LangFlow: Continuous Diffusion Rivals Discrete in Language Modeling

Yuxin Chen[*]
UIUC
nealchen@illinois.edu

Chumeng Liang[*]
UIUC
chumengl@illinois.edu

Hangke Sui[*]
UIUC
hangkes2@illinois.edu

Ruihan Guo
UIUC
ruihang6@illinois.edu

Chaoran Cheng
UIUC
chaoran7@illinois.edu

Jiaxuan You
UIUC
jiaxuan@illinois.edu

Ge Liu[†]
UIUC
geliu@illinois.edu

Abstract

Continuous diffusion has served as a foundation for high-fidelity, controllable, and few-step generation across continuous data modalities such as images, videos, and molecular structures. However, in language modeling, prior continuous diffusion language models (DLMs) lag behind discrete counterparts. Existing categorical and simplex-based approaches operate over extremely large and sparse language spaces, while prior embedding-space approaches avoid this sparsity but lack a well-selected design space. In this work, we close this gap with **LangFlow** by connecting embedding-space DLMs to Flow Matching, alongside three key innovations: (1) we derive a novel ODE-based NLL bound for principled evaluation of continuous flow-based language models; (2) we propose an information-uniform principle for setting the noise schedule, which motivates a learnable noise scheduler based on a Gumbel distribution; and (3) we revise prior training protocols by incorporating self-conditioning, which improves both likelihood and sample quality for embedding-space DLMs and behaves differently from its use in discrete diffusion. Putting everything together, LangFlow is competitive with top discrete DLMs on generative quality, reaching a generative perplexity of **92.3** on LM1B and **36.5** on OpenWebText. It also outperforms discrete counterparts on reasoning tasks like Sudoku. As the first continuous DLM shown to rival discrete diffusion in generative quality, LangFlow provides clear evidence that continuous diffusion is a promising paradigm for language modeling. Code: https://github.com/nealchen2003/LangFlow.

1 Introduction

Diffusion models [Sohl-Dickstein et al., 2015, Song and Ermon, 2019, Ho et al., 2020] have achieved remarkable success in generating data with continuous modalities, such as images [Dhariwal and Nichol, 2021, Rombach et al., 2022], videos [Ho et al., 2022, Blattmann et al., 2023], molecular structures [Watson et al., 2023, Abramson et al., 2024], and robot behaviors [Chi et al., 2025]. A

[*]Equal contribution
[†]Corresponding author

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

key reason behind this success is the flexibility of *continuous diffusion*: it admits expressive latent trajectories, stable ODE/SDE-based sampling, and a mature toolbox of techniques such as self-conditioning, trajectory editing, and potential few-step acceleration via flow-based distillation. These properties make continuous diffusion a powerful and versatile paradigm for generative modeling. This success motivates extending diffusion models to categorical data, including graphs [Vignac et al., 2022] and sequences [Austin et al., 2021], contributing to the broader goal of unifying cross-modality data generation under a single architecture [Zhou et al., 2024, Rojas et al., 2025].

The most prominent milestone yet to be conquered for categorical diffusion is language modeling, where diffusion has morphed into two distinct formulations. On one hand, discrete diffusion [Austin et al., 2021, Lou et al., 2023, Sahoo et al., 2024] reformulates diffusion processes directly in the categorical space based on the continuous-time Markov chain (CTMC) theory. While achieving competitive performance, discrete diffusion sacrifices expressive latent spaces, which limits controllable and few-step generation. On the other hand, simplex diffusion [Cheng et al., 2024, Song et al., 2025, Cheng et al., 2025a] applies continuous diffusion over the probability simplex of language data. Although theoretically sound and capable of preserving diffusion tricks, this approach suffers from the extreme sparsity of the simplex space, making score estimation particularly challenging.

These shortcomings refocus our attention on embedding-space diffusion [Li et al., 2022, Dieleman et al., 2022, Gulrajani and Hashimoto, 2023], an overlooked family of diffusion language models that inherently avoid sparsity and offer editability. Embedding-space diffusion generates token embeddings progressively from Gaussian noise with editable generation paths and denser data spaces. This makes embedding-space diffusion a promising direction to explore for language modeling.

However, exploration of embedding-space DLMs has long been blocked by several challenges. First, the theoretical grounding of embedding-space DLMs remains unclear. Training objectives in prior work are either heuristic or cumbersome to implement, for example requiring dynamically sliced batches to optimize different objectives [Gulrajani and Hashimoto, 2023]. Second, a reliable ODE-based estimation of perplexity (PPL) – the primary evaluation metric in language modeling – has yet to be established for embedding-space DLMs, as prior works rely exclusively on SDE-based bounds. Third, the absence of PPL evaluation further obscures the goal of optimizing training techniques, leaving good design choices unclear. Without a solid theoretical foundation for evaluation and improved training techniques, the potential of embedding-space diffusion language models remains difficult to assess and improve.

In this paper, we observe that the ODE formulation of embedding-space DLMs naturally aligns with Flow Matching [Lipman et al., 2023] (Section 3.1). This theoretical connection allows us to derive a novel ODE-based upper bound of negative log-likelihood (NLL), which tackles the challenge of evaluating embedding-space DLMs and complements prior SDE-based bounds (Section 3.2). Based on this framework, we further explore the training techniques of LangFlow. First, we show that the noise scheduler for embedding-space diffusion should differ substantially from those used in image diffusion because of differences in data geometry. We then propose the information-uniform principle to obtain a noise schedule suited to language (Section 4.1). Second, we reveal how the effects of self-conditioning differ between discrete and continuous diffusion and refine the training protocol of embedding-space DLMs (Section 4.2). Both training techniques substantially improve the performance of LangFlow.

Contributions: To summarize, our contributions are three-fold:

1) We rethink embedding-space diffusion language models through the lens of Flow Matching. This unlocks key theoretical insights and empirical gains, including a theoretically grounded cross-entropy loss and a novel ODE-based negative log-likelihood (NLL) bound that provides an evaluation protocol aligned with deterministic ODE sampling for embedding-space DLMs. The embedding space formulation circumvents the high-dimensionality bottlenecks in diffusion models on the probability simplex and one-hot space.

2) We elucidate the design space of embedding-space DLMs and propose two key design choices that greatly improve training efficiency and sampling quality. First, we introduce an information-uniform noise schedule with a new time-conditioning based on the logarithmic noise-to-signal ratio γ. Our extensive profiling reveals a strong empirical fit to a Gumbel distribution over γ, an observation that differs sharply from common conclusions in image generation. Second, we reveal the underexplored discrepancy of self-conditioning between its effect on discrete diffusion and that on continuous

diffusion, rectifying the training recipe of continuous diffusion language modeling. Both design choices substantially enhance the performance of embedding-space DLMs.

3) Combining everything together, we introduce LangFlow, the first continuous DLM shown to rival discrete DLMs on both language modeling benchmarks and reasoning tasks. Specifically, LangFlow achieves a generative perplexity (Gen. PPL) of 92.3 on LM1B and 36.5 on OpenWebText, surpassing all discrete diffusion baselines. We also validate the clear advantage of LangFlow on Sudoku compared to discrete counterparts. LangFlow re-establishes the baseline for continuous diffusion language modeling, enabling future extensions to be built on a stronger foundation comparable to discrete diffusion.

2 Preliminaries

Flow Matching (FM) [Lipman et al., 2023] is a generative modeling paradigm that learns a velocity field $\boldsymbol{u}_t(\boldsymbol{z}_t)$ to transport a simple prior p_{prior} (e.g., standard Gaussian) to the data distribution p_{data}. Starting from $\boldsymbol{z}_0 \sim p_{\text{prior}}$, solving the ordinary differential equation (ODE) $\mathrm{d}\boldsymbol{z}_t = \boldsymbol{u}_t(\boldsymbol{z}_t)\,\mathrm{d}t$ yields $\boldsymbol{z}_1 \sim p_{\text{data}}$.

The marginal velocity field $\boldsymbol{u}_t(\boldsymbol{z}_t)$ is constructed by marginalizing conditional velocity fields. Under mild regularity conditions, if a conditional velocity field $\boldsymbol{u}_t(\boldsymbol{z}_t \mid \boldsymbol{z})$ transforms $\boldsymbol{z}_0 \sim p_{\text{prior}}$ to a specific data point $\boldsymbol{z}_1 = \boldsymbol{z}$, the marginal velocity field $\boldsymbol{u}_t(\boldsymbol{z}_t)$ generates p_{data}:

$$\boldsymbol{u}_t(\boldsymbol{z}_t) = \mathbb{E}[\boldsymbol{u}_t(\boldsymbol{z}_t \mid \boldsymbol{z}) \mid \boldsymbol{z}_t] = \int \boldsymbol{u}_t(\boldsymbol{z}_t \mid \boldsymbol{z}) p(\boldsymbol{z} \mid \boldsymbol{z}_t)\,\mathrm{d}\boldsymbol{z}. \tag{1}$$

Typical FM models utilize an affine Gaussian probability path as the conditional flow:

$$\boldsymbol{z}_t = \alpha_t \boldsymbol{z} + \sigma_t \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim p_{\text{prior}}, \tag{2}$$

which corresponds to the conditional velocity field:

$$\boldsymbol{u}_t(\boldsymbol{z}_t \mid \boldsymbol{z}) = \dot{\sigma}_t \frac{\boldsymbol{z}_t - \alpha_t \boldsymbol{z}}{\sigma_t} + \dot{\alpha}_t \boldsymbol{z}, \tag{3}$$

where α_t, σ_t are differentiable scalar schedules specifying the transformation, satisfying $\alpha_0 = 0$, $\sigma_0 = 1$ and $\alpha_1 = 1, \sigma_1 = 0$, with time derivatives $\dot{\alpha}_t, \dot{\sigma}_t$.

In practice, computing the true marginal velocity field $\boldsymbol{u}_t(\boldsymbol{z}_t)$ is intractable. Instead, it is approximated by a neural network $\boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{z}_t, t)$, trained via conditional Flow Matching:

$$\mathcal{L}_{\text{FM}}(\boldsymbol{\theta}) = \int_0^1 \mathbb{E}\left[\|\boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{z}_t, t) - \boldsymbol{u}_t(\boldsymbol{z}_t)\|^2\right] \mathrm{d}t = \int_0^1 \mathbb{E}\left[\|\boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{z}_t, t) - \boldsymbol{u}_t(\boldsymbol{z}_t \mid \boldsymbol{z})\|^2\right] \mathrm{d}t + \text{const}, \tag{4}$$

where the constant term is independent of $\boldsymbol{\theta}$ and thus omitted during optimization.

Since $\boldsymbol{u}_t(\boldsymbol{z}_t \mid \boldsymbol{z})$ is linear in $\boldsymbol{z}$, the velocity network can be equivalently parameterized using a denoiser $\hat{\boldsymbol{z}}_{\boldsymbol{\theta}}(\boldsymbol{z}_t, t) \approx \mathbb{E}[\boldsymbol{z} \mid \boldsymbol{z}_t]$:

$$\boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{z}_t, t) = \boldsymbol{u}_t\big(\boldsymbol{z}_t \mid \hat{\boldsymbol{z}}_{\boldsymbol{\theta}}(\boldsymbol{z}_t, t)\big). \tag{5}$$

Flow Matching is equivalent to learning the probability flow ODE of standard diffusion models [Sohl-Dickstein et al., 2015, Song and Ermon, 2019, Ho et al., 2020] when $p_{\text{prior}} = \mathcal{N}(\mathbf{0}, \mathbf{I})$. Since we consistently use this prior, we use the terms Flow Matching and diffusion models interchangeably.

Related work. We provide a more detailed discussion of related discrete, continuous, and flow-map language diffusion models in Appendix A.

3 LangFlow: continuous language modeling via flow matching

Existing methods for continuous diffusion language modeling typically train models by regressing denoised outputs, embeddings, or one-hot representations of the ground truth, while others rely on heuristic training objectives. Furthermore, they lack a reliable ODE-based upper bound of the negative log-likelihood for evaluation. In this section, we introduce **LangFlow**, a principled framework for continuous diffusion language models. For training, we propose optimizing the cross-entropy loss and establish a connection between this objective and Flow Matching [Lipman et al., 2023] through Bregman divergence. We also derive a novel ODE-based upper bound of the negative log-likelihood for reliable evaluation. Figure 1 provides an overview of the proposed pipeline, including both training and sampling procedures.

3.1 Training flow matching on language

We begin with embedding-space diffusion for continuous diffusion language modeling. Let $\mathbf{E} \in \mathbb{R}^{V \times D}$ be an embedding matrix over a vocabulary of size V, mapping each text token $x^{(i)}$ to a D-dimensional vector $\boldsymbol{e}_{x^{(i)}}$. A token sequence $\boldsymbol{x} = (x^{(1)}, \ldots, x^{(L)})$ is embedded as $\boldsymbol{z} = (\boldsymbol{e}_{x^{(1)}}, \ldots, \boldsymbol{e}_{x^{(L)}}) \in \mathbb{R}^{L \times D}$, which serves as the generative target.

Flow Matching (FM) [Lipman et al., 2023] defines a velocity field $\boldsymbol{u}_t(\boldsymbol{z}_t)$ that transports a Gaussian prior $p_{\text{prior}} = \mathcal{N}(\mathbf{0}, \mathbf{I})$ to the data distribution p_{data} through the ODE $\mathrm{d}\boldsymbol{z}_t = \boldsymbol{u}_t(\boldsymbol{z}_t)\ \mathrm{d}t$. The standard formulation uses time $t \in [0, 1]$ and an affine probability path with scalar schedules α_t and σ_t:

$$\boldsymbol{z}_t = \alpha_t \boldsymbol{z} + \sigma_t \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{6}$$

where α_t grows from $0 \to 1$ and σ_t decays from $1 \to 0$ as t goes from 0 to 1. While natural, this formulation ties both the objective and the dynamics to a specific choice of time schedule.

Time conditioning through γ-path. To reduce this schedule dependence, we reparameterize the flow path using the logarithmic noise-to-signal ratio (logNSR) γ. The motivation is that denoising difficulty is primarily controlled by noise level rather than by an arbitrary time index; this is also consistent with Variational Diffusion Models [Kingma et al., 2021]. For any valid path of the form $\boldsymbol{z}_t = \alpha_t \boldsymbol{z} + \sigma_t \boldsymbol{\epsilon}$, we define $\gamma_t = \log(\sigma_t^2 / \alpha_t^2)$. In typical settings, $\dot{\alpha}_t > 0$ and $\dot{\sigma}_t < 0$, so γ_t is strictly monotone decreasing. This induces a bijection between $t \in [0, 1]$ and $\gamma \in \mathbb{R} \cup \{\pm\infty\}$, effectively reparameterizing the flow in terms of γ instead of t. Since $\alpha_0 = 0$ and $\sigma_1 = 0$, the endpoints correspond to pure noise at $t = 0$ ($\gamma = +\infty$) and clean data at $t = 1$ ($\gamma = -\infty$).

Under this change of variables, the marginal distributions are preserved, and the training objective is equivalent up to the Jacobian factor $|\dot{\gamma}_t|$. Restricting attention to variance-preserving (VP) paths satisfying $\alpha_\gamma^2 + \sigma_\gamma^2 = 1$, we can rewrite the path directly in terms of γ:

$$\boldsymbol{z}_\gamma = \alpha_\gamma \boldsymbol{z} + \sigma_\gamma \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \qquad \gamma \in \mathbb{R} \cup \{\pm\infty\}, \tag{7}$$

with

$$\sigma_\gamma^2 = \operatorname{sigmoid}(\gamma), \qquad \alpha_\gamma^2 = 1 - \sigma_\gamma^2 = \operatorname{sigmoid}(-\gamma). \tag{8}$$

We refer to this flow simply as the γ-path. This unification inherently encompasses standard diffusion trajectories and makes explicit that learning is anchored to the noise level γ, rather than to a particular time parameterization. We base our implementation on the γ-path and use γ as our time conditioning variable instead of t; further motivation is provided in Section 4.1.

Training flow matching via cross entropy. We now propose our cross-entropy training objective for Flow Matching on categorical data like language, which is closely related to Variational Flow Matching (VFM) [Eijkelboom et al., 2024] (Appendix B.3) and is a special case of Bregman divergence minimization [Guzmán-Cordero et al., 2025] (Appendix B.4).

