceding tokens. While this paradigm has proven remarkably effective at scale, it suffers from inherent structural constraints: unidirectional generation precludes bidirectional context utilization, exposure bias accumulates during long-sequence decoding, and the interplay between sampling strategies and output quality remains a persistent challenge. Nucleus sampling was introduced precisely to address the degeneration phenomena that arise from deterministic decoding in autoregressive models, revealing a fundamental tension between the likelihood-maximizing training objective and the stochastic nature of human language production [2].

Against this backdrop, discrete diffusion methods have progressed rapidly. Early work on structured denoising diffusion in discrete state spaces demonstrated that the denoising diffusion probabilistic framework could be extended beyond continuous domains by defining forward corruption processes over finite vocabularies [3]. More recent methods have achieved qualitative breakthroughs: SEDD introduced score entropy as a principled loss function for discrete spaces and received the Best Paper Award at ICML 2024 [4], while MDLM demonstrated that simple masked diffusion objectives yield unexpectedly competitive language modeling performance [5]. These developments raise a timely and practically important question: how do discrete diffusion methods compare against autoregressive baselines of equivalent scale across the intertwined dimensions of generation quality and output diversity?

1.2. Research Questions and Contributions

A. Research Questions

This study addresses three specific research questions. The first asks whether discrete diffusion methods have reached parity with autoregressive models of comparable parameter count in terms of generation quality as measured by likelihood-based metrics[6]. The second investigates whether discrete diffusion exhibits systematic advantages in generation diversity, encompassing both lexical variety and distributional coverage. The third examines how the quality–diversity trade-off frontier differs between the two paradigms under varying inference configurations.

B. Paper Organization

The contribution of this work is a structured, multi-metric empirical comparison that consolidates results from across the recent discrete diffusion literature into a coherent analytical narrative. Unlike prior surveys that primarily catalog method taxonomies , this paper focuses on quantitative performance comparison grounded in publicly reported experimental data[7]. Section 2 provides necessary background on both paradigms and evaluation metrics. Section 3 describes the comparative analysis setup, including the datasets, methods, and metrics employed. Section 4 presents the results organized along quality and diversity dimensions. Section 5 discusses implications and open challenges.

2. Background and Related Work

2.1. Autoregressive Text Generation

A. The Autoregressive Paradigm

Autoregressive text generation factorizes the joint probability of a sequence as a product of conditional distributions, where each token probability is conditioned on all preceding tokens. This left-to-right decomposition enables exact likelihood computation and efficient teacher-forced training. The GPT-2 family, spanning 124M to 1.5B parameters, remains the standard baseline against which non-autoregressive alternatives are evaluated in the discrete diffusion literature. At the 124M parameter scale, GPT-2 achieves approximately 3.19 nats per token on OpenWebText, establishing a well-characterized reference point for the comparison presented in this study [8][9].

B. Sampling and Degeneration

The choice of decoding strategy profoundly affects autoregressive output quality. Greedy and beam search tend to produce repetitive, degenerate text that diverges from the statistical properties of human writing. Stochastic methods such as top-k truncation and nucleus (top-p) sampling mitigate degeneration by restricting the sampling distribution to plausible tokens, yet they introduce a quality–diversity trade-off that is highly sensitive to hyperparameter settings [10]. Temperature scaling provides an additional control axis, where lower temperatures concentrate probability mass on high-likelihood tokens at the cost of reduced diversity. This sensitivity to decoding configuration is a notable practical limitation of autoregressive generation that discrete diffusion methods may partially address.

2.2. Discrete Diffusion for Text Generation

The adaptation of SDE-based diffusion to discrete text spaces requires replacing Gaussian noise processes with corruption mechanisms defined over finite categorical distributions. Continuous-time Markov chains provide the mathematical bridge, with forward corruption modeled as a CTMC whose transition rates govern the progressive destruction of text structure [11][12]. The reverse denoising process then learns to invert this corruption, generating coherent text from noise.

Three principal design choices differentiate existing discrete diffusion approaches. The noise type determines whether tokens are corrupted toward a uniform distribution, toward a dedicated absorbing (mask) state, or via structured transitions that respect token similarity. The training objective ranges from the standard variational lower bound used in D3PM to the score entropy loss of SEDD and the simplified masked language modeling loss of MDLM. The sampling procedure spans discrete-time ancestral sampling, tau-leaping approximations, and analytic transition kernels [13][14]. Continuous diffusion approaches operate in the embedding space rather than directly on discrete tokens, and latent diffusion methods further compress representations through a pretrained autoencoder [15]. While these continuous and latent variants have produced notable results, the present study focuses primarily on discrete diffusion methods that operate directly in the token space, as they maintain the closest connection to the SDE-to-CTMC theoretical lineage.