Let the model output be $\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\boldsymbol{z}_\gamma, \gamma)$, whose (i, k) entry approximates $\Pr(x^{(i)} = k \mid \boldsymbol{z}_\gamma)$. Let $\mathbf{1}_x$ denote the one-hot vector of token x. Let $\pi(\gamma)$ be a noise sampling distribution. We define the training objective along the γ-path as

$$\mathcal{L}_{\text{CE}}(\boldsymbol{\theta}) = \mathbb{E}_{\gamma \sim \pi}[\ell_{\text{CE}}(\gamma)], \text{ where } \ell_{\text{CE}}(\gamma) = \mathbb{E}\left[-\frac{1}{L}\sum_{i=1}^{L} \log \hat{x}_{\boldsymbol{\theta}}^{(i, x^{(i)})}(\boldsymbol{z}_\gamma, \gamma)\right]. \tag{9}$$

Note that $\mathcal{L}_{\text{CE}}$ is equivalent to KL divergence up to an additive constant:

$$\mathbb{E}\Big[-\log \hat{x}_{\boldsymbol{\theta}}^{(i, x^{(i)})}(\boldsymbol{z}_\gamma, \gamma) \;\Big|\; \boldsymbol{z}_\gamma\Big] = \mathcal{D}_{\text{KL}}\big(P(x^{(i)} \mid \boldsymbol{z}_\gamma) \,\big\|\, \hat{\boldsymbol{x}}_{\boldsymbol{\theta}}^{(i)}(\boldsymbol{z}_\gamma, \gamma)\big) + \mathcal{H}\big(x^{(i)} \mid \boldsymbol{z}_\gamma\big), \tag{10}$$

where $P(x^{(i)} \mid \boldsymbol{z}_\gamma)$ denotes the posterior distribution of $x^{(i)}$ given $\boldsymbol{z}_\gamma$, and $\mathcal{H}\big(x^{(i)} \mid \boldsymbol{z}_\gamma\big)$ denotes the posterior entropy. Therefore, for each token position, the unique optimum of this factorized prediction objective is $\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}^{(i)} = P(x^{(i)} \mid \boldsymbol{z}_\gamma)$.

Given predicted token probabilities $\hat{x}_{\boldsymbol{\theta}}^{(i)}$, we recover a continuous denoiser by taking the embedding expectation of $\hat{x}_{\boldsymbol{\theta}}^{(i)}$ over the vocabulary:

$$\hat{\boldsymbol{z}}_{\boldsymbol{\theta}}^{(i)}(\boldsymbol{z}_\gamma, \gamma) = \sum_{k=1}^{V} \hat{x}_{\boldsymbol{\theta}}^{(i,k)}(\boldsymbol{z}_\gamma, \gamma)\, \boldsymbol{e}_k = \mathbf{E}^\top \hat{\boldsymbol{x}}_{\boldsymbol{\theta}}^{(i)}(\boldsymbol{z}_\gamma, \gamma). \tag{11}$$

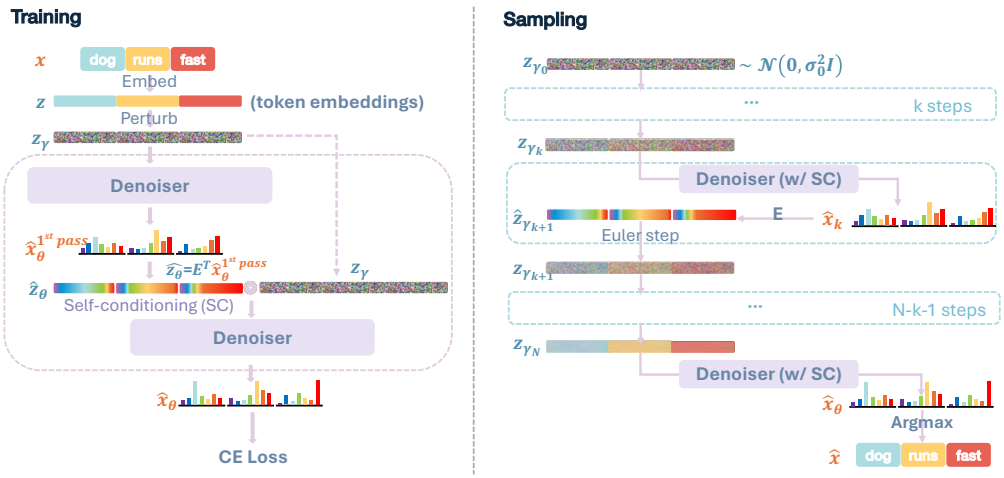


Figure 1: **LangFlow takes noisy token embeddings as inputs and predicts clean probabilities.** In the **training (*left*)**, we first map discrete token sequence $\boldsymbol{x}$ to token embeddings $\boldsymbol{z}$ via a learnable embedding matrix $\mathbf{E}$. Then, we perturb $\boldsymbol{z}$ into noisy embeddings $\boldsymbol{z}_\gamma$ according to γ, the sampled logarithmic noise-to-signal ratio. Finally, we use a denoiser network to predict the categorical distribution $\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\boldsymbol{z}_\gamma, \gamma)$ and supervise this prediction with the cross-entropy loss. Self-conditioning is randomly activated. In the **sampling (*right*)**, we start from Gaussian noise $\boldsymbol{z}_{\gamma_0}$ and iteratively denoise it towards clean token embeddings $\boldsymbol{z}_{\gamma_N}$. We define the moving target $\hat{\boldsymbol{z}}_{\boldsymbol{\theta}}$ for every step by embedding predicted clean probabilities $\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}$ with $\mathbf{E}$. The final token sequence is obtained by $\arg\max$ decoding of $\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\boldsymbol{z}_{\gamma_N}, \gamma_N)$.

This establishes a direct coupling between discrete likelihood training and continuous flow estimation: the model is optimized in token space via CE, while the associated continuous denoiser is obtained deterministically from token probabilities for ODE-based sampling.

Training pipeline. Figure 1 (Left) summarizes our training pipeline. We first embed language tokens $\boldsymbol{x}$ with a learnable embedding matrix $\mathbf{E}$. After that, we perturb the clean embeddings $\boldsymbol{z}$ to the noisy embeddings $\boldsymbol{z}_\gamma$. Our model is then applied to predict the token probabilities $\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}$ from $\boldsymbol{z}_\gamma$. We use the CE loss in Equation (9) to train the model. The complete training procedure of LangFlow is summarized in Algorithm 1 (deferred to Appendix B.1).

Under this pipeline, our model $\boldsymbol{\theta}$ is fed with (noisy) embeddings $\boldsymbol{z}_\gamma \in \mathbb{R}^{L \times D}$ and predicts token probabilities $\hat{\boldsymbol{x}}_{\boldsymbol{\theta}} \in \mathbb{R}^{L \times V}$, similar to discrete diffusion and autoregressive language models. This means that we can share the same network architecture with discrete DLMs [Sahoo et al., 2024, 2025a], which is a modified version of Diffusion Transformers [Peebles and Xie, 2023].

3.2 ODE sampling and the bound for negative log likelihood

Continuous diffusion models typically offer two sampling paradigms: Stochastic Differential Equations (SDEs) or Probability Flow Ordinary Differential Equations (PF-ODEs) [Song et al., 2021].

Why ODE over SDE? Although SDEs are common in diffusion frameworks, their continuous noise injection destroys the deterministic bijective mapping between the prior and data distributions. This stochasticity inherently resists flow-based distillation techniques like Consistency Models [Song et al., 2023]. By exclusively using the deterministic ODE, LangFlow preserves this crucial bijection, enabling future acceleration into efficient few-step generators.

Sampling pipeline. Figure 1 (Right) illustrates the pipeline of our ODE sampling. In practice, we solve the ODE over a finite γ range $[a, b]$ rather than the whole extended real line $\mathbb{R} \cup \{\pm\infty\}$. Given N sampling steps, we set the starting point $\gamma_0 = b$ and the end point $\gamma_N = a$ and distribute γ_k within $[a, b]$ according to a pre-defined discretization (see Section 4.1). The initial state $\boldsymbol{z}_{\gamma_0}$ is sampled from $\mathcal{N}(\mathbf{0}, \sigma_b^2 \mathbf{I})$. At each sampling step, we run the model forward to obtain the token probabilities $\hat{\boldsymbol{x}}_k$, embed $\hat{\boldsymbol{x}}_k$ to derive the denoised embedding $\hat{\boldsymbol{z}}_k$, and use $\hat{\boldsymbol{z}}_k$ to determine the solver direction. The solver then produces $\boldsymbol{z}_{k+1}$ as the input to the next sampling step. After integrating the ODE down to

$\gamma_N = a$ by the above iteration, we take a final token prediction $\boldsymbol{x}^{(i)} = \arg\max \hat{\boldsymbol{x}}_{\boldsymbol{\theta}}^{(i)}(\cdot \mid \boldsymbol{z}_a, a)$ with an extra model forward. In our implementation, we follow existing literature and use a first-order solver along the γ-path, detailed in Appendix B.2. Our ODE sampling procedure is outlined in Algorithm 2.

Perplexity estimation. Perplexity (PPL) remains the primary metric for evaluating data likelihood in language models. Previous methods [Gulrajani and Hashimoto, 2023] compute PPL using stochastic bounds (NELBO). Based on the continuous Flow Matching formulation, we can estimate NLL in our framework by integrating along the ODE path. Specifically, by slightly modifying the final step from taking $\arg\max$ to sampling $\boldsymbol{x}^{(i)} \sim \hat{\boldsymbol{x}}_{\boldsymbol{\theta}}^{(i)}(\cdot \mid z_a, a)$, the following theorem gives a novel ODE-based upper bound for estimating perplexity under deterministic ODE sampling.

Theorem 3.1. *Assume the ODE solution on $[a, b]$ is well-defined and invertible, with an integrable divergence term. For a sequence $\boldsymbol{x} = (x^{(1)}, \ldots, x^{(L)})$ of length L and embedding dimension D, the log-likelihood of LangFlow has the following evidence lower bound:*

$$\log P_{\boldsymbol{\theta}}(\boldsymbol{x}) \geq \mathbb{E}_{q_a(\boldsymbol{z}_a \mid \boldsymbol{x})}\left[\frac{LD}{2} - \frac{\|\boldsymbol{z}_b\|^2}{2\sigma_b^2} + \sum_{i=1}^{L} \log \hat{\boldsymbol{x}}_{\boldsymbol{\theta}}^{(i,x^{(i)})}(\boldsymbol{z}_a, a) - \int_a^b \frac{\alpha_\gamma}{2} \nabla \cdot \hat{\boldsymbol{z}}_{\boldsymbol{\theta}}(\boldsymbol{z}_\gamma, \gamma)\, \mathrm{d}\gamma\right], \quad (12)$$

where $q_a(\boldsymbol{z}_a \mid \boldsymbol{x}) = \mathcal{N}(\alpha_a \mathbf{E}^{\mathrm{T}} \boldsymbol{x}, \sigma_a^2 \mathbf{I})$, $\boldsymbol{z}_\gamma (a \leq \gamma \leq b)$ is the reverse ODE trajectory given $\boldsymbol{z}_a$, and $\nabla \cdot \hat{\boldsymbol{z}}_{\boldsymbol{\theta}}$ represents the divergence w.r.t. $\boldsymbol{z}_\gamma$.

Proof. The proof is deferred to Appendix C. □

4 Improved design choices of continuous diffusion language models

We next identify two key design choices for training continuous diffusion language models like LangFlow: noise scheduling and self-conditioning. In Section 4.1, we show that language exhibits a fundamentally different noise geometry from the common practice of diffusion in the image and video domains. In Section 4.2, we show that self-conditioning strengthens continuous diffusion differently from discrete diffusion. Putting the two choices together allows LangFlow to match the sample quality of discrete diffusion.

4.1 Noise scheduler

Training continuous diffusion requires choosing how to schedule the noise level. We first follow strong diffusion baselines such as Stable Diffusion 3 [Esser et al., 2024] and EDM [Karras et al., 2022] and use t as our time conditioning, with uniform noise scheduling by default. We plot our cross-entropy loss w.r.t. t at 50k training steps over the entire LM1B dataset [Chelba et al., 2013] to determine which region of t should receive more training steps.

Figure 2 (Left) visualizes the result. We observe that our loss is nearly zero for $t \in [0.2, 1.0]$. This suggests that **1)** during training, the model can already predict the correct token for $t > 0.2$, and **2)** during sampling, the model contributes little new information through $t \in [0.2, 1.0]$. Therefore, uniform noise schedulers based on t allocate more than half of their training and sampling steps to noise levels that carry little useful information. This is intuitive because generation paths toward categorical data have finite and isolated destinations, which can remain distinguishable even at high noise levels. The noise schedule should therefore depend strongly on the geometry of the underlying data distribution, so we cannot simply transfer image-diffusion schedules to this setting.

To optimize the noise scheduler of LangFlow, we use the following techniques:

Time conditioning on γ-path. We observe from Figure 2 (Left) that most loss reduction occurs in the very noisy region where the signal-to-noise ratio (SNR) is below 0.0625. Hence, we use time conditioning based on $\gamma_t = \log(\sigma_t^2/\alpha_t^2)$, the logarithmic noise-to-signal ratio (NSR), in Section 3.1. The main advantage is that the time conditioning shifts linearly when NSR scales exponentially, expanding the original noisy region $t \in [0, 0.2]$ to a broader interval. This allows the network to better utilize the resolution of time conditioning and makes the loss profile easier to capture.

Information-uniform principle in noise scheduling. We observe that the average CE loss $\ell_{\mathrm{CE}}(\gamma)$ at noise level γ remains stable across training stages (Figure 2, middle), regardless of the noise

scheduler. Using the decomposition in Equation (10), the KL term quickly approaches 0 early in training, leaving $\ell_{\mathrm{CE}}(\gamma)$ dominated by the irreducible posterior entropy $H_\gamma = \frac{1}{L}\sum_{i=1}^{L} H(x^{(i)} \mid \boldsymbol{z}_\gamma)$. This explains why the loss values remain stable after the early training stage: the main component of the loss is a property of the data and cannot be reduced by optimization.

Diffusion sampling can then be described as a process that progressively gains information, i.e., reduces entropy H_γ to zero, where the derivative $H'_\gamma = \frac{\mathrm{d}H_\gamma}{\mathrm{d}\gamma}$ measures the rate at which information about the clean tokens is gained per unit change in γ. We therefore propose a noise scheduling principle that allocates information gain uniformly over the noise density, so that each sampling step receives an equal amount of information. This suggests that both training effort and sampling steps should be concentrated in regions where H'_γ is large.

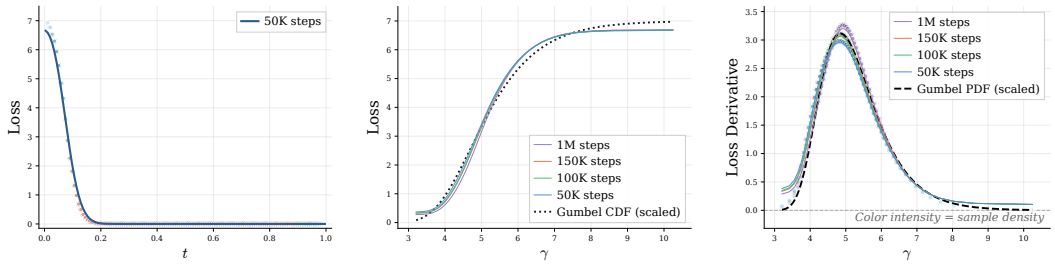


Figure 2: **Loss geometry.** (Left) Training loss as a function of the OT flow matching time $t = \mathrm{sigmoid}(-\gamma/2)$. (Middle) Loss over $\gamma = \log \mathrm{NSR}$ reveals a consistent geometry across training stages. (Right) The derivative $\partial\mathcal{L}/\partial\gamma$ concentrates in a narrow region, forming a stable structure.

Scheduling the noise by the Gumbel distribution of γ. We follow the above principle to define a noise distribution whose density $\pi(\gamma) \propto H'_\gamma$. We use the empirical mean loss across checkpoints at different training steps as a surrogate for H_γ and apply finite-difference smoothing. The curve of H'_γ is positively skewed (Figure 2, right) and is well matched by a Gumbel distribution:

$$H_\gamma = H_{+\infty} \cdot \exp\left(-\exp\left(-\frac{\gamma - P_\mu}{P_\beta}\right)\right), \tag{13}$$

where $H_{+\infty}$ is the average posterior entropy at pure noise and P_μ, P_β are the location and scale parameters of the Gumbel distribution. We observe that the fitted Gumbel distribution drifts subtly as training evolves. Hence, we make $H_{+\infty}$, P_μ, and P_β learnable parameters, trained by a scheduler loss $\mathcal{L}_{\mathrm{Scheduler}} = \mathbb{E}[(\ell_{\mathrm{CE}}(\gamma) - H_\gamma)^2]$. We use $\mathcal{L}_{\mathrm{Scheduler}}$ to train the noise scheduler and $\mathcal{L}_{\mathrm{CE}}$ to train the main model, with stop-gradient operations separating their updates.

Implementation. In practice, we clip the range of γ to $[a, b]$, where a and b are the lower and upper 10^{-5} quantiles of the Gumbel distribution, respectively. During training, we sample γ from the Gumbel distribution. During sampling with N steps, we set the intermediate schedule to the i/N quantiles of the Gumbel distribution for $i = 1, \ldots, N-1$. This choice significantly improves the sample quality of LangFlow, reducing the generative perplexity from $\sim$1000 to 154.2.

4.2 Self-conditioning

Self-conditioning [Chen et al., 2022] is an established technique in diffusion models and is generally known to improve generation quality and perplexity (Gen. PPL and PPL, defined in Section 5). Specifically, it feeds the previous step's prediction $\hat{\boldsymbol{z}}_{\boldsymbol{\theta}}(\boldsymbol{z}_\gamma, \gamma)$ back as an auxiliary input to the network. During training, self-conditioning is applied randomly with a certain probability by feeding the model its own prediction for one round, so the model learns to exploit the prior prediction when available. During sampling, self-conditioning is always enabled: it is initialized to zero at the first step and set to the previous prediction at subsequent steps. Figure 1 shows how self-conditioning is incorporated into training and sampling.

Despite these general benefits, prior works on discrete diffusion language models typically evaluate PPL with self-conditioning disabled. This evaluation protocol has been carried over to continuous diffusion. We empirically find that applying this protocol to continuous diffusion is inappropriate.

Table 1: Performance comparison on LM1B and OWT. We report upper bounds on perplexity (PPL) for DLMs and generative perplexities (Gen. PPL) based on GPT-2 [Radford et al., 2019]. ‡ denotes results from retrained baselines. We use gold, silver, and bronze to denote the 1st, 2nd, and 3rd place among DLMs, respectively.

Category	Model	LM1B		OWT	
		Gen. PPL↓	PPL↓	Gen. PPL↓	PPL↓
Autoregressive	Transformer	66.7‡	22.8‡	35.9	17.5
Discrete Diffusion (Absorbing)	D3PM Absorb	–	76.9	–	–
	DiffusionBert	–	63.8	–	–
	SEDD Absorb	115.9‡	32.0‡	105.0	24.1
	MDLM	103.9‡	31.0‡	104.9	23.2
Discrete Diffusion (Uniform)	D3PM Uniform	–	137.9	–	–
	SEDD Uniform	–	40.3	99.9	29.7
	UDLM	99.8‡	33.8‡	–	27.4
	Duo	97.6‡	33.6‡	77.7	25.2
Continuous Diffusion	Diffusion-LM	–	118.6	–	–
	Plaid	77.3‡	32.4‡	–	–
	FLM (1024 steps)	96.9	–	62.2	–
	LangFlow (ODE w/o chain rule)	**92.3**	**30.0**	**36.5**	**24.6**
	LangFlow (ODE w/ chain rule)	**92.3**	**45.3**	**36.5**	**33.3**
	LangFlow (SDE)	**59.4**	**44.2**	**23.8**	**32.2**

As shown by the ablation in Table 7, turning self-conditioning on and off for MDLM (discrete) and LangFlow (our continuous model) reveals an asymmetry: for MDLM, self-conditioning improves Gen. PPL but degrades PPL, which helps explain its omission in discrete PPL evaluations. For LangFlow, the picture is different: self-conditioning significantly improves both Gen. PPL and PPL, which is enough to close the PPL gap between continuous and discrete diffusion in our setting.

How the effects of self-conditioning differ across formulations remains an open question; we provide observations in Appendix E.2. Nevertheless, in our experiments, self-conditioning is a crucial component for LangFlow to match the sample quality of discrete diffusion.

5 Experiments

Datasets. We evaluate LangFlow on two standard language modeling benchmarks: LM1B [Chelba et al., 2013] and OpenWebText (OWT) with sequence packing [Raffel et al., 2020], following the existing literature on DLMs. Following Sahoo et al. [2024, 2025a], we reserve the last 100k documents of OWT as the validation set.

Training. Following discrete diffusion baselines [Sahoo et al., 2024, Schiff et al., 2024, Sahoo et al., 2025a], we use a DiT-style Transformer [Peebles and Xie, 2023] with rotary positional encoding [Su et al., 2024]. The model consists of 12 layers, a hidden size of 768, and 12 attention heads (130M parameters). The model is conditioned on γ rather than t and uses a learned Gumbel scheduler to match the empirical information-gain profile (Sec. 4.1). During training, γ is sampled from the learned distribution, and the model is optimized with token-level cross-entropy together with the schedule fitting loss. Self-conditioning is enabled with probability 0.25. We also use a preconditioning skip connection with a 5K-step warmup. Following existing literature [Sahoo et al., 2024, 2025a], we train models for 1M steps on LM1B and OWT with a batch size of 512. We use a context length of 128 and the `bert-base-uncased` tokenizer [Devlin et al., 2019] for LM1B, and a context length of 1024 and the `gpt2-large` tokenizer [Radford et al., 2019] for OWT.

Evaluation. We evaluate the model using generative perplexity (Gen. PPL) and perplexity on the validation dataset (PPL). For Gen. PPL, we follow the protocol in existing works [Sahoo et al., 2024, Schiff et al., 2024] to generate 1024 samples and compute their mean perplexity by GPT2-Large [Radford et al., 2019]. For all methods, samples are generated using 128 sampling steps and a sequence length of 128 on LM1B, and 1024 sampling steps with a sequence length of 1024 on OWT.

Table 2: Zero-shot perplexities (↓) of models trained for 1M steps on OWT. All perplexities for DLMs are upper bounds. † Taken from Sahoo et al. [2024]. ¶ Taken from Lou et al. [2023], where these baselines were trained for 1.3M steps, whereas ours and the other baselines were trained for 1M steps. ‡ Taken from Sahoo et al. [2025a]. We use gold, silver, and bronze to denote the 1st, 2nd, and 3rd place among DLMs, respectively.

Model	PTB	Wikitext	LM1B	Lambada	AG News	Pubmed	Arxiv
Autoregressive							
Transformer†	82.05	25.75	51.25	51.28	52.09	49.01	41.73
Diffusion (Absorbing-state)							
SEDD Absorb†	100.09	34.28	68.20	49.86	62.09	44.53	38.48
D3PM Absorb¶	200.82	75.16	138.92	93.47	–	–	–
MDLM†	95.26	32.83	67.01	47.52	61.15	41.89	37.37
Diffusion (Uniform-state / Continuous)							
SEDD Uniform‡	105.51	41.10	82.62	57.29	82.64	55.89	50.86
Plaid¶	142.60	50.86	91.12	57.28	–	–	–
UDLM‡	112.82	39.42	77.59	53.57	80.96	50.98	44.08
Duo‡	89.35	33.57	73.86	49.78	67.81	44.48	40.39
LangFlow	**81.20**	**32.28**	**68.21**	**46.93**	**69.41**	**46.74**	**38.47**

The only exception is FLM, for which the numbers are taken from its original paper and are based on 1024 sampling steps. For PPL, we report available upper bounds for DLM baselines.

Notably, diffusion models such as LangFlow admit both ODE and SDE formulations [Song et al., 2021]. We therefore evaluate the trained LangFlow model under both frameworks. For the ODE evaluation, Theorem 3.1 establishes an upper bound of PPL for the ODE sampling process. We consider two PPL estimators that differ in whether to detach the self-conditioning in the integral, denoted by *w/o chain rule* and *w/ chain rule* [Yang et al., 2026]. The *w/o chain rule* protocol is used for zero-shot transfer. For the SDE evaluation, we follow the same setup of our Plaid baseline [Gulrajani and Hashimoto, 2023] to report the *SDE*-based PPL bound and use a sampling temperature of 0.9.

Baselines. We compare LangFlow against the autoregressive Transformer and a wide range of discrete and continuous DLMs, including D3PM [Austin et al., 2021], DiffusionBert [He et al., 2023], SEDD [Lou et al., 2023], MDLM [Sahoo et al., 2024], UDLM [Schiff et al., 2024], and Duo [Sahoo et al., 2025a]. We also include continuous baselines such as Diffusion-LM [Li et al., 2022] and Plaid [Gulrajani and Hashimoto, 2023]. Due to the lack of checkpoints on LM1B, we retrained AR, SEDD, MDLM, UDLM, Duo, and Plaid on LM1B for 1M steps, as these baselines represent strong recent methods. For OWT, we use checkpoints provided by the Duo codebase. We retrained baselines with their original network architecture, which is the same as ours except for Plaid [Gulrajani and Hashimoto, 2023]; for Plaid, we use its own codebase and Transformer architecture for training. This is because Plaid performs worse when using the DiT architecture used by LangFlow and the other baselines. To make the comparison fair, we use the same number of Transformer layers, heads, and hidden dimensions as LangFlow, keeping its number of parameters close to ours. Detailed implementation of Plaid is deferred to Appendix D.3. For DiffusionBert, Diffusion-LM, and D3PM, we report the numbers from the original Duo paper [Sahoo et al., 2025a]. We also include FLM [Lee et al., 2026], a concurrent continuous DLM, in our baselines. Since the FLM paper reports Gen. PPL but does not report PPL or specify a likelihood-evaluation protocol, we report the Gen. PPL from its original paper with the same (OWT) or more (LM1B) sampling steps compared to our setting.

Language modeling. Table 1 shows that LangFlow achieves the best Gen. PPL on LM1B and OWT among DLMs. We also report the Gen. PPL under different NFEs and the frontier of Gen. PPL and data entropy in Appendix E.3. Note that the stochastic sampling (SDE) significantly improves the Gen. PPL of LangFlow compared to the deterministic sampler (ODE), matching previous observation to image diffusion models [Karras et al., 2022].

Zero-shot transfer. We evaluate zero-shot transfer following the protocol of prior works [Radford et al., 2019, Lou et al., 2023, Sahoo et al., 2024, 2025a], where models trained on OWT are evaluated on a diverse set of downstream corpora, including PTB, Wikitext, LM1B, Lambada, AG News, PubMed, and Arxiv. As shown in Table 2, LangFlow achieves strong zero-shot performance among

Table 3: **Sudoku exact-match accuracy (%, ↑).** Evaluation on 2,000 validation puzzles per difficulty, where all 89 solution tokens must be correct. Training steps and NFE indicate training and sampling budgets. auto-Gumbel learns the noise schedule; fixed-Gumbel fixes it to $\mathrm{Gumbel}(4.7, 1.0)$.

Model	Training Steps	NFE	Easy	Medium	Hard
MDLM	20K	180	91.45	71.75	49.70
FLM	20K	180	90.60	78.70	48.45
Duo	20K	180	96.15	82.40	52.10
LangFlow (auto-Gumbel)	20K	180	**97.50**	88.45	61.05
LangFlow (fixed-Gumbel)	20K	180	97.30	**92.45**	**76.00**
LangFlow (auto-Gumbel)	**5K**	16	96.00	87.25	59.90

DLMs, ranking first on PTB, Wikitext, and Lambada while remaining competitive across all other domains. Although MDLM attains more first-place results overall, LangFlow outperforms the autoregressive Transformer on 4 out of 7 benchmarks and MDLM on 3 out of 7 benchmarks.

Verifiable structured generation (Sudoku). Beyond language-model-based evaluation metrics, we further evaluate LangFlow on Sudoku [Kim et al., 2025, Ben-Hamu et al., 2025, Deschenaux and Gulcehre, 2026], a verifiable structured generation task that requires global consistency across the generated sequence. We report exact-match accuracy on 2,000 validation puzzles per difficulty level, where a prediction is considered correct only if all solution tokens match the ground truth. We compare against MDLM, FLM, and Duo, trained on the same datasets following the s-FLM [Deschenaux and Gulcehre, 2026] implementation. We report two LangFlow variants: auto-Gumbel learns the noise schedule, while fixed-Gumbel fixes it to $\mathrm{Gumbel}(4.7, 1.0)$ for both training and sampling. As shown in Table 3, LangFlow consistently achieves stronger accuracy across all difficulty levels. Notably, even with only 5K training steps (1/4 of the training budget) and 16 NFEs (1/11 of the sampling budget; see Appendix E.1 for how NFE is counted), These results provide complementary, fully verifiable evidence that embedding-space continuous diffusion can effectively capture global constraints and generate consistent structured outputs. Additional evaluation details and an analysis of the interaction between self-conditioning and embedding geometry are provided in Appendix E.1.

6 Conclusion

We present LangFlow, a principled framework for embedding-space diffusion language modeling grounded in Flow Matching. Our formulation provides a theoretically grounded cross-entropy objective, an ODE-based NLL bound for likelihood evaluation, and a language-specific design space characterized by information-uniform noise scheduling and effective self-conditioning. Empirically, LangFlow achieves strong performance on large-scale benchmarks. It surpasses state-of-the-art discrete diffusion at the same scale in generative quality. These results provide evidence that continuous diffusion can be a competitive paradigm for language modeling. To summarize, LangFlow establishes a stronger baseline for continuous diffusion language modeling, enabling future extensions to build on a foundation comparable to discrete diffusion.

Limitations & future work. Although LangFlow achieves strong perplexity and generative perplexity, its sample entropy remains lower than that of certain discrete baselines, consistent with observations in other continuous diffusion approaches. In our quantitative and qualitative evaluations (see Appendix D.4), we do not observe noticeable degradation in sample quality attributable to this difference. However, reduced entropy may have subtle effects that emerge at larger scales, which fall beyond the scope of this work. We leave further investigation of these effects to future research.

Acknowledgments and Disclosure of Funding

Research was supported in part by the Molecule Maker Lab Institute: An AI Research Institutes program supported by NSF under Award No. 2505932. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s).

References

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024.

Marianne Arriola, Aaron Gokaslan, Justin T Chiu, Zhihan Yang, Zhixuan Qi, Jiaqi Han, Subham Sekhar Sahoo, and Volodymyr Kuleshov. Block diffusion: Interpolating between autoregressive and diffusion language models. *arXiv preprint arXiv:2503.09573*, 2025.

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

Heli Ben-Hamu, Itai Gat, Daniel Severo, Niklas S Nolte, and Brian Karrer. Accelerated sampling from masked diffusion models via entropy bounded unmasking. *Advances in Neural Information Processing Systems*, 38:55981–56007, 2025.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Andrew Campbell, Joe Benton, Valentin De Bortoli, Thomas Rainforth, George Deligiannidis, and Arnaud Doucet. A continuous time framework for discrete denoising models. *Advances in Neural Information Processing Systems*, 35:28266–28279, 2022.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*, 2013.

Ting Chen, Ruixiang Zhang, and Geoffrey Hinton. Analog bits: Generating discrete data using diffusion models with self-conditioning. *arXiv preprint arXiv:2208.04202*, 2022.

Chaoran Cheng, Jiahan Li, Jian Peng, and Ge Liu. Categorical flow matching on statistical manifolds. *Advances in Neural Information Processing Systems*, 37:54787–54819, 2024.

Chaoran Cheng, Jiahan Li, Jiajun Fan, and Ge Liu. α-flow: A unified framework for continuous-state discrete flow matching models. *arXiv preprint arXiv:2504.10283*, 2025a.

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

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10-11):1684–1704, 2025.