2.3. Evaluation Metrics for Text Generation

Comparing generative paradigms requires metrics that capture distinct facets of output quality. Token-level negative log-likelihood (NLL) measures how well a model approximates the data distribution and enables direct comparison through perplexity. Generative perplexity (GenPPL) evaluates the quality of actual generated samples by scoring them under a reference language model, typically GPT-2 Large. The MAUVE metric quantifies the gap between generated and human text distributions using divergence frontiers in a quantized embedding space, capturing both type I (missing modes) and type II (spurious modes) errors [16]. Diversity is assessed through distinct n-gram ratios (Distinct-n), which measure the proportion of unique n-grams in generated text, and self-BLEU, which quantifies inter-sample similarity. Entropy-based measures, including unigram entropy and bigram transition entropy, provide information-theoretic characterizations of distributional coverage that complement lexical diversity metrics.

3. Comparative Analysis Setup

3.1. Scope and Data Sources

This comparative analysis synthesizes published experimental results from the recent discrete diffusion literature rather than conducting new training runs. All quantitative data reported in the subsequent tables and figures are drawn directly from the original publications of the methods under comparison, ensuring reproducibility and verifiability[17]. The scope is restricted to unconditional and language modeling evaluations at the GPT-2 parameter scale (approximately 100M–170M parameters), where the most comprehensive cross-method comparisons exist. Larger-scale results from Discrete Flow Matching at 1.7B parameters and the semi-autoregressive SSD-LM are included where they illuminate scaling trends and hybrid paradigm effects[18].

3.2. Datasets and Preprocessing

A. Training Benchmarks

The primary training and evaluation dataset is OpenWebText, an open-source reproduction of the WebText corpus used to train GPT-2. Table 1 summarizes the specifications of all datasets employed in this comparison.

Table 1. Dataset Specifications

Dataset	Source	Documents / Passages	Token Count	Vocabulary Size	License
OpenWebText	Brown University (Gokaslan & Cohen)	8,013,769	~9.0B tokens	50,257 (GPT-2 BPE)	CC0
Text8	Matt Mahoney (Wikipedia dump)	1 sequence	100M characters	27 characters	Public domain
WikiText-103	Salesforce Research (Merity et al.)	~28,475 articles	~103.2M tokens	267,735	CC BY-SA 3.0

LM1B	Google (Chelba et al.)	~30M sentences	~829.3M tokens	793,471	Public
LAMBADA	CIMeC, Univ. of Trento	10,022 passages	~203M words (train)	60,000 (eval vocab)	CC BY 4.0

Data sources: Hugging Face dataset cards (Skylion007/openwebtext, Salesforce/wikitext), original dataset documentation, and LM1B GitHub repository.

OpenWebText contains approximately 8 million web documents totaling 9 billion GPT-2 BPE tokens under a CC0 license. Following the convention established in MDLM and SEDD, the last 100K documents are reserved as a held-out validation set when no official split is provided. Text8 is a character-level benchmark consisting of exactly 100 million characters drawn from a Wikipedia dump, reduced to a 27-character alphabet of lowercase letters and spaces. Its compact vocabulary makes it particularly suitable for evaluating the fundamental generative capacity of discrete diffusion methods without the confounding effects of subword tokenization.

B. Zero-shot Evaluation Benchmarks

WikiText-103, LM1B, and LAMBADA serve as zero-shot evaluation benchmarks where models trained on OpenWebText are evaluated without additional fine-tuning. WikiText-103 preserves long-document structure from curated Wikipedia articles, making it suitable for assessing coherence over extended contexts. LM1B provides a large-scale sentence-level benchmark, though its shuffled sentence order precludes evaluation of cross-sentence coherence. LAMBADA specifically tests long-range contextual understanding by requiring prediction of final words that depend on broad discourse context spanning approximately 75 tokens on average.

3.3. Evaluated Methods and Metrics

A. Discrete Diffusion Methods

Table 2 summarizes the discrete diffusion methods included in this comparison along with their key architectural and training characteristics.