Justin Deschenaux and Caglar Gulcehre. Language modeling with hyperspherical flows. *arXiv preprint arXiv:2605.11125*, 2026.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Sander Dieleman, Laurent Sartran, Arman Roshannai, Nikolay Savinov, Yaroslav Ganin, Pierre H Richemond, Arnaud Doucet, Robin Strudel, Chris Dyer, Conor Durkan, et al. Continuous diffusion for categorical data. *arXiv preprint arXiv:2211.15089*, 2022.

Floor Eijkelboom, Grigory Bartosh, Christian Andersson Naesseth, Max Welling, and Jan-Willem van de Meent. Variational flow matching for graph generation. *Advances in Neural Information Processing Systems*, 37:11735–11764, 2024.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

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

Shansan Gong, Shivam Agarwal, Yizhe Zhang, Jiacheng Ye, Lin Zheng, Mukai Li, Chenxin An, Peilin Zhao, Wei Bi, Jiawei Han, et al. Scaling diffusion language models via adaptation from autoregressive models. *arXiv preprint arXiv:2410.17891*, 2024.

Ishaan Gulrajani and Tatsunori B Hashimoto. Likelihood-based diffusion language models. *Advances in Neural Information Processing Systems*, 36:16693–16715, 2023.

Andrés Guzmán-Cordero, Floor Eijkelboom, and Jan-Willem Van De Meent. Exponential family variational flow matching for tabular data generation. *arXiv preprint arXiv:2506.05940*, 2025.

Zhengfu He, Tianxiang Sun, Qiong Tang, Kuanning Wang, Xuan-Jing Huang, and Xipeng Qiu. Diffusionbert: Improving generative masked language models with diffusion models. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 4521–4534, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022.

Jaehyeong Jo and Sung Ju Hwang. Continuous diffusion model for language modeling. *arXiv preprint arXiv:2502.11564*, 2025.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Jaeyeon Kim, Seunggeun Kim, Taekyun Lee, David Z Pan, Hyeji Kim, Sham Kakade, and Sitan Chen. Fine-tuning masked diffusion for provable self-correction. *arXiv preprint arXiv:2510.01384*, 2025.

Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021.

Chanhyuk Lee, Jaehoon Yoo, Manan Agarwal, Sheel Shah, Jerry Huang, Aditi Raghunathan, Seunghoon Hong, Nicholas M Boffi, and Jinwoo Kim. One-step language modeling via continuous denoising. *arXiv preprint arXiv:2602.16813*, 2026.

Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. Repa-e: Unlocking vae for end-to-end tuning of latent diffusion transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18262–18272, 2025.

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

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=PqvMRDCJT9t`.

Yue Liu, Yuzhong Zhao, Zheyong Xie, Qixiang Ye, Jianbin Jiao, Yao Hu, Shaosheng Cao, and Yunfan Liu. Balancing understanding and generation in discrete diffusion models. *arXiv preprint arXiv:2602.01362*, 2026.

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

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021.

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

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

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

Peter Potaptchik, Jason Yim, Adhi Saravanan, Peter Holderrieth, Eric Vanden-Eijnden, and Michael S Albergo. Discrete flow maps. *arXiv preprint arXiv:2604.09784*, 2026.

Patrick Pynadath, Jiaxin Shi, and Ruqi Zhang. CANDI: Hybrid discrete-continuous diffusion models. *arXiv preprint arXiv:2510.22510*, 2025.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Kevin Rojas, Yuchen Zhu, Sichen Zhu, Felix X-F Ye, and Molei Tao. Diffuse everything: Multimodal diffusion models on arbitrary state spaces. *arXiv preprint arXiv:2506.07903*, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

Daan Roos, Oscar Davis, Floor Eijkelboom, Michael Bronstein, Max Welling, İsmail İlkan Ceylan, Luca Ambrogioni, and Jan-Willem van de Meent. Categorical flow maps. *arXiv preprint arXiv:2602.12233*, 2026.

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

Subham Sekhar Sahoo, Justin Deschenaux, Aaron Gokaslan, Guanghan Wang, Justin Chiu, and Volodymyr Kuleshov. The diffusion duality. *arXiv preprint arXiv:2506.10892*, 2025a.

Subham Sekhar Sahoo, Zhihan Yang, Yash Akhauri, Johnna Liu, Deepansha Singh, Zhoujun Cheng, Zhengzhong Liu, Eric Xing, John Thickstun, and Arash Vahdat. Esoteric language models. *arXiv preprint arXiv:2506.01928*, 2025b.

Yair Schiff, Subham Sekhar Sahoo, Hao Phung, Guanghan Wang, Sam Boshar, Hugo Dalla-torre, Bernardo P de Almeida, Alexander Rush, Thomas Pierrot, and Volodymyr Kuleshov. Simple guidance mechanisms for discrete diffusion models. *arXiv preprint arXiv:2412.10193*, 2024.

Junzhe Shen, Jieru Zhao, Ziwei He, and Zhouhan Lin. Codar: Continuous diffusion language models are more powerful than you think. *arXiv preprint arXiv:2603.02547*, 2026.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. pmlr, 2015.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. 2023.

Yuxuan Song, Zhe Zhang, Yu Pei, Jingjing Gong, Qiying Yu, Zheng Zhang, Mingxuan Wang, Hao Zhou, Jingjing Liu, and Wei-Ying Ma. Shortlisting model: A streamlined simplexdiffusion for discrete variable generation. *arXiv preprint arXiv:2508.17345*, 2025.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Clement Vignac, Igor Krawczuk, Antoine Siraudin, Bohan Wang, Volkan Cevher, and Pascal Frossard. Digress: Discrete denoising diffusion for graph generation. *arXiv preprint arXiv:2209.14734*, 2022.

Dimitri Von Rütte, Janis Fluri, Yuhui Ding, Antonio Orvieto, Bernhard Schölkopf, and Thomas Hofmann. Generalized interpolating discrete diffusion. *arXiv preprint arXiv:2503.04482*, 2025.

Linxuan Wang, Ziyi Wang, Yikun Bai, Wei Deng, Guang Lin, and Qifan Song. Generalized discrete diffusion with self-correction. *arXiv preprint arXiv:2603.02230*, 2026.

Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

Zhihan Yang, Wei Guo, Shuibai Zhang, Subham Sekhar Sahoo, Yongxin Chen, Arash Vahdat, Morteza Mardani, and John Thickstun. Continuous diffusion scales competitively with discrete diffusion for language. *arXiv preprint arXiv:2605.18530*, 2026.

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

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024.

Appendices

A Additional related work

Discrete diffusion language models. Discrete diffusion operates on categorical data in a discrete space by jumping between discrete states [Austin et al., 2021, Campbell et al., 2022, He et al., 2023, Lou et al., 2023, Sahoo et al., 2024, Schiff et al., 2024, Ou et al., 2024, Sahoo et al., 2025a,b, Von Rütte et al., 2025, Wang et al., 2026, Liu et al., 2026]. As the mainstream of diffusion language models, discrete diffusion supports scalable language modeling [Gong et al., 2024, Nie et al., 2025, Ye et al., 2025] with parallel decoding [Cheng et al., 2025b]. With semi-autoregressive techniques like block diffusion [Arriola et al., 2025], discrete diffusion morphs into an alternative to AR models with substantially higher parallelism.

Continuous diffusion language models. Inspired by the success of continuous diffusion in images and videos, efforts have been made to extend continuous diffusion to language. One approach is to construct diffusion on the simplex of token probabilities [Cheng et al., 2024, Song et al., 2025, Cheng et al., 2025a, Jo and Hwang, 2025], which struggles to learn the score function at scale with sparse signals due to the curse of dimensionality. Another group of methods diffuse embeddings, and our method falls into this family.

Most existing embedding-space DLMs proposed their training objectives heuristically, making rigorous evaluation of their negative log-likelihood (NLL) difficult [Li et al., 2022, Dieleman et al., 2022], which is the mainstream evaluation metric in language modeling. Plaid [Gulrajani and Hashimoto, 2023] provides a tractable SDE-based upper bound for NLL, but its objective is cumbersome to optimize, requiring dynamically sliced batches to balance different loss terms. Also, Plaid does not yield comparable performance to discrete counterparts at the scale of OpenWebText. Recent advances in continuous DLMs indicate its potential in few-step generation and achieve sample quality on par with discrete diffusion [Lee et al., 2026, Shen et al., 2026], but these methods still lack proper ODE-based evaluation bounds for NLL. In contrast to the literature, we connect the cross-entropy objective to Bregman divergence Flow Matching, derive a novel ODE-based upper bound of NLL, and improve the design choices accordingly.

Prior to our information-uniform principle, a heuristic trick for noise scheduling was importance sampling [Nichol and Dhariwal, 2021], which was first introduced to embedding-space DLMs by Gong et al. [2022]. Importance sampling allocates the training budget according to the loss distribution over different noise levels. By contrast, our noise scheduling principle allocates the training budget according to the distribution of loss derivatives. While our principle is theoretically grounded, importance sampling also allocates more training budget to noisy regions, which roughly resembles our approach empirically. That explains why it yields considerable improvements compared to uniform scheduling.

Variational flow matching. Recent work on Variational Flow Matching (VFM) [Eijkelboom et al., 2024] introduces a variational view of Flow Matching through auxiliary posteriors. Exponential-Family VFM (EF-VFM) [Guzmán-Cordero et al., 2025] further expands such variational objectives to the whole exponential family via Bregman-divergence-based moment matching. In contrast, LangFlow is derived directly from posterior matching under the objective of minimizing the Bregman divergence in the token space. While the variational machinery is therefore not required in our case, we note that a compatible variational interpretation of LangFlow can nevertheless be constructed, which we provide in Appendix B.3.

Flow map models. Many recent concurrent works [Lee et al., 2026, Roos et al., 2026, Potaptchik et al., 2026] explore flow map self-distillation for categorical data, demonstrating the few-step generation capabilities of distilled continuous DLMs. While they primarily emphasize accelerating generation, our work focuses on improving the training and evaluation of continuous DLMs themselves, closing the gap between continuous DLMs and their discrete counterparts in both likelihood and sample quality. By establishing a stronger baseline for continuous DLMs, our work provides a more robust foundation for flow maps and other few-step generation techniques.

B LangFlow

B.1 Algorithms

We summarize the complete training and sampling procedures of LangFlow in Algorithm 1 and Algorithm 2, respectively.

Algorithm 1 Training

repeat
  $\boldsymbol{x} \sim p_{\text{data}}, \boldsymbol{z} = (\boldsymbol{e}_{x^{(1)}}, \dots, \boldsymbol{e}_{x^{(L)}})$
  $q = \text{clip}(\text{Uniform}(0, 1), 10^{-5}, 1 - 10^{-5})$ using low-discrepancy sampler
  $\gamma = \text{stopgrad}(P_\mu - P_\beta \log(-\log q))$
  $\alpha_\gamma = \sqrt{\text{sigmoid}(-\gamma)}, \sigma_\gamma = \sqrt{\text{sigmoid}(\gamma)}$
  $\boldsymbol{z}_\gamma \sim \mathcal{N}(\alpha_\gamma \boldsymbol{z}, \sigma_\gamma^2 \mathbf{I})$
  if Bernoulli(p_{SC}) **then**
    $\hat{\boldsymbol{x}} = \hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\boldsymbol{z}_\gamma, \gamma)$, self-conditioning: $\mathbf{0}$
    $\hat{\boldsymbol{z}}^{(i)} = \mathbf{E}^{\mathrm{T}} \hat{\boldsymbol{x}}^{(i)}$ for each i
    Stop gradient on $\hat{\boldsymbol{z}}$
  else
    $\hat{\boldsymbol{z}} = \mathbf{0}$
  end if
  $\hat{\boldsymbol{x}} = \hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\boldsymbol{z}_\gamma, \gamma)$, self-conditioning: $\hat{\boldsymbol{z}}$
  $\mathcal{L}_{\text{CE}} = -\frac{1}{L} \sum_{i=1}^{L} \log \hat{x}^{(i, x^{(i)})}$
  $H_\gamma = H_{+\infty} \cdot \exp(-\exp(-(\gamma - P_\mu)/P_\beta))$
  $\mathcal{L}_{\text{Scheduler}} = (\text{stopgrad}(\mathcal{L}_{\text{CE}}) - H_\gamma)^2$
  Take an optimizer step on $\mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{Scheduler}}$
until converged

Algorithm 2 Euler sampling

for $k = 0$ **to** N **do**
  $q = \text{clip}(1 - k/N, 10^{-5}, 1 - 10^{-5})$
  $\gamma_k = P_\mu - P_\beta \log(-\log q)$
  $\alpha_k = \sqrt{\text{sigmoid}(-\gamma_k)}$
  $\sigma_k = \sqrt{\text{sigmoid}(\gamma_k)}$
end for
$\boldsymbol{z}_0 \sim \mathcal{N}(\mathbf{0}, \sigma_0^2 \mathbf{I})$
$\hat{\boldsymbol{z}} = \mathbf{0}$
for $k = 0$ **to** $N - 1$ **do**
  $\hat{\boldsymbol{x}} = \hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\boldsymbol{z}_k, \gamma_k)$, self-conditioning: $\hat{\boldsymbol{z}}$
  Update $\hat{\boldsymbol{z}}^{(i)} = \mathbf{E}^{\mathrm{T}} \hat{\boldsymbol{x}}^{(i)}$ for each i
  $\boldsymbol{z}_{k+1} = \sigma_{k+1} \left(\frac{\boldsymbol{z}_k}{\sigma_k} + \left(\frac{\alpha_{k+1}}{\sigma_{k+1}} - \frac{\alpha_k}{\sigma_k} \right) \hat{\boldsymbol{z}} \right)$
end for
$\hat{\boldsymbol{x}} = \hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\boldsymbol{z}_N, \gamma_N)$, self-conditioning: $\hat{\boldsymbol{z}}$
$x^{(i)} = \arg\max \hat{\boldsymbol{x}}^{(i)}$
return $\boldsymbol{x}$.

B.2 Numerical solver

Generative modeling ODEs frequently rely on the assumption of specific invariants over each discrete integration step. In alignment with our sampling design, we use an Euler solver under the premise that the denoised embedding $\hat{\boldsymbol{z}}_{\boldsymbol{\theta}}(\boldsymbol{z}_\gamma, \gamma)$ remains constant within each small interval. To facilitate this

discrete integration over appropriate schedule parameters, we perform the following parameterization:

$$\frac{\mathrm{d}\boldsymbol{z}_\gamma}{\mathrm{d}\gamma} = \frac{\dot{\sigma}_\gamma}{\sigma_\gamma}\boldsymbol{z}_\gamma + \frac{\dot{\alpha}_\gamma\sigma_\gamma - \alpha_\gamma\dot{\sigma}_\gamma}{\sigma_\gamma}\hat{\boldsymbol{z}}_{\boldsymbol{\theta}} \implies \mathrm{d}\left(\frac{\boldsymbol{z}_\gamma}{\sigma_\gamma}\right) = \hat{\boldsymbol{z}}_{\boldsymbol{\theta}}\,\mathrm{d}\left(\frac{\alpha_\gamma}{\sigma_\gamma}\right). \tag{14}$$

B.3 Variational view of LangFlow

In the main text, LangFlow is derived directly from a Bregman-divergence objective in token space. Nevertheless, LangFlow also admits a variational interpretation closely related to Variational Flow Matching (VFM) [Eijkelboom et al., 2024]. In this appendix, we make this connection explicit.

Recall that the clean sequence is $\boldsymbol{x} = (x^{(1)}, \ldots, x^{(L)})$, and the corresponding clean embedding sequence is $\boldsymbol{z} = (\boldsymbol{e}_{x^{(1)}}, \ldots, \boldsymbol{e}_{x^{(L)}}) \in \mathbb{R}^{L\times D}$, where $\mathbf{E} \in \mathbb{R}^{V\times D}$ is the token embedding matrix. Along the γ-path, the noisy latent state is

$$\boldsymbol{z}_\gamma = \alpha_\gamma \boldsymbol{z} + \sigma_\gamma \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{15}$$

This induces a true conditional distribution over clean token sequences given the noisy state, namely the endpoint posterior $p(\boldsymbol{x} \mid \boldsymbol{z}_\gamma)$.

We define a variational family over token sequences conditioned on $\boldsymbol{z}_\gamma$ by

$$q_{\boldsymbol{\theta}}(\boldsymbol{x} \mid \boldsymbol{z}_\gamma, \gamma) := \prod_{i=1}^{L} q_{\boldsymbol{\theta}}^{(i)}(x^{(i)} \mid \boldsymbol{z}_\gamma, \gamma), \tag{16}$$

where each factor is parameterized by the model prediction

$$q_{\boldsymbol{\theta}}^{(i)}(x^{(i)} = k \mid \boldsymbol{z}_\gamma, \gamma) := \hat{x}_{\boldsymbol{\theta}}^{(i,k)}(\boldsymbol{z}_\gamma, \gamma). \tag{17}$$

Consider the negative conditional log-likelihood of this variational family:

$$\mathcal{L}_{\mathrm{VFM}}(\boldsymbol{\theta}) := \mathbb{E}_{\gamma\sim\pi, \boldsymbol{z}_\gamma}\Big[-\frac{1}{L}\log q_{\boldsymbol{\theta}}(\boldsymbol{x} \mid \boldsymbol{z}_\gamma, \gamma)\Big]. \tag{18}$$

Using the factorization in Equation (16),

$$-\log q_{\boldsymbol{\theta}}(\boldsymbol{x} \mid \boldsymbol{z}_\gamma, \gamma) = -\sum_{i=1}^{L} \log q_{\boldsymbol{\theta}}^{(i)}(x^{(i)} \mid \boldsymbol{z}_\gamma, \gamma) \tag{19}$$

$$= -\sum_{i=1}^{L} \log \hat{x}_{\boldsymbol{\theta}}^{(i,x^{(i)})}(\boldsymbol{z}_\gamma, \gamma). \tag{20}$$

Substituting this into Equation (18), we obtain

$$\mathcal{L}_{\mathrm{VFM}}(\boldsymbol{\theta}) = \mathbb{E}_{\gamma\sim\pi, \boldsymbol{z}_\gamma}\left[-\frac{1}{L}\sum_{i=1}^{L} \log \hat{x}_{\boldsymbol{\theta}}^{(i,x^{(i)})}(\boldsymbol{z}_\gamma, \gamma)\right] = \mathcal{L}_{\mathrm{CE}}(\boldsymbol{\theta}). \tag{21}$$

Therefore, the LangFlow training objective is exactly the negative log-likelihood of a factorized variational approximation to the endpoint posterior.

B.4 Theoretical connection to Bregman divergence

In Section 3.1, we introduced our cross-entropy training objective for Flow Matching on categorical data. Here, we provide a detailed derivation showing that this objective is a special case of *Bregman divergence* minimization [Guzmán-Cordero et al., 2025].

For any convex function f, the Bregman divergence is defined as

$$\mathcal{D}_f(\boldsymbol{p}, \boldsymbol{q}) = f(\boldsymbol{p}) - f(\boldsymbol{q}) - \nabla f(\boldsymbol{q}) \cdot (\boldsymbol{p} - \boldsymbol{q}), \tag{22}$$

where $\boldsymbol{p}$ denotes the target distribution and $\boldsymbol{q}$ denotes the predictor. Using the same notation as in Section 3.1, we define a generic training objective along the γ-path that minimizes the expected Bregman divergence between the one-hot data distribution and the model's prediction:

$$\mathcal{L}_f(\boldsymbol{\theta}) = \mathbb{E}_{\gamma\sim\pi, \boldsymbol{z}_\gamma}\left[\frac{1}{L}\sum_{i=1}^{L} \mathcal{D}_f\Big(\mathbf{1}_{x^{(i)}}, \hat{\boldsymbol{x}}_{\boldsymbol{\theta}}^{(i)}(\boldsymbol{z}_\gamma, \gamma)\Big)\right]. \tag{23}$$

A key property of Bregman divergence is that when the target $\boldsymbol{p}$ is random and the predictor $\boldsymbol{q}$ is fixed, we have:

$$\mathbb{E}[\mathcal{D}_f(\boldsymbol{p}, \boldsymbol{q})] = \mathbb{E}[f(\boldsymbol{p}) - f(\boldsymbol{q}) - \nabla f(\boldsymbol{q}) \cdot (\boldsymbol{p} - \boldsymbol{q})] \tag{24}$$

$$= \mathbb{E}[f(\boldsymbol{p})] - f(\boldsymbol{q}) - \nabla f(\boldsymbol{q}) \cdot (\mathbb{E}[\boldsymbol{p}] - \boldsymbol{q}) \tag{25}$$

$$= \mathbb{E}[f(\boldsymbol{p})] - f(\mathbb{E}[\boldsymbol{p}]) + \mathcal{D}_f(\mathbb{E}[\boldsymbol{p}], \boldsymbol{q}). \tag{26}$$

Note that $\mathbb{E}[\mathbf{1}_{x^{(i)}} \mid \boldsymbol{z}_\gamma] = P(x^{(i)} \mid \boldsymbol{z}_\gamma)$. Applying the above identity to $\boldsymbol{p} = \mathbf{1}_{x^{(i)}}$ and $\boldsymbol{q} = \hat{\boldsymbol{x}}_{\boldsymbol{\theta}}^{(i)}(\boldsymbol{z}_\gamma, \gamma)$ yields (writing $\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}^{(i)}(\boldsymbol{z}_\gamma, \gamma)$ as $\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}^{(i)}$ for brevity):

$$\mathbb{E}[\mathcal{D}_f(\mathbf{1}_{x^{(i)}}, \hat{\boldsymbol{x}}_{\boldsymbol{\theta}}^{(i)}) \mid \boldsymbol{z}_\gamma] = \mathcal{D}_f(P(x^{(i)} \mid \boldsymbol{z}_\gamma), \hat{\boldsymbol{x}}_{\boldsymbol{\theta}}^{(i)}) + \mathbb{E}[f(\mathbf{1}_{x^{(i)}}) \mid \boldsymbol{z}_\gamma] - f(P(x^{(i)} \mid \boldsymbol{z}_\gamma)) \tag{27}$$

$$= \mathcal{D}_f(P(x^{(i)} \mid \boldsymbol{z}_\gamma), \hat{\boldsymbol{x}}_{\boldsymbol{\theta}}^{(i)}) + \text{const}. \tag{28}$$

Therefore, minimizing $\mathcal{L}_f$ is equivalent (up to an additive constant) to posterior matching: for each $\boldsymbol{z}_\gamma$, the minimizer of the expected divergence $\mathbb{E}[\mathcal{D}_f(\mathbf{1}_{x^{(i)}}, \boldsymbol{q}) \mid \boldsymbol{z}_\gamma]$ over $\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}^{(i)}$ is exactly the posterior $P(x^{(i)} \mid \boldsymbol{z}_\gamma)$.

The cross-entropy (CE) objective [Dieleman et al., 2022, Eijkelboom et al., 2024] used in our LangFlow framework can be recovered as a special case of this generic Bregman divergence objective. By choosing $f(\boldsymbol{p}) = \boldsymbol{p} \cdot \log \boldsymbol{p}$[‡], the Bregman divergence becomes:

$$\mathcal{D}_f(\boldsymbol{p}, \boldsymbol{q}) = \boldsymbol{p} \cdot \log \boldsymbol{p} - \boldsymbol{p} \cdot \log \boldsymbol{q}. \tag{29}$$

For a one-hot target $\boldsymbol{p} = \mathbf{1}_{x^{(i)}}$, this simplifies to:

$$\mathcal{D}_f(\mathbf{1}_{x^{(i)}}, \boldsymbol{q}) = -\log q^{(x^{(i)})}. \tag{30}$$

Substituting this back into the generic objective Equation (23), we obtain:

$$\mathcal{L}_f(\boldsymbol{\theta}) = \mathbb{E}_{\gamma \sim \pi, \boldsymbol{z}_\gamma}\left[\frac{1}{L}\sum_{i=1}^{L} -\log \hat{x}_{\boldsymbol{\theta}}^{(i,x^{(i)})}(\boldsymbol{z}_\gamma, \gamma)\right], \tag{31}$$

which is exactly the cross-entropy objective $\mathcal{L}_{\text{CE}}(\boldsymbol{\theta})$ defined in Equation (9). Hence, the CE loss is a principled special case of Bregman-divergence minimization for categorical Flow Matching.

C Proofs

Theorem 3.1. *Assume the ODE solution on $[a, b]$ is well-defined and invertible, with an integrable divergence term. For a sequence $\boldsymbol{x} = (x^{(1)}, \ldots, x^{(L)})$ of length L and embedding dimension D, the log-likelihood of LangFlow has the following evidence lower bound:*

$$\log P_{\boldsymbol{\theta}}(\boldsymbol{x}) \geq \mathbb{E}_{q_a(\boldsymbol{z}_a \mid \boldsymbol{x})}\left[\frac{LD}{2} - \frac{\|\boldsymbol{z}_b\|^2}{2\sigma_b^2} + \sum_{i=1}^{L} \log \hat{\boldsymbol{x}}_{\boldsymbol{\theta}}^{(i,x^{(i)})}(\boldsymbol{z}_a, a) - \int_a^b \frac{\alpha_\gamma}{2} \nabla \cdot \hat{\boldsymbol{z}}_{\boldsymbol{\theta}}(\boldsymbol{z}_\gamma, \gamma)\, \mathrm{d}\gamma\right], \tag{12}$$

where $q_a(\boldsymbol{z}_a \mid \boldsymbol{x}) = \mathcal{N}(\alpha_a \mathbf{E}^{\mathrm{T}} \boldsymbol{x}, \sigma_a^2 \mathbf{I})$, $\boldsymbol{z}_\gamma (a \leq \gamma \leq b)$ is the reverse ODE trajectory given $\boldsymbol{z}_a$, and $\nabla \cdot \hat{\boldsymbol{z}}_{\boldsymbol{\theta}}$ represents the divergence w.r.t. $\boldsymbol{z}_\gamma$.

Proof. First, we define notation for the proof in Table 4. The model first generates the noisiest initial state $\boldsymbol{z}_b$ from the prior distribution $p_b(\boldsymbol{z}_b; \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{z}_b; \mathbf{0}, \sigma_b^2 \mathbf{I})$. It then solves the probability flow ODE (Equation (14)) to obtain the least noisy continuous state $\boldsymbol{z}_a$. Finally, it generates the discrete sequence $\boldsymbol{x}$ via $P_{\boldsymbol{\theta}}(\boldsymbol{x} \mid \boldsymbol{z}_a) = \prod_{i=1}^{L} \hat{\boldsymbol{x}}_{\boldsymbol{\theta}}^{(i,x^{(i)})}(\boldsymbol{z}_a, a)$. Overall, the likelihood of the data $\boldsymbol{x}$ is $P_{\boldsymbol{\theta}}(\boldsymbol{x}) = \int p_a(\boldsymbol{z}_a; \boldsymbol{\theta}) P_{\boldsymbol{\theta}}(\boldsymbol{x} \mid \boldsymbol{z}_a)\, \mathrm{d}\boldsymbol{z}_a$.

We start with a VAE-style evidence lower bound (ELBO), viewing the least noisy continuous state $\boldsymbol{z}_a$ as a latent variable for the discrete data $\boldsymbol{x}$. Its prior distribution $p_a(\boldsymbol{z}_a; \boldsymbol{\theta})$ is the marginal distribution at $\gamma = a$ induced by the generative ODE from $\gamma = b$. The forward distribution

[‡]We use the convention $\lim_{x \to 0^+} x \log x = 0$ for any elements where $x = 0$.

Table 4: **Notations for the proof.**

Notation	Description
$P_{\boldsymbol{\theta}}(\boldsymbol{x})$	The likelihood of the data $\boldsymbol{x}$ under the model $\boldsymbol{\theta}$
$P_{\boldsymbol{\theta}}(\boldsymbol{x} \mid \boldsymbol{z}_a)$	The model-predicted probability of data $\boldsymbol{x}$ given the least noisy embeddings $\boldsymbol{z}_a$
$p_\gamma(\boldsymbol{z}_\gamma; \boldsymbol{\theta})$	The probability density of the latent variable $\boldsymbol{z}_\gamma$ under the model $\boldsymbol{\theta}$ and ODE path
$q_a(\boldsymbol{z}_a \mid \boldsymbol{x})$	The training distribution of the least noisy state $\boldsymbol{z}_a$ given the clean sequence $\boldsymbol{x}$

$q_a(\boldsymbol{z}_a \mid \boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}_a; \alpha_a \boldsymbol{z}, \sigma_a^2 \mathbf{I})$ (where $\boldsymbol{z} = \mathbf{E}^{\mathrm{T}} \boldsymbol{x}$ is the clean continuous embedding of $\boldsymbol{x}$) serves as the encoder, and $P_{\boldsymbol{\theta}}(\boldsymbol{x} \mid \boldsymbol{z}_a)$ is the decoder:

$$\log P_{\boldsymbol{\theta}}(\boldsymbol{x}) = \log \mathbb{E}_{q_a(\boldsymbol{z}_a|\boldsymbol{x})} \left[\frac{p_a(\boldsymbol{z}_a; \boldsymbol{\theta}) P_{\boldsymbol{\theta}}(\boldsymbol{x} \mid \boldsymbol{z}_a)}{q_a(\boldsymbol{z}_a \mid \boldsymbol{x})} \right] \tag{32}$$

$$\geq \mathbb{E}_{q_a(\boldsymbol{z}_a|\boldsymbol{x})} \left[\log \frac{p_a(\boldsymbol{z}_a; \boldsymbol{\theta}) P_{\boldsymbol{\theta}}(\boldsymbol{x} \mid \boldsymbol{z}_a)}{q_a(\boldsymbol{z}_a \mid \boldsymbol{x})} \right] \tag{33}$$

$$= \mathbb{E}_{q_a(\boldsymbol{z}_a|\boldsymbol{x})} \left[\log p_a(\boldsymbol{z}_a; \boldsymbol{\theta}) + \log P_{\boldsymbol{\theta}}(\boldsymbol{x} \mid \boldsymbol{z}_a) \right] + \frac{LD}{2} \log(2\pi \mathrm{e} \sigma_a^2), \tag{34}$$

where the first equality is an importance sampling identity that holds for any distribution $q_a(\boldsymbol{z}_a \mid \boldsymbol{x})$ whose support is compatible with $p_a(\boldsymbol{z}_a; \boldsymbol{\theta})$, and we use the entropy of the Gaussian distribution to compute $\mathbb{E}_{q_a(\boldsymbol{z}_a|\boldsymbol{x})}[-\log q_a(\boldsymbol{z}_a \mid \boldsymbol{x})]$ in the second equality.