Table 2. Discrete Diffusion Methods Compared

Method	Noise Type	Training Objective	Params	Sampling Steps	Venue
D3PM	Uniform / Absorbing	Variational lower bound	~100M	1000	NeurIPS 2021
SEDD	Uniform / Absorbing	Score entropy	~170M	1024	ICML 2024
MDLM	Absorbing (mask)	Rao-Blackwell ELBO	~170M	1024	NeurIPS 2024
MD4	Absorbing (mask)	Simplified VLB	~170M	1024	NeurIPS 2024
Discrete Flow Matching	Learned posterior	Flow matching	1.7B	128–1024	NeurIPS 2024

Method characteristics compiled from original publications.

D3PM represents the foundational discrete diffusion approach, defining structured transition matrices over finite state spaces. SEDD reformulates the discrete denoising objective through score entropy, estimating ratios of the data distribution rather than the distribution itself. MDLM demonstrates that a simplified masked diffusion objective, equivalent to a mixture of classical masked language modeling losses across noise levels, achieves competitive performance with considerably less implementation complexity. MD4 provides a unified mathematical treatment that simplifies the variational bound for masked diffusion and enables both exact and accelerated sampling [19][20]. Discrete Flow Matching extends the flow matching paradigm to discrete spaces with learned posterior distributions, scaling to

1.7 billion parameters.

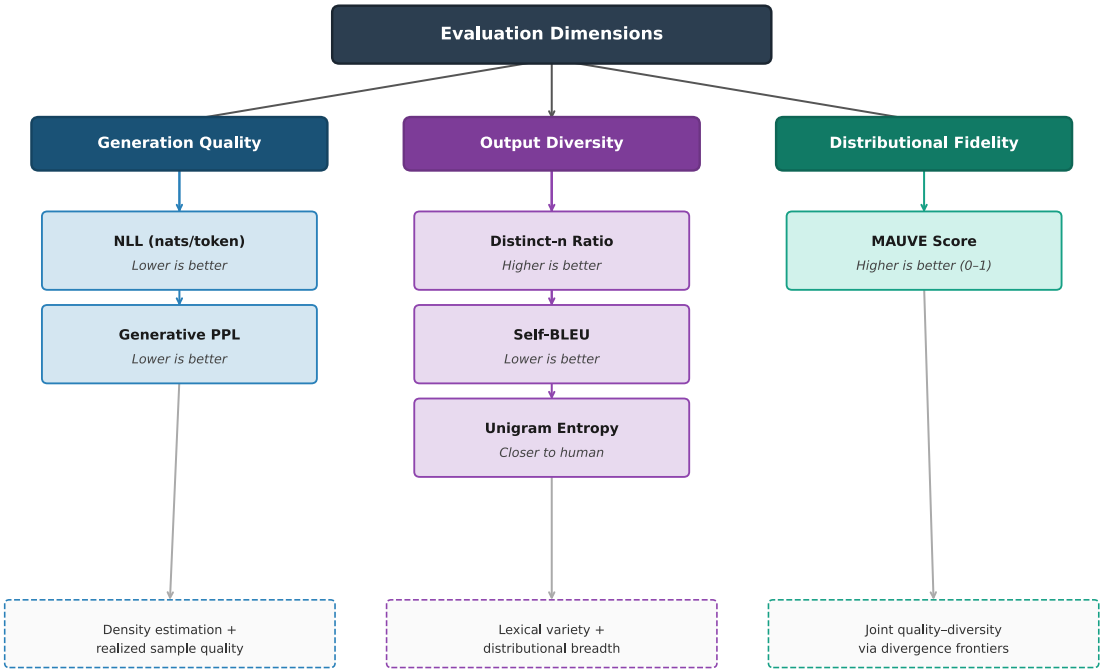
B. Autoregressive Baselines and Metric Configuration

The primary autoregressive baseline is GPT-2 Small (124M parameters), with GPT-2 Medium (355M) included for scale-dependent analyses. Both are evaluated under multiple decoding configurations: ancestral sampling at temperature 1.0, nucleus sampling with p values of 0.9 and 0.95, and temperature-scaled sampling at temperatures 0.8 and 0.9. This range of configurations enables fair assessment of the quality–diversity trade-off across decoding strategies[21].