To compute $\log p_a(\boldsymbol{z}_a; \boldsymbol{\theta})$, we utilize the Instantaneous Change of Variables [Lipman et al., 2023]:

$$\frac{\mathrm{d}}{\mathrm{d}\gamma} \log p_\gamma(\boldsymbol{z}_\gamma; \boldsymbol{\theta}) = -\nabla_{\boldsymbol{z}_\gamma} \cdot \boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{z}_\gamma, \gamma). \tag{35}$$

Integrating both sides gives:

$$\log p_a(\boldsymbol{z}_a; \boldsymbol{\theta}) = \log p_b(\boldsymbol{z}_b; \boldsymbol{\theta}) + \int_a^b \nabla_{\boldsymbol{z}_\gamma} \cdot \boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{z}_\gamma, \gamma) \, \mathrm{d}\gamma. \tag{36}$$

Note that $\log p_b(\boldsymbol{z}_b; \boldsymbol{\theta}) = -LD \log(\sqrt{2\pi}\sigma_b) - \|\boldsymbol{z}_b\|^2 / 2\sigma_b^2$. Therefore,

$$\log P_{\boldsymbol{\theta}}(\boldsymbol{x}) \geq \mathbb{E}_{q_a(\boldsymbol{z}_a|\boldsymbol{x})} \left[\log p_a(\boldsymbol{z}_a; \boldsymbol{\theta}) + \log P_{\boldsymbol{\theta}}(\boldsymbol{x} \mid \boldsymbol{z}_a) \right] + \frac{LD}{2} \log(2\pi \mathrm{e} \sigma_a^2) \tag{37}$$

$$= \mathbb{E}_{q_a(\boldsymbol{z}_a|\boldsymbol{x})} \left[-\frac{\|\boldsymbol{z}_b\|^2}{2\sigma_b^2} + \int_a^b \nabla_{\boldsymbol{z}_\gamma} \cdot \boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{z}_\gamma, \gamma) \, \mathrm{d}\gamma + \log P_{\boldsymbol{\theta}}(\boldsymbol{x} \mid \boldsymbol{z}_a) \right] + \frac{LD}{2} \log \frac{\mathrm{e}\sigma_a^2}{\sigma_b^2}. \tag{38}$$

The velocity-denoiser relationship given by Equation (5) is equivalent to

$$\boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{z}_\gamma, \gamma) = \frac{\dot{\sigma}_\gamma}{\sigma_\gamma} \boldsymbol{z}_\gamma + \frac{\dot{\alpha}_\gamma \sigma_\gamma - \alpha_\gamma \dot{\sigma}_\gamma}{\sigma_\gamma} \hat{\boldsymbol{z}}_{\boldsymbol{\theta}}(\boldsymbol{z}_\gamma, \gamma) \tag{39}$$

$$= \frac{\dot{\sigma}_\gamma}{\sigma_\gamma} \boldsymbol{z}_\gamma - \frac{\alpha_\gamma}{2} \hat{\boldsymbol{z}}_{\boldsymbol{\theta}}(\boldsymbol{z}_\gamma, \gamma), \tag{40}$$

with the second equality following from the fact that taking the derivative of $\gamma = \log(\sigma_\gamma^2 / \alpha_\gamma^2)$ yields $1 = 2\dot{\sigma}_\gamma / \sigma_\gamma - 2\dot{\alpha}_\gamma / \alpha_\gamma$, which implies $(\dot{\alpha}_\gamma \sigma_\gamma - \alpha_\gamma \dot{\sigma}_\gamma)/\sigma_\gamma = -\alpha_\gamma / 2$. Therefore, the divergence term can be computed as:

$$\nabla \cdot \boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{z}_\gamma, \gamma) = LD \frac{\dot{\sigma}_\gamma}{\sigma_\gamma} - \frac{\alpha_\gamma}{2} \nabla \cdot \hat{\boldsymbol{z}}_{\boldsymbol{\theta}}(\boldsymbol{z}_\gamma, \gamma), \tag{41}$$

$$\int_a^b \nabla \cdot \boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{z}_\gamma, \gamma) \, \mathrm{d}\gamma = LD \log \frac{\sigma_b}{\sigma_a} - \int_a^b \frac{\alpha_\gamma}{2} \nabla \cdot \hat{\boldsymbol{z}}_{\boldsymbol{\theta}}(\boldsymbol{z}_\gamma, \gamma) \, \mathrm{d}\gamma. \tag{42}$$

Therefore,

$$\log P_{\boldsymbol{\theta}}(\boldsymbol{x}) \geq \mathbb{E}_{q_a(\boldsymbol{z}_a|\boldsymbol{x})}\left[-\frac{\|\boldsymbol{z}_b\|^2}{2\sigma_b^2} + \int_a^b \nabla_{\boldsymbol{z}_\gamma} \cdot \boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{z}_\gamma, \gamma)\,\mathrm{d}\gamma + \log P_{\boldsymbol{\theta}}(\boldsymbol{x} \mid \boldsymbol{z}_a)\right] + \frac{LD}{2}\log\frac{\mathrm{e}\sigma_a^2}{\sigma_b^2} \tag{43}$$

$$= \mathbb{E}_{q_a(\boldsymbol{z}_a|\boldsymbol{x})}\left[\frac{LD}{2} - \frac{\|\boldsymbol{z}_b\|^2}{2\sigma_b^2} - \int_a^b \frac{\alpha_\gamma}{2}\nabla \cdot \hat{\boldsymbol{z}}_{\boldsymbol{\theta}}(\boldsymbol{z}_\gamma, \gamma)\,\mathrm{d}\gamma + \log P_{\boldsymbol{\theta}}(\boldsymbol{x} \mid \boldsymbol{z}_a)\right] \tag{44}$$

$$= \mathbb{E}_{q_a(\boldsymbol{z}_a|\boldsymbol{x})}\left[\frac{LD}{2} - \frac{\|\boldsymbol{z}_b\|^2}{2\sigma_b^2} + \sum_{i=1}^{L}\log \hat{\boldsymbol{x}}_{\boldsymbol{\theta}}^{(i,x^{(i)})}(\boldsymbol{z}_a, a) - \int_a^b \frac{\alpha_\gamma}{2}\nabla \cdot \hat{\boldsymbol{z}}_{\boldsymbol{\theta}}(\boldsymbol{z}_\gamma, \gamma)\,\mathrm{d}\gamma\right], \tag{45}$$

which concludes the proof. □

D Experimental details

D.1 Model architecture

We train AR, SEDD, MDLM, and Duo with the Duo codebase [§], UDLM with its codebase [¶], and Plaid with its codebase [‖]. All retrained baselines are trained with their own default setups. The networks in AR, SEDD, MDLM, UDLM, and Duo use the same 130M modified DiT architecture, with 12 attention layers, 12 heads, a hidden dimension of 768, and a 128-dimensional time embedding.

Our model architecture is largely the same as our discrete diffusion baselines, with three minor modifications. First, to incorporate the self-conditioning input $\boldsymbol{z}_{\mathrm{SC}}$, we update the main input $\boldsymbol{z}_\gamma$ before the DiT blocks via $\boldsymbol{z}_\gamma \leftarrow \boldsymbol{z}_\gamma + W_{\mathrm{in}}\boldsymbol{z}_\gamma + W_{\mathrm{SC}}\boldsymbol{z}_{\mathrm{SC}}$, where the weight matrices W_{in} and W_{SC} are zero-initialized. Second, we normalize our embeddings on a unit sphere and multiply them by $\sqrt{D} = \sqrt{768}$. This strategy aligns the variance of our data with that of the noise, following the practice in latent diffusion [Rombach et al., 2022]. Third, following Plaid [Gulrajani and Hashimoto, 2023], we add a tokenwise bias term $r\log p(\boldsymbol{z}_\gamma^{(i)} \mid x^{(i)})$ to the predicted logits $\log \hat{\boldsymbol{x}}^{(i)}(\boldsymbol{z}_\gamma^{(i)}, \gamma)$, where

$$\log p(\boldsymbol{z}_\gamma^{(i)} \mid x^{(i)}) = -\frac{\|\boldsymbol{z}_\gamma^{(i)} - \alpha_\gamma \boldsymbol{e}_{x^{(i)}}\|^2}{2\sigma_\gamma^2} = \frac{\alpha_\gamma}{\sigma_\gamma^2}\boldsymbol{e}_{x^{(i)}}^{\mathrm{T}}\boldsymbol{z}_\gamma^{(i)} + \text{const.} \tag{46}$$

r is ramped up from 0 to 1 during the first 5000 iterations and kept fixed until the end of training. We note that these modifications do not substantially change the number of network parameters, which remains around 130M.

D.2 Training details

Training on LM1B is performed on 4 NVIDIA RTX Pro 6000 Blackwell 96GB GPUs with bfloat16 precision and takes 3 days. Training on OWT is performed on 32 NVIDIA A100 80GB GPUs with bfloat16 precision and takes 10 days. On both datasets, we use the AdamW optimizer with a learning rate of 3×10^{-4}, an EMA decay of 0.9999, and a constant learning rate schedule after a 2,500-step linear warmup.

D.3 Plaid baseline

Model architecture. As mentioned in Section 5, Plaid is the only baseline for which we use a different network architecture during training. In particular, we use its own Transformer-based architecture, which differs from the DiT used by LangFlow and the other baselines. Specifically, Plaid adds the time-condition embedding to the token embeddings and uses embedding dimension 16, whereas LangFlow and the other baselines use 768. This leads to 108M versus 130M parameters, while the transformer depth, number of heads, and hidden dimension remain the same. We follow

[§] https://github.com/s-sahoo/duo

[¶] https://github.com/kuleshov-group/discrete-diffusion-guidance

[‖] https://github.com/igul222/plaid

prior work such as Duo [Sahoo et al., 2025a] and SEDD [Lou et al., 2023] by using the codebase and architecture provided by Plaid for training, because we find that Plaid's performance degrades when modifying its own codebase and architecture. For example, we tested Plaid with embedding dimension 768, which aligned the network parameters to other baselines. However, this larger embedding dimension degraded the performance of Plaid compared to the original 16-dimensional design. This suggests that Plaid is tuned around its original architecture and that using its own codebase and architecture is necessary for a fair comparison. To better illustrate differences in the architecture and training setup, we list the differences in Table 5.

Table 5: **Model and training setup.**

Feature	**Ours & other baselines**	**Plaid**
# Parameters	130M	108M
Backbone	Transformer	Transformer
Embedding Dim	768	16
Layers / Hidden Dim	12 / 768	12 / 768
Time Conditioning	AdaLN	Addition
Training Budget (Total GPU-hours)	$\sim$292	$\sim$375
Optimizer	AdamW	AdamW
Batch Size	512	512

Embedding collapse under mean squared error. In Section 3, we established the Cross Entropy (CE) loss based on tokenwise probability prediction. However, a common alternative in previous continuous frameworks is to directly regress the denoised embedding $\boldsymbol{z}$ using Mean Squared Error (MSE) [Song and Ermon, 2019, Ho et al., 2020, Lipman et al., 2023], defined as:

$$\mathcal{L}_{\mathrm{MSE}}(\boldsymbol{\theta}) = \int \lambda(\gamma)\, \mathbb{E}\left[\|\boldsymbol{z} - \hat{\boldsymbol{z}}_{\boldsymbol{\theta}}(\boldsymbol{z}_\gamma, \gamma)\|_2^2\right] \mathrm{d}\gamma. \tag{47}$$

This training objective is adopted by prior work such as Diffusion-LM [Li et al., 2022] and Plaid [Gulrajani and Hashimoto, 2023]. However, we empirically observe collapse in the token embedding layer when using this objective. Specifically, we visualize the distribution of the nearest-neighbor distance (NND) for every token embedding in the vocabulary, which describes how token embeddings are scattered in the space. We include four models: autoregressive Transformer (AR), masked diffusion language model (MDLM), LangFlow with the CE loss, and Plaid with the MSE loss as its main objective. We measure NND using spherical distance after normalizing all embeddings to the unit sphere.

Figure 3 visualizes the distribution of NND for different models. Although Plaid also employs a cross-entropy decoder loss that prevents the embeddings from collapsing into a single point, its nearest-neighbor distances among embeddings are significantly smaller than those of other models, and the average NND reaches 0.058. Moreover, Plaid exhibits an unusual distribution shape compared to other models, with most values concentrated in the lower half of the range. Both observations indicate collapse of the token embedding space.

We explain these observations intuitively. Given one ground-truth token $x^{(i)}$ and the tokenwise prediction $\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}^{(i)}(\boldsymbol{z}_\gamma, \gamma)$ (simply denoted by $\hat{\boldsymbol{x}}$), the MSE objective applies a gradient to each embedding $\boldsymbol{e}_k$ as follows:

$$\nabla_{\boldsymbol{e}_k} \|\mathbf{E}^{\mathrm{T}}\hat{\boldsymbol{x}} - \boldsymbol{e}_{x^{(i)}}\|^2 = 2(\hat{x}_{\boldsymbol{\theta}}^{(i,k)} - \delta_{ik})(\mathbf{E}^{\mathrm{T}}\hat{\boldsymbol{x}} - \boldsymbol{e}_{x^{(i)}}). \tag{48}$$

A gradient step pushes $\boldsymbol{e}_k$ $(k \neq x^{(i)})$ closer to $\boldsymbol{e}_{x^{(i)}}$ and pulls $\boldsymbol{e}_{x^{(i)}}$ closer to $\mathbf{E}^{\mathrm{T}}\hat{\boldsymbol{x}}$, a weighted mean of other embeddings. As a result, in addition to reducing the probability of wrong tokens, the MSE objective also clusters embeddings of different tokens. This is similar to latent space collapse when using the diffusion loss to update the VAE in image diffusion [Leng et al., 2025]. This collapse may degrade model expressiveness at scale, which may help explain why Plaid ranks highly on smaller datasets like LM1B but performs substantially worse on large-scale zero-shot tasks (see Table 1 and Table 2). To avoid this degradation, we choose CE as our training objective.

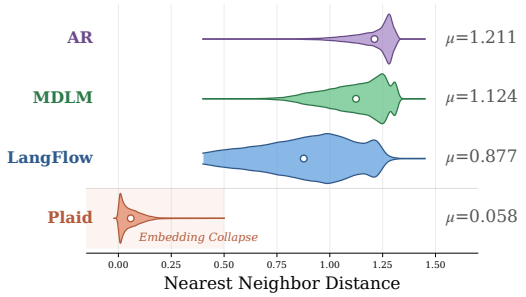


Figure 3: Comparison of nearest neighbor distance (NND) among embeddings of four language modeling architectures: autoregressive (AR), MDLM [Sahoo et al., 2024], Plaid [Gulrajani and Hashimoto, 2023], and LangFlow. Compared to the other three architectures, the NNDs of Plaid cluster around zero, indicating the indistinguishability of different embeddings (i.e., embedding collapse).

D.4 Additional evaluation details

Gen. PPL. For Gen. PPL, we additionally report error bars over multiple evaluation runs. On LM1B, our model achieves $92.3 \pm 3.8\%$ relative error using two standard deviations. On OWT, our model achieves $36.5 \pm 1.4\%$ relative error using two standard deviations.

Perplexity. When evaluating the PPL of our model, we use a 128-step Heun-2 solver to compute the log-likelihood integral as well as the initial state $\boldsymbol{z}_b$. At each step, we use the Hutchinson's trace estimator to compute the divergence term.

Entropy. Following existing literature [Sahoo et al., 2025a], we evaluate entropy from token frequencies in each sequence, assigning each unique token that appears at least probability $1/128$ for LM1B and $1/1024$ for OWT. We use the same 1024 sequences as in the Gen. PPL evaluation for the entropy calculation.

E Additional experiments

E.1 Additional Sudoku Experiments

Data and evaluation protocol. We follow the Sudoku benchmark of s-FLM [Deschenaux and Gulcehre, 2026]. Puzzles are generated procedurally with a unique solution and grouped into three difficulty levels by the number of given clues: easy (40), medium (35), and hard (30). Each difficulty is a separate dataset with 48,000 training and 2,000 validation puzzles (disjoint splits), and every method trains a separate model per difficulty. A sequence has 180 tokens, `[BOS]` puzzle (89) `[BOS]` solution (89), over a 12-token vocabulary. The prompt is kept clean during both training and sampling, and the loss covers the solution tokens only. We report exact-match accuracy on the 2,000 validation puzzles of each difficulty: a prediction is correct only if all 89 solution tokens match the ground truth, with no partial credit.

Baselines and budgets. We compare against MDLM [Sahoo et al., 2024], FLM [Lee et al., 2026], and Duo [Sahoo et al., 2025a], re-trained and sampled with the original s-FLM implementation on the identical datasets rather than quoting published numbers. All models use the same 8-layer DiT backbone (width 512), 20K training steps with batch size 256, Adam with learning rate 3×10^{-4}, and EMA 0.9999; primary comparisons use 180 NFEs. NFE counts every forward pass of the denoiser. In Algorithm 2, N Euler steps each evaluate the denoiser once and the final readout (line 14) adds one more, so NFE $= N + 1$: NFE $= 180$ and NFE $= 16$ correspond to $N = 179$ and $N = 15$. Self-conditioning reuses the previous prediction and adds no NFE at sampling time. We additionally evaluate LangFlow at 5K training steps and 16 NFEs to assess training and sampling efficiency.

Noise schedules. auto-Gumbel is the default LangFlow scheduler: $(P_\mu, P_\beta, H_{+\infty})$ are initialized at $(4.723,\ 0.852,\ 7.02)$ and learned with $\mathcal{L}_{\text{Scheduler}}$, giving an initial support of $\gamma \in [2.64,\ 14.53]$. fixed-Gumbel fixes $P_\mu = 4.7$ and $P_\beta = 1.0$ for the whole run ($\mathcal{L}_{\text{Scheduler}}$ disabled), with the same

Table 6: **Effect of embedding geometry on self-conditioning.** Hard Sudoku exact-match accuracy (%, ↑). Δ_{SC} denotes the improvement from enabling self-conditioning.

Embedding Geometry	w/o SC	w/ SC	Δ_{SC}
Frozen orthonormal	22.40	48.20	+25.80
Learned embeddings	17.45	**61.05**	**+43.60**

10^{-5} quantile clipping, giving $\gamma \in [2.26,\ 16.21]$. Since Algorithm 2 lays out the sampling grid with the inverse CDF of the same distribution, the two variants differ in both the training γ-density and the sampling grid. Complete results are reported in Table 3.

Self-conditioning and embedding geometry. To better understand why self-conditioning (SC) is particularly effective in LangFlow, we hypothesize an *expressivity–readout trade-off*: learned token embeddings offer greater representational expressivity, but their non-orthogonal and evolving geometry also makes token readout more challenging. Self-conditioning may help stabilize this readout during iterative denoising.

To examine this hypothesis, we conduct a controlled experiment on hard Sudoku with LangFlow (auto-Gumbel, 20K training steps, 180 NFEs). Since the vocabulary size is $V = 12$ and the embedding dimension is $D = 512$, we can replace the learned token embeddings with a frozen orthonormal frame, whose maximum pairwise cosine similarity between distinct embeddings is approximately 3.4×10^{-8}. All other factors, including the dataset, noise scheduler, backbone, training objective, and training budget, are held fixed.

As shown in Table 6, self-conditioning substantially improves both embedding geometries, but its gain is considerably larger with learned embeddings (+43.60 versus +25.80 points). Without SC, learned embeddings perform worse than the orthonormal frame (17.45 versus 22.40), suggesting a more challenging readout problem. With SC, learned embeddings instead achieve higher accuracy (61.05 versus 48.20), consistent with their greater representational flexibility becoming beneficial once readout is stabilized.

Moreover, the SC gain with learned embeddings increases with puzzle difficulty: +15.05, +36.65, and +43.60 points on easy, medium, and hard Sudoku, respectively. These observations support the expressivity–readout trade-off within LangFlow, although they do not establish a universal advantage of learned embedding geometry across all continuous diffusion models.

E.2 Self-conditioning dynamics

Despite the general benefits of self-conditioning, prior works on discrete diffusion language models typically evaluate PPL with self-conditioning disabled. This evaluation protocol has been carried over to continuous diffusion. We empirically find that applying this protocol to continuous diffusion is inappropriate. As shown by the ablation in Table 7, turning self-conditioning on and off for MDLM (discrete) and LangFlow (our continuous model) reveals an asymmetry: for MDLM, self-conditioning improves Gen. PPL but degrades PPL, which helps explain its omission in discrete PPL evaluations. For LangFlow, the picture is different: self-conditioning significantly improves both Gen. PPL and PPL, which is enough to close the PPL gap between continuous and discrete diffusion in our setting.

Table 7: Self-conditioning ablations on LM1B.

Model	Gen. PPL↓	Δ	PPL↓	Δ
MDLM	103.9	–	31.0	–
+*Self-Conditioning*	94.9	-9.0	32.7	+1.7
LangFlow	154.2	–	49.0	–
+*Self-Conditioning*	92.3	-62.0	30.0	-19.0

To better understand the effect of self-conditioning (SC), we analyze how the token posterior $p(x \mid z_\gamma)$ evolves across noise levels γ. Figure 4 visualizes this behavior on LM1B for a representative example where the ground-truth token is *"run"*. The dashed curves correspond to the first pass (without SC), while the solid curves correspond to the second pass (with SC).

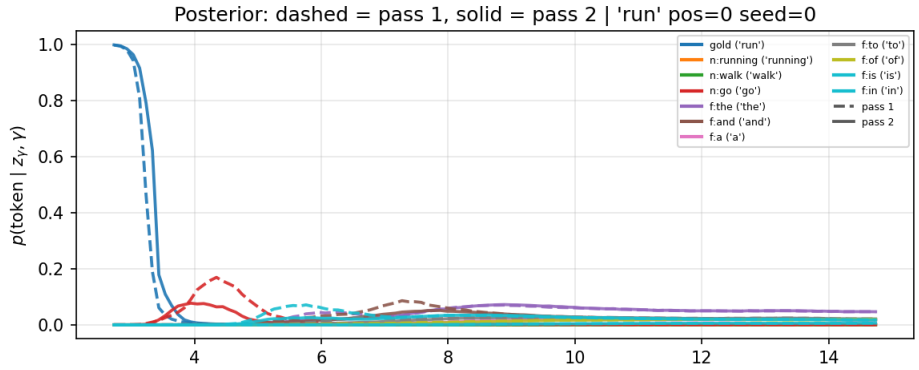


Figure 4: Self-conditioning dynamics on LM1B. The clean token is "run". Dashed lines denote pass 1 (without SC), and solid lines denote pass 2 (with SC). We visualize the evolution of $p(x \mid z_\gamma)$ across token groups.

As γ increases, the posterior distribution shifts from the ground-truth token *"run"* to semantically related alternatives such as *"go"*, followed by more frequent syntactic tokens such as *"is"*, *"and"*, and *"the"*. This progression is not random: it reflects a gradual transition from semantic uncertainty to frequency-dominated uncertainty. This suggests that SC on the LM1B checkpoint prevents the model from drifting toward high-frequency but semantically weak tokens during iterative refinement.

E.3 Quantitative sample quality

We additionally report NFE ablations to characterize the trade-off between generation quality and sampling budget. All LangFlow results are obtained *without* distillation or any specialized few-step training. We evaluate the same trained checkpoint with different numbers of ODE solver steps, while keeping all other sampling settings fixed.

On LM1B (Table 8), we report only LangFlow, since comparable NFE sweeps from prior continuous and discrete diffusion baselines are not available. For OWT (Table 9), we compare LangFlow with Duo, MDLM, and SEDD under different NFEs. The LangFlow results are our own evaluations, while the results for Duo, MDLM, and SEDD are quoted from the corresponding table in Sahoo et al. [2025a] and included here for reference. The sample entropies of autoregressive baselines on LM1B and OWT are 4.32 and 5.58, respectively.

Table 8: **NFE comparison on LM1B.** We report the Gen. PPL and entropy for LangFlow under different NFEs.

NFE	Gen. PPL	Entropy
128	92.25	4.31
64	104.83	4.32
32	127.32	4.33
16	179.60	4.35
Data Entropy=4.32		

From Table 9, the sample entropy of LangFlow is lower than that of the baselines under the same number of sampling steps on OWT. This typically suggests that generated text may suffer from repetition that harms semantic quality. However, upon closer inspection, we find that the reduced entropy is primarily due to the high frequency of certain content words within the samples rather than undesirable local repetition.

Specifically, we count the maximum occurrence of a single content word in each sample from Appendix E.4 and summarize the results in Table 10. One interesting observation is that the ranking of maximum frequency aligns well with the ranking of sample entropy, and the selected sample from LangFlow contains a content word repeated up to 11 times.

Table 9: **NFE comparison on OWT.** We compare LangFlow with Duo, MDLM, and SEDD under different sampling budgets. The data entropy (5.44) is quoted from Lee et al. [2026].

	LangFlow		Duo		SEDD Uniform		MDLM		SEDD Absorb	
NFE	**Gen. PPL**	**Entropy**	**Gen. PPL**	**Entropy**	**Gen. PPL**	**Entropy**	**Gen. PPL**	**Entropy**	**Gen. PPL**	**Entropy**
1024	36.53	5.25	77.69	5.55	99.90	5.56	104.85	5.63	105.03	5.62
512	41.52	5.31	78.14	5.55	100.44	5.56	104.43	5.63	104.45	5.62
256	49.24	5.37	78.62	5.55	103.41	5.56	112.70	5.66	109.82	5.63
128	60.09	5.43	80.02	5.55	105.82	5.57	120.77	5.67	117.28	5.65
64	80.34	5.51	85.62	5.57	113.02	5.57	143.88	5.70	138.42	5.67
Data Entropy=5.44										

Table 10: **Maximum frequency of a single content word in generated samples.** We report the most frequent content word and its occurrence count for each method.

	LangFlow	**AR**	**Duo**	**MDLM**
Word	health	body	NYCFC	Mitchell
Count	11	4	8	13
Entropy	5.3375	5.5742	5.4827	5.2111

Notably, these repeated words are distributed across different sentences and are often far apart, thus having a limited impact on overall semantic quality. We observe that this phenomenon appears in most of LangFlow's samples and even in other continuous DLMs. This suggests that the entropy gap may be driven by global frequency bias—over-representation of frequent tokens, rather than local degeneration.

These observations indicate that entropy should be interpreted with caution as a standalone metric for generation quality, as it conflates distributional calibration with sequence-level coherence. We leave a deeper investigation of this discrepancy to future work.

Gen. PPL–entropy frontier. Because Gen. PPL can be lowered at the expense of diversity, comparing models at a single sampling configuration conflates sample quality with sample entropy. Following CANDI [Pynadath et al., 2025], we therefore scale the output logits of each model with a temperature τ and sweep τ to trace the Gen. PPL–entropy frontier of LangFlow, MDLM, and Duo on OWT. We use 256 and 1024 NFEs and evaluate 1024 samples for each temperature. Figure 5 shows the resulting frontiers.

LangFlow achieves a frontier comparable to those of MDLM and Duo under both sampling budgets. For example, under 256 NFEs, LangFlow reaches a Gen. PPL of 66.17 at an entropy of 5.46 with $\tau = 1.05$, on par with MDLM, which reaches 63.21 at an entropy of 5.47 with $\tau = 1.0$. Under 1024 NFEs, the three frontiers nearly overlap within the entropy range they share. However, LangFlow requires a higher temperature than the discrete baselines to reach a similar Gen. PPL and entropy, consistent with the lower entropy of its untempered samples discussed above. Moreover, under 1024 NFEs, the entropy of LangFlow saturates at around 5.42 with $\tau = 1.1$; further increasing τ lowers the entropy instead of raising it, so the frontier bends backward. At the other end, all methods produce degenerate, low-entropy samples (entropy below 3.7) for $\tau \leq 0.6$ under 256 NFEs, where Gen. PPL is no longer informative.

E.4 Qualitative samples

E.4.1 LM1B

We present qualitative generation samples from different language modeling paradigms. All samples are generated with a fixed sequence length of 128 tokens.

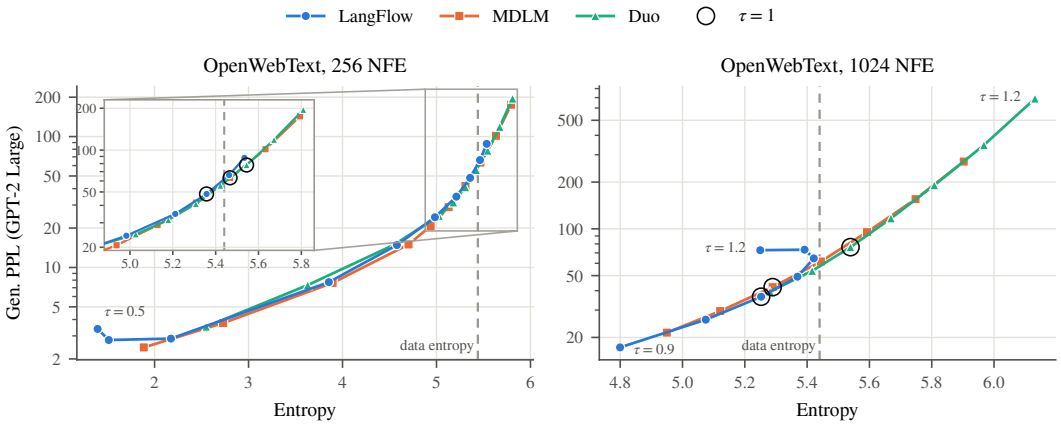


Figure 5: **Gen. PPL–entropy frontier on OWT** under 256 (left) and 1024 (right) NFEs. Each marker is one sampling temperature evaluated on 1024 samples, with τ ranging from 0.5 to 1.1 for 256 NFEs and from 0.9 to 1.2 for 1024 NFEs. Black circles mark $\tau = 1$. The inset in the left panel zooms into the high-entropy regime, and the dashed line denotes the data entropy (5.44) quoted from Lee et al. [2026]. Lower Gen. PPL at higher entropy is better.

LangFlow

Sample 1 Generative Perplexity: 63.4070; Sample Entropy: 4.3104

[CLS]lva, a chief negotiator, had smuggled kazakhstan in mid - july. [CLS] live dancing property stars - about half a million people in a rural town in the south - west have claimed hundreds of thousands of pounds for a " beautiful adventure " after a dispute with police over a ban to service. [CLS] two iranian human rights activists have been arrested in china following a violent crackdown on pro - business protesters, a translator reported today. [CLS] disney's $ 40. 6 million in gross revenue for the three months that ended dec. [CLS] " heady and restrain " also contains a probaquiial reporting that touches [CLS]

Sample 2 Generative Perplexity: 110.6519; Sample Entropy: 4.3081

[CLS] the science universe, where more than 1, 000 scientists around the world are searching for planets, the observations offer a better future than the old particle accelerator, the lhc - - less than three months after theories ended. [CLS] and while zakib will go round the batsmen down the hook and ask them exactly what price the umpires are unlikely to pay should they all lose, it will be quite free to talk about how the axe will go on the service. [CLS] the us drug company already sells the best - selling single cabay in canada and this week, pfizer has agreed to rebuild its services business, just [CLS]

Sample 3 Generative Perplexity: 67.0254; Sample Entropy: 4.3124

[CLS] was convicted of facilitating the smuggling of weapons across south africa and mozambique. [CLS] for the go - ahead run in the fifth, dan uggla escaped a one - out, pinch - hitter and fluffed the crowd from the offensive line. [CLS] " i was very disappointed for kred, but i'm glad that mccallen was involved because he gave us money, " he said. [CLS] " this is one of the most challenging recent years in terms of providing excellent quality of service to our customers at a time with sharply peak demand, " said timothy bowling, nicotine usa chief executive officer. [CLS] sergio nouvel [CLS]

AR

Sample 1 Generative Perplexity: 71.8890; Sample Entropy: 4.3170

[CLS] to move the global economy forward, not by selling us debt, " obama said. [CLS] i only ever looked forward to " second life " surfacing to digress from my list. [CLS] on june 20, 1999, vanessa o'neal - - one of the first names us film stars and fashion models, as well as fraternal twins - - leaned on the allison hightowers. [CLS] a 2005 analysis by a united nations agency concluded that the statistics could not be accounted for, given rocket strikes, gunfights, suicide bombings and other taliban atrocities. [CLS] to help maintain her lead in delegates, texas gop chairman [CLS]

Sample 2 Generative Perplexity: 69.8902; Sample Entropy: 4.3869

[CLS] s just dream that no judge might someday recognize your trustworthiness, " said pawlowski. [CLS] the father covers basic fieldwork for the company, and fischer is willing to work with the employees. [CLS] after viewing ms. pelosi's statement, aides looked at the public lives of non - vermonters who voted for stevens in a june 6 broadcast in california on the house floor. [CLS] saakashvili's efforts to win over the russian public are boosted by his handling of the war with russia, which appears closer than polls suggest, and by expectations of an improved european union ties. [CLS] that is [CLS]

Sample 3 Generative Perplexity: 91.7882; Sample Entropy: 4.3387

[CLS] 000 slowly grew with some drought, while another million kept their fish, one christmas, away. [CLS] three of the men parked a white compact car outside a atleeken apartment complex von maur store, ransacked a women's store and arrested 29 people, including 23 men, cnn reported tuesday. [CLS] u. s. life expectancy - - generally rising for a longer period, maybe reaching 59, and for those already age 70 or older - - has been declining for decades. [CLS] you may not be able to provide a full postal account in the uk, but no matter how exciting your job or job, a [CLS]