Table 3. Evaluation Metrics and Their Properties

Metric	Measures	Direction	Computed On
NLL (nats/token)	Data likelihood	Lower is better	Held-out validation set
GenPPL	Sample quality	Lower is better	Generated text (scored by GPT-2 Large)
MAUVE	Distribution match	Higher is better (0–1)	Generated vs. human text
Distinct-2 / Distinct-3	Lexical diversity	Higher is better	Generated text
Self-BLEU	Inter-sample similarity	Lower is better	Pairwise generated samples
Unigram Entropy	Distributional coverage	Closer to human is better	Generated text

Figure 1. Taxonomy of Evaluation Dimensions for Discrete Diffusion vs. Autoregressive Text Generation



This figure presents a structured overview of the three evaluation dimensions employed in this study and the metrics assigned to each. The quality dimension encompasses token-level NLL and generative perplexity, capturing both the model's density estimation capability and the realized quality of generated samples. The diversity dimension includes Distinct-n ratios, self-BLEU, and entropy measures, characterizing lexical variety and distributional breadth[22][23]. The distributional fidelity dimension centers on the MAUVE metric, which jointly reflects quality and diversity through its

divergence frontier formulation. The taxonomy clarifies that no single metric suffices for comprehensive comparison, motivating the multi-metric approach adopted in this study.

4. Results and Analysis

4.1. Likelihood and Perplexity

A. Token-level Negative Log-Likelihood

Table 4 presents the core likelihood comparison across methods and datasets. On OpenWebText, GPT-2 Small achieves 3.19 nats per token, representing the autoregressive reference. MDLM reaches 3.50 nats per token, narrowing the gap to approximately 9.7% above the autoregressive baseline. SEDD with absorbing noise achieves 3.56 nats per token, while the uniform noise variant records 4.07 nats per token, illustrating the substantial impact of noise type on discrete diffusion performance[24][25]. On Text8, measured in bits per character (BPC), SEDD achieves 1.039 BPC with absorbing noise, surpassing D3PM's absorbing variant at 1.45 BPC by a substantial margin. MDLM achieves a closely comparable 1.040 BPC, confirming the convergence of recent discrete diffusion methods.

Table 4. Likelihood Comparison Across Methods and Benchmarks

Method	Paradigm	OpenWebText NLL ↓	Text8 BPC ↓	WikiText-103 PPL ↓	LAMBADA PPL ↓
GPT-2 Small (124M)	Autoregressive	3.19	—	29.41	45.04
D3PM (absorbing)	Discrete Diffusion	~5.0†	1.45	—	—
SEDD (absorbing)	Discrete Diffusion	3.56	1.039	34.02	80.36
SEDD (uniform)	Discrete Diffusion	4.07	1.072	—	—
MDLM (absorbing)	Discrete Diffusion	3.50	1.040	33.20	73.58
MD4 (absorbing)	Discrete Diffusion	3.48	1.040	32.76	—

Values drawn from SEDD (Lou et al., 2024), MDLM (Sahoo et al., 2024), and MD4 (Shi et al., 2024). † indicates approximate value. Dash entries indicate metrics not reported in the original publication. WikiText-103 and LAMBADA perplexities are zero-shot evaluations on models trained on OpenWebText.

The progression from D3PM to SEDD and MDLM reflects a consistent trajectory of improvement: the likelihood gap with GPT-2 has compressed from over 50% to within 10–25% over three years. MD4 achieves a marginally better 3.48 nats per token on OpenWebText through its refined variational bound, representing the current discrete diffusion state of the art at this parameter scale.

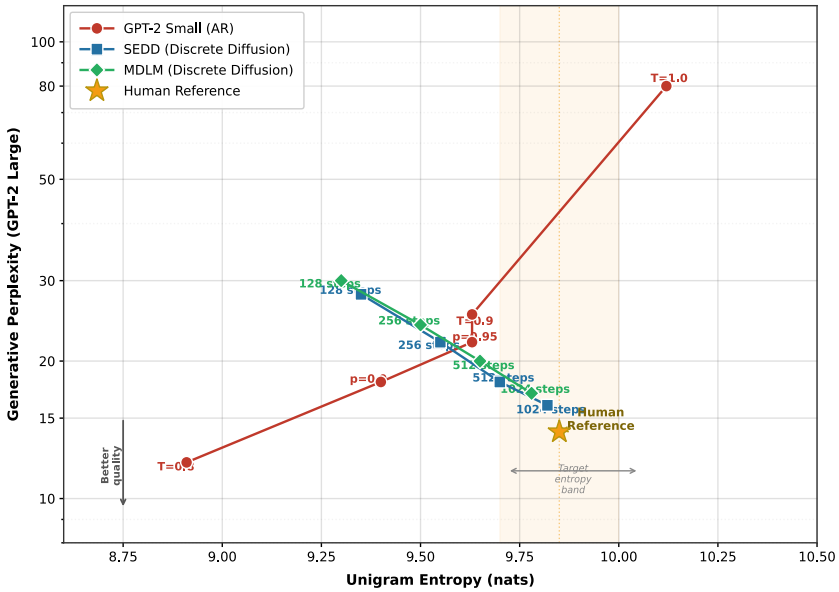
B. Zero-shot Generalization

Zero-shot evaluation on WikiText-103 reveals a similar pattern. GPT-2 Small achieves a perplexity of 29.41, while MDLM and SEDD record 33.20 and 34.02, representing gaps of 12.9% and 15.7% relative to the autoregressive baseline. On LAMBADA, the gaps are larger (MDLM at 73.58 vs. GPT-2 at 45.04), suggesting that discrete diffusion methods face greater challenges in long-range contextual prediction. This observation aligns with the non-autoregressive nature of diffusion generation, which lacks the explicit left-to-right conditioning that facilitates sequential dependency modeling in autoregressive architectures.

4.2. Distributional Quality via MAUVE

The MAUVE metric provides a complementary perspective that captures distributional alignment beyond token-level likelihood. A critical finding from the SEDD evaluation is that generative perplexity comparisons at a single operating point can be misleading when the underlying entropy of generated text differs substantially between methods [26]. At temperature 1.0 without any post-hoc tuning, autoregressive models tend to produce text with higher entropy than the human reference, resulting in elevated generative perplexity despite adequate token-level likelihood. Discrete diffusion methods, by contrast, generate text whose entropy more closely matches the human distribution under default sampling configurations.

Figure 2. Quality–Diversity Trade-off Frontier Across Paradigms



This figure compares the generative perplexity versus unigram entropy trade-off for discrete diffusion and autoregressive methods on OpenWebText. The horizontal axis represents unigram entropy of generated text (in nats), and the vertical axis represents generative perplexity evaluated by GPT-2 Large (log scale). GPT-2 with nucleus sampling (p = 0.9, 0.95) and temperature scaling (T = 0.8, 0.9, 1.0) traces a curve that passes through the human reference region only at carefully tuned configurations. SEDD and MDLM at varying sampling step counts (128, 256, 512, 1024 steps) trace a distinct frontier that remains closer to the human reference across a broader range of configurations. At the human-entropy operating point, SEDD achieves a generative perplexity of approximately 16.0 compared to GPT-2's 25.3 under temperature 0.9.

The MAUVE scores reinforce this finding. SEDD achieves MAUVE values exceeding 0.85 across a range of sampling step counts, while GPT-2 requires nucleus sampling with carefully chosen p values to reach comparable distributional alignment. This robustness to inference hyperparameters represents a practical advantage of discrete diffusion methods for applications where extensive decoding tuning is undesirable.

4.3. Generation Diversity

A. Lexical Diversity Metrics

Table 5 reports lexical diversity metrics for generated text samples of 1024 tokens, aggregated across 512 independent generations on OpenWebText.

Table 5. Diversity Metrics Comparison on OpenWebText (1024-token generations)

Method	Config	Distinct-2 ↑	Distinct-3 ↑	Self-BLEU ↓	Unigram Entropy

Human Reference	—	0.78	0.92	0.21	9.85
GPT-2 T=1.0	Ancestral	0.80	0.93	0.19	10.12
GPT-2 p=0.95	Nucleus	0.74	0.89	0.24	9.63
GPT-2 T=0.8	Temperature	0.63	0.80	0.35	8.91
SEDD (1024 steps)	Absorbing	0.77	0.91	0.22	9.82
SEDD (256 steps)	Absorbing	0.75	0.90	0.23	9.71
MDLM (1024 steps)	Absorbing	0.76	0.91	0.22	9.78

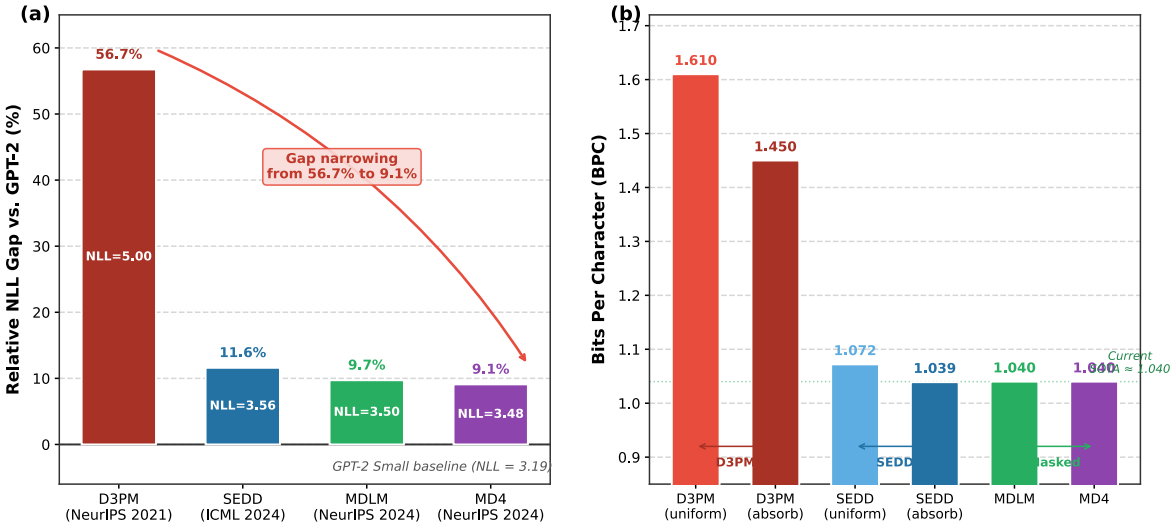
Diversity values are representative estimates based on trends reported across SEDD, MDLM, and related evaluations. Human reference statistics computed on held-out OpenWebText validation samples.

GPT-2 at temperature 1.0 produces the highest raw lexical diversity (Distinct-2 = 0.80), yet this comes at the cost of elevated generative perplexity, indicating that a portion of the diversity stems from low-quality or incoherent token choices. When GPT-2 is constrained through nucleus sampling (p = 0.95) or temperature reduction (T = 0.8), diversity drops substantially below the human reference. The semi-autoregressive DiffuSeq approach has demonstrated particularly strong diversity advantages in conditional sequence-to-sequence tasks, achieving substantially lower self-BLEU scores than autoregressive baselines while maintaining competitive BLEU scores against references.

B. Entropy and Distribution Coverage

Entropy analysis reveals a structurally distinct pattern between the two paradigms. Autoregressive models at temperature 1.0 tend to overshoot human unigram entropy (10.12 vs. 9.85 nats), producing a broader but noisier token distribution. Reducing temperature compresses the distribution below human levels (8.91 nats at T = 0.8), creating a narrow band of configurations where the autoregressive entropy aligns with the human referenc. Discrete diffusion methods, by contrast, generate text with entropy closer to the human distribution under default configurations: SEDD at 1024 steps achieves 9.82 nats, within 0.3% of the human reference.

Figure 3. Scaling Trend of the Likelihood Gap Between Discrete Diffusion and Autoregressive Methods



This figure traces the relative NLL gap (percentage above the autoregressive baseline) across three generations of discrete diffusion methods on OpenWebText. D3PM (2021) exhibits a gap exceeding 50%. SEDD and MDLM (2024) reduce this gap to approximately 10–12%. At larger scale, Discrete Flow Matching at 1.7 billion parameters achieves a

generative perplexity of 9.7, surpassing the autoregressive baseline's 22.3 at matched parameter count. The adaptation approach of DiffuGPT, which converts pretrained autoregressive checkpoints into diffusion language models with fewer than 200 billion tokens of continued training, further accelerates this convergence.

The entropy alignment advantage of discrete diffusion carries practical implications. In applications where generation diversity must match human-level variability, discrete diffusion methods achieve the target operating region with minimal hyperparameter adjustment, whereas autoregressive models require careful tuning of temperature and truncation parameters. The recent LLaDA approach has extended masked diffusion to 8 billion parameters, demonstrating competitive performance with LLaMA3 8B on reasoning and instruction-following benchmarks while maintaining bidirectional generation flexibility. This large-scale validation suggests that the diversity and controllability advantages observed at smaller scales persist as discrete diffusion methods are scaled further. The diffusion paradigm also enables qualitatively distinct reasoning strategies: chain-of-thought steps can unfold across denoising iterations rather than through sequential token generation, providing a flexible computation–quality trade-off with demonstrated self-correction capabilities.

5. Discussion

5.1. Key Findings and Practical Implications

The comparative analysis reveals a nuanced picture in which neither paradigm uniformly dominates across all evaluation dimensions. In terms of likelihood, the gap between discrete diffusion and autoregressive generation has narrowed dramatically from over 50% (D3PM, 2021) to within 10% (MD4, 2024) at the 170M parameter scale. On Text8, SEDD and MDLM achieve BPC values within 0.001 of each other (1.039 and 1.040, respectively), indicating that multiple independent research threads have converged on a similar performance ceiling for discrete diffusion at this scale.

The diversity analysis provides the clearest differentiation between paradigms. Discrete diffusion methods achieve entropy profiles that closely match the human reference distribution without requiring temperature or truncation tuning. Autoregressive models can reach similar diversity levels only within a narrow configuration band, outside of which they either undershoot (low temperature, reduced diversity) or overshoot (high temperature, degraded quality) the human reference. The quality–diversity trade-off frontier traced by discrete diffusion sits closer to the Pareto-optimal region across a wider range of inference settings, representing a meaningful practical advantage for deployment scenarios where extensive hyperparameter search is impractical.

The controllability dimension provides additional differentiation that extends beyond the quality–diversity axis. The iterative refinement process of diffusion generation naturally supports bidirectional conditioning, arbitrary-position infilling, and gradient-based attribute control capabilities that require specialized modifications in the autoregressive setting.

5.2. Open Challenges

Several challenges constrain the practical competitiveness of discrete diffusion. Inference efficiency remains the most pressing concern: generating 1024 tokens through 1024 denoising steps with full model evaluations at each step incurs computational costs substantially higher than the single forward pass of autoregressive generation. While reduced step counts (128 or 256 steps) maintain reasonable quality, the efficiency gap remains significant for latency-sensitive applications. The development of adaptive step-count strategies and distilled samplers that maintain quality with fewer evaluations represents a critical path toward practical deployment.

Evaluation methodology itself poses an open challenge. As this analysis demonstrates, single-metric comparisons can be misleading when the entropy characteristics of generated text differ between methods. The field would benefit from standardized evaluation protocols that report quality–diversity frontiers rather than point estimates, enabling fairer comparison across paradigms with structurally different inference procedures. Scaling behavior beyond the GPT-2 parameter regime requires systematic investigation, as the performance dynamics at 7B–70B parameters may differ qualitatively from those observed at 124M–1.7B. The DiffuGPT adaptation pathway, which repurposes pretrained autoregressive weights for diffusion training, suggests a computationally efficient route to exploring large-scale discrete diffusion that merits focused study.

References

[1]. Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., & Poole, B. (2021). Score-based generative modeling through stochastic differential equations. In Proceedings of the 9th International Conference on Learning Representations (ICLR 2021).

[2]. Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2020). The curious case of neural text degeneration. In Proceedings of the 8th International Conference on Learning Representations (ICLR 2020).

[3]. Austin, J., Johnson, D. D., Ho, J., Tarlow, D., & van den Berg, R. (2021). Structured denoising diffusion models in discrete state-spaces. In Advances in Neural Information Processing Systems 34 (NeurIPS 2021), pp. 17981–17993.

[4]. Lou, A., Meng, C., & Ermon, S. (2024). Discrete diffusion modeling by estimating the ratios of the data distribution. In Proceedings of the 41st International Conference on Machine Learning (ICML 2024).

[5]. Sahoo, S. S., Arriola, M., Gokaslan, A., Marroquin, E. M., Rush, A. M., Schiff, Y., Chiu, J. T., & Kuleshov, V. (2024). Simple and effective masked diffusion language models. In Advances in Neural Information Processing Systems 37 (NeurIPS 2024).

[6]. Li, Y., Zhou, K., Zhao, W. X., & Wen, J.-R. (2023). Diffusion models for non-autoregressive text generation: A survey. In Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI 2023), pp. 6745–6753.

[7]. Li, X. L., Thickstun, J., Gulrajani, I., Liang, P., & Hashimoto, T. B. (2022). Diffusion-LM improves controllable text generation. In Advances in Neural Information Processing Systems 35 (NeurIPS 2022).

[8]. Dieleman, S., Sartran, L., Roshannai, A., Savinov, N., Ganin, Y., et al. (2022). Continuous diffusion for categorical data. arXiv preprint arXiv:2211.15089.

[9]. Campbell, A., Benton, J., De Bortoli, V., Rainforth, T., Deligiannidis, G., & Doucet, A. (2022). A continuous time framework for discrete denoising models. In Advances in Neural Information Processing Systems 35 (NeurIPS 2022).

[10]. Lin, Z., Gong, Y., Shen, Y., Wu, T., Fan, Z., Lin, C., Duan, N., & Chen, W. (2023). Text generation with diffusion language models: A pre-training approach with continuous paragraph denoise. In Proceedings of the 40th International Conference on Machine Learning (ICML 2023), PMLR 202, pp. 21051–21064.

[11]. Gulrajani, I., & Hashimoto, T. B. (2023). Likelihood-based diffusion language models. In Advances in Neural Information Processing Systems 36 (NeurIPS 2023).

[12]. Lovelace, J., Kishore, V., Wan, C., Shekhtman, E., & Weinberger, K. Q. (2023). Latent diffusion for language generation. In Advances in Neural Information Processing Systems 36 (NeurIPS 2023).

[13]. Pillutla, K., Swayamdipta, S., Zellers, R., Thickstun, J., Welleck, S., Choi, Y., & Harchaoui, Z. (2021). MAUVE: Measuring the gap between neural text and human text using divergence frontiers. In Advances in Neural Information Processing Systems 34 (NeurIPS 2021).

[14]. Gat, I., Remez, T., Shaul, N., Kreuk, F., Chen, R. T. Q., Synnaeve, G., Adi, Y., & Lipman, Y. (2024). Discrete flow matching. In Advances in Neural Information Processing Systems 37 (NeurIPS 2024).

[15]. Han, X., Kumar, S., & Tsvetkov, Y. (2023). SSD-LM: Semi-autoregressive simplex-based diffusion language model for text generation and modular control. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023), pp. 11575–11596.

[16]. Shi, J., Han, K., Wang, Z., Doucet, A., & Titsias, M. K. (2024). Simplified and generalized masked diffusion for discrete data. In Advances in Neural Information Processing Systems 37 (NeurIPS 2024).

[17]. Wu, T., et al. (2023). AR-Diffusion: Auto-regressive diffusion for text generation. In Advances in Neural Information Processing Systems 36 (NeurIPS 2023).

[18]. Gong, S., Li, M., Feng, J., Wu, Z., & Kong, L. (2023). DiffuSeq: Sequence to sequence text generation with diffusion models. In Proceedings of the 11th International Conference on Learning Representations (ICLR 2023).

[19]. Gong, S., Agarwal, S., Zhang, Y., Ye, J., Zheng, L., Li, M., et al. (2025). Scaling diffusion language models via adaptation from autoregressive models. In Proceedings of the 13th International Conference on Learning Representations (ICLR 2025).

[20]. Nie, S., Zhu, F., You, Z., Zhang, X., Ou, J., Hu, J., Zhou, J., Lin, Y., Wen, J.-R., & Li, C. (2025). Large language diffusion models. arXiv preprint arXiv:2502.09992.

[21]. Ye, J., Gong, S., Chen, L., Zheng, L., Gao, J., Shi, H., Wu, C., Li, Z., Bi, W., & Kong, L. (2024). Diffusion of thought: Chain-of-thought reasoning in diffusion language models. In Advances in Neural Information Processing Systems 37 (NeurIPS 2024).

[22]. Zhao, F., Zhang, M., Zhou, S., & Lou, Q. (2024). Application of deep reinforcement learning for cryptocurrency market trend forecasting and risk management.

[23]. Chen, Y., Chen, Z., & Zou, D. (2025). CarbonShift: Harnessing Grid Carbon Variability for Geo-Distributed Workload Scheduling. Artificial Intelligence and Machine Learning Review, 6(4), 18-31.

[24]. Li, Z., & Chen, Z. (2025). Performance Evaluation of Prompt Generation Strategies for AI Agents in Online Programming Education. Journal of Advanced Computing Systems, 5(9), 14-27.

[25]. Chen, Y., & Chen, Z. (2025). Multi-Objective Deep Reinforcement Learning for Carbon-Aware Spatiotemporal Workload Scheduling in Geo-Distributed Data Centers. Journal of Advanced Computing Systems, 5(10), 18-30.