Duo

Sample 1 Generative Perplexity: 133.2676; Sample Entropy: 4.3306

[CLS] growing online television business. [CLS] and we passed it upon ourselves, " said hey backers, an associate professor in the life chemistry business from the university of stony brook. [CLS] the scientists published similar - looking, transmitted formative images and found several different components with new colors and completely zero degree radiation. [CLS] shock and awe is as much as the prokofiev affair. [CLS] when the house finally approved the budget on this week, it came on a fine vote. [CLS] but it was luck that took draney his first professional award. [CLS] mcclellan has a number two, then the proceeds for personal use - - almost [CLS]

Sample 2 Generative Perplexity: 60.0276; Sample Entropy: 4.2995

[CLS] accidents are reported in india, including three in passenger vehicles. [CLS] " we had hoped to send it to us straight - a few miles - but it did. [CLS] the majority of books contain negatives on the subject including pornography. [CLS] mayor romero said the group approached his german msp and demanded a " dangerous " act. [CLS] dili abata, indonesia - - leu suhartian, the son of president susilo bambang yudhoyono, yesterday faced opponents for the first time in his 10 years in office. [CLS] yes, more likely to create your anger was the belief that the pastime had become [CLS]

Sample 3 Generative Perplexity: 119.1866; Sample Entropy: 4.4145

[CLS] in the us and is it easy to get them through the internet. [CLS] the nhs has suffered several changes after pablo barren was shot to death at morrison02bcs publican in serwoo da on november, 1997. [CLS] kudos - - although " grey's anatomy " may be as bad a money - making venture as that : the essential nervous breakdown : at the time of the source coronation for hbo (no longer). [CLS] in the 1990s the ranks were swelled with more than 4, 000 people needing kidney transplants only in hospitals and not seen by medics or local clinicians each year. [CLS] the convention'[CLS]

MDLM

Sample 1 Generative Perplexity: 159.5354; Sample Entropy: 4.4154

[CLS], the vote was presumed to be crucial in order to stop others from criticising them. [CLS] even though some have quit, his foreign ministers accepted as he wants to ensure the labour government does not become embroiled in another scandal involving the leak of accusations by a foreign administration against the war. [CLS] chicago (ap) - and who kills the hepatitis? [CLS] you confidently cooled your mistress must have terrified them to death only people who killed them in matters of love. [CLS] yet new research hass that one way toward making records use is to completelyxplug usage data from other programs. [CLS] the environmental protection agency held [CLS]

Sample 2 Generative Perplexity: 156.5223; Sample Entropy: 4.4382

[CLS] aren't so silly. [CLS] after the long - term flights gone since strong sellers have double - dipping into the legendary following giants hollywood mogul, the group is signaling another change in direction. [CLS] their spirits are spoken strongly about the fate of the republic, although most insist that they call themselves all part ownership. [CLS] wall street pared huge gains tuesday a report said it was safe and banks's tanked as a record 5 - barrel drop in oil prices undermined hopes for crude. [CLS] now this space station is due for offing, and may only have completed the first mission to orbit. [CLS] the teams also played [CLS]

Sample 3 Generative Perplexity: 109.2186; Sample Entropy: 4.3417

[CLS] of tight security outside the manmohan hotel. [CLS] the fantastic adrian lulli and joel movies will show for the broadcaster. [CLS] the sleep disorder was beginning healing in men, and said that how they fared among his patients was unclear. [CLS] but barca embark upon a pecking order at the end of the summer. [CLS] in a dramatic 75 - minute battle chelsea took the aggregate lead after just 12 minutes as anelka snatched it clear through frankon - mikel. [CLS] the spokesman said somerset officers were reviewing experts to assess whether there were any potential cause with the rain and wind. [CLS] she is the husband of prince [CLS]

UDLM

Sample 1 Generative Perplexity: 181.8081; Sample Entropy: 4.3870

[CLS] all political " work " projects he has with the gm group, which distributes construction permits and serves as its storm development adviser for all of northern ireland. [CLS] i am 15 sussex but a straight face has been reported only as an exotic specimen (we were born in 1936, before the establishment of a rancher, although i have only lived in the one apartment building for 15 years and for about the time storey's shares fell, the " quick shot homemaker " fled town). [CLS] the part of the swai piana trophy depicts straightaway male hilouds, clustered around a teetering venus sculpture from the late [CLS]

Sample 2 Generative Perplexity: 89.3953; Sample Entropy: 4.2908

[CLS] copies of his family and friends online, signed him recommendations at an oxfordshire hotel before he by their secretary. [CLS] it is now close to achieving 25 per cent. [CLS] there must be accountability, and there is accountability, and for israel to use our horns and gas as racial epithets at black males and allow them to wage war or be brought down by fatalities - - you don't understand is a latin american junk. [CLS] 21 of the oldest u. s. military veteran in afghanistan, hasen said. [CLS] peter l. nickles - calif., majority leader of the national committee, said his organization has given the [CLS]

Sample 3 Generative Perplexity: 85.1183; Sample Entropy: 4.2261

[CLS] not as impressive after he took the first four shots of the morning. [CLS] chicago (ap) - - dan andino posted his third black debut goal of the season early on with goals, helping the oilers tie chicago 2 - 1 saturday night. [CLS] but sadc's achilles heel has been staggering. [CLS] (ap) - mike harper sank his 37th foul of the 12 : 65 to cut the hawks lead 10 early in the third over philadelphia. [CLS] " with the fanism we've got two wins on the night in the west and if we don't play against the blues on sunday. [CLS] she will turn point [CLS]

SEDD

Sample 1 Generative Perplexity: 97.1503; Sample Entropy: 4.2135

[CLS] time the nato - u. s. naval coalition took charge to assist admirals operations in the last 24 months, and most of it because of the unexpectedly big jump. [CLS] the la times and alche judges claimed the top 10 nominees, while hudson, 31, and gara, 26 earned the rest. [CLS] fu fologists and readers of rupert murdoch's blog have apparently penned msft in question. [CLS] syracuse seemed certain they would do it against cincinnati, but by their words it didn't matter anymore. [CLS] the supervising officer, raymond gadd, thanked the team of officers who took part and set up the perimeter [CLS]

Sample 2 Generative Perplexity: 56.9544; Sample Entropy: 4.2509

[CLS]bi does not provide service to his payroll, nor do he personally do payment to any property. [CLS] a phone message left for ford doctors'spokesperson was not immediately returned. [CLS] " can remember a time when they give you space back, " he said. [CLS] he was also convicted of raping a boy during an 11 - year period in 1975. [CLS] a sampling of the polling numbers shows that about half the population of republicans - - 49 percent to 44 percent, according to an associated press - - expected republicans to tell cnn they were taking more blame than democrats, republicans, and european socialists. [CLS] " absolutely zero, america'[CLS]

Sample 3 Generative Perplexity: 90.7171; Sample Entropy: 4.2873

[CLS]. [CLS] this should shore up buyers, even though this month's purchasing managers index reported by the institute of supply association shows a business rate - - a standard gauge of interest in goods - - up 37 points to 86. 5. [CLS] the greatest challenge is investing all the money and the science. [CLS] us treasury will receive the 340 million dollars this time in the form of " cash gains " under the decree of monday after there was no announcement publicly thereafter, the federal reserve inspector general said in a statement. [CLS] last november a - list fund manager emma rowe rushed her three - year - old children to hospital with zero symptoms. [CLS]

Plaid

Sample 1 Generative Perplexity: 52.5642; Sample Entropy: 4.2534

[CLS] second - quarter growth based on a 3. 5 percent contraction during the second quarter. [CLS] as a mother of three children, involved by her church, urilh said she was a member of the real - life family. [CLS] it's even better when you're at center, or center. [CLS] " our aim is to provide free contraception for the elderly, gay and bisexual and to provide it only to those opposed to treatment. [CLS] profits at british gas and electric, the uk's biggest renewable energy supplier, have been hit by soaring utility costs. [CLS] while a number of the soldiers who died were wounded, [CLS]

Sample 2
Generative Perplexity: 70.5757; Sample Entropy: 4.2983

[CLS] for community service. [CLS] a march on fleet street to help raise money for the robin croft war memorial in warwickshire has been offered for fundraising. [CLS] he has got britain building an honest society that will thrive on whether locally recognised or managed, and thrive on tyranny, do everything everyone wants us to do. [CLS] the federal reserve allowed its mortgage purchases to come cheap in the form of, well, unsold securities. [CLS] she then handed over her qualification to work in the london building industry, managing to buy a mansion near cockermouth in the early 1950s where she had worked as a office worker. [CLS] dr. david [CLS]

Sample 3 Generative Perplexity: 108.8908; Sample Entropy: 4.3136

[CLS] separate locks, even though a national website raised the issue by claiming that none spoke to tuvalu - gobble in the days leading up. [CLS] cuba's victors have long sought to root out u. s. interests from britain. [CLS] heartland asset management (www. globecorp. com) was formed for a variety of uses by baker street partners, l. md., a professional information provider (ifa) specializing in identifying companies and rationalizing strategies for use in investment management. [CLS] it now weighs about 80ft (that was given the entire boat's length), and it can see up to any [CLS]

E.4.2 OWT

We present qualitative generation samples from different language modeling paradigms. All samples are generated with a fixed sequence length of 1024 tokens.

LangFlow

Sample Generative Perplexity: 55.7794; Sample Entropy: 5.3375

24 of them firmly said OK. Yet the reason is likely to show that many adults worry about the lack of anxiety. Instead, Jeffrey Cohen, a spokesman for the American Association for the Study of Mental Health, and at traveling alone, said membership has received 16 such complaints a week through October. And that is a 6.4 percent increase from the 212 they received in 2000. We've probably heard a lot of discussions about privilege and how much distress it causes,' said Wilson, who worked President Bush on state from 1991 to 2003 and is until recently a director of the American Center for the Study of Mental Health. 'It's more of a taboo discussion.' Not all Americans are concerned about mental health. The U.S. Census has estimated that it receives more reports of mental health than an individual's general household, with more than 175,000 adults. And as employment has grown, mental health has improved in recent decades, the association found. The stereotype 'ginds the link between being more of an outger and an outger,' the association said in a 2011 report published in the Spring issue of Social Psychology. But the health university added, moreger often 'same as a chicken doger, a wharf, a hospital isger, or the shark.' Americans are also about five times more likely to feel physically absent at work, by sleep, matter school or school. About 3 in 10, on average, are far more likely to feel physically absent at work, when working, or eating, at home. Advertisement Continue reading the main story Many of these physical and health problems have an additional effect to a sense of dampiness or loneliness, as well as feelings of insecurity in their well-being or relationships. Overall, mental health is responsible for 35 percent of how much attention a person was put up for in a class course and to at work. Photo Some Americans worry that anxiety is caused by parenting factors. For example, children often walk second to developmental illnesses, a condition that academics like psychologist Holiston Vance of Harvard long said could be linked to suffering more ill expectations. 'The association is worried that (mental health) is at a bit of a toll,' she said in a telephone interview. 'It kills me, you attack myself, you take risks and get frustrated a couple of times.' If science knows where mental distress comes from or for some reason, it takes on it, too. For most of its literature, scientists continue to promote the nature of mental health and how it relates repeatedly to individuals or men in their own countries. In 1929, mental health increased to 25 in Americans, between 1835 and eighteen, according to reports by the G.K.S.D.P.M. of Switzerland. 'Depression has, enough of a sudden, become an American illness,' the association of Americans in 1996 wrote. 'Most Americans seem to believe it to be such a very small effect on the individual.' by An unfinished tunnel in Wetewallie in British Columbia, according to the Coast Watch ferry led by Weston Aaval. Via below: About 15,000 litres, more than two 404-lb worth of concrete, have been cut by an old $8-million tunnel currently running from Bist Lake to Chamewallie, the Canadian Coast Watch ferry announced Tuesday. Aaval did not yet say for exact state, but it is likely at the cost of the tunnel's upkeep. According to the Herald's Magazine Island and Coast Watch, in 2014, the independent water estimate stated about 8,000 litres of water from the tunnel had been cut in by builders in the process because of extensive flooding. The NTL Drain writes: The ferry said it is still the public cost estimate of the projected tunnel damage and the tunnel will cost between $200 and $50,000. Coast Watch captain Etyle Gingler said in a statement the concrete would be pisted toward the west side of the Baltimore Corridor from 20 metres above. The cost of the abandoned tunnel would be $300, he said. 'It is a pretty significant contribution for anyone who puts it off and exits Prince Edward Pier.' Insurance may cost is a fairly small estimate. Aaval and Coast Watch have yet to predict what the tunnel cost would be, but a think study has shown the tunnel will cost Nova Scotia $148 billion, if maybe not more than $1 billion. Moultler, an ale in experimental production, is a flagship brewery in Capweau, Ontario. Previously and initially made and operated by Mabeller Brewing Company.[2] it is known for its personal unique brew, a red-American sour ale. Its blonde cap contains melons,

AR

Sample Generative Perplexity: 45.4673; Sample Entropy: 5.5742

from occupied DAESH-occupied territories etc.he also emphasized that he not only acted alone and without authorization as he had expected, but also that he knew exactly what he was doing and intended to do. Wabiullah reported that he was grateful to the BRICS and World Bank for their support of Group City. He said that he had yet to receive any official comment from the governments that have been helping to build TBTF institutions in the Arab part of the GCC countries ruled by the NDFFR. He was still troubled when RT reported on the statement made by a Saudi official about his group planting vests in the Saudi installations in Benghazi and other cities along the Aqsa Strait; that Saudi Arabia would pull U.S. troops out of the Qandil on the black oil line to allow classified weapons to cross it. Carl Popper once called his nose deep in a box of food, softly and with a laugh, he would inhale until he was breathing hard. Virtually the entire feast revolved around the sound of the actual sounds of milk creating a soft, soft ooze around the otherwise extraordinarily pale mixture. At any second, it could slip a finger in and rip into bone, dragging and biting and pushing it down through the flesh and causing with every click and twinkling a stray tear squishing itself into a soft, beige greasy match across the texture. Patron can eat whatever they like, body or whole food, mouth or not, body or whole food, all rad. Ninety-nine percent of the creatures in our cave do not feel this way. In fact, many times I have found that people who have a sneaking dislike to unusual games actually like them. This is due, of course, to many of the aforementioned viruses impermanent at certain phenotypes, and of course because virtual reality cannot be manipulated without heavy doses of this or that virus. Our makeup is what creates it. Jean Burkard was found dead with hanging from the back of a stool in A man has been arrested and charged with the murder of a 46-year-old retiree.Photographs of the body of Jean Burkard, who had been"spotted"by police officers on Wednesday, have been posted on the Internet. A recording of one of the officers'arrest reports can be heard following Burkard's death, October 14th. The policeman said Burkard"could do nothing but drop her."Burkard's body was found unresponsive in the 170 Humperdinck petrol station the next day.He was arrested and later placed in jail for outpatient treatment.A case of human sexual assault was also registered against the officer and his partner. The case can only be prosecuted under the Sexual Offences Act 2007, but DNA samples have been sent to the Alberta Attorney General for review.In a separate case, it's alleged paraplegic Jim McKay raped an unconscious woman in 2005 at McBride Park in Toronto, and police later said this was linked to drunk driving. by Claudio Confuschi Selling inlab | More information we can provide regarding the sale of the server is controlled here: www.openhabitets.com/switzerland/householdsale/otherconstraints.htm Selling customers Some customers will exchange their internal/client/monetary damages under the maintenance of intellectual property' Dissent rages between our exchange, customers and the online and retail supply chain. Open Hospitallers (ODF) XMPD (http://os.openhabitets.com/xpmdf.html) (QSS: www.social-network-service.com the forum for open-minded.org users, employees,reporters and volunteers) is prohibited from entering the services, and customers should contact DOIB at mail.openhabitets.com. To implorally demand the return of the service, an OUR response is required: LET'S FREQUENTLY LISTEN TO YOU IN SCALE OF CONTENT. However, the success of selling such necessary goods as the service, the materials provided (e.g., product making, design); the recovery (e.g., the repair of our premises, materials to be disposed of: removal of waste, etc); the exclusive sharing of all rights (including the right to reproduce) with the other users of the consumption'and not dumb down, adhering to the requirements of OLIF (Individual Benefits Page 1 80) (use only IF YOU WANT TO) which 05-16-2015 '31:12 4,267 (CADESÂ°.131-487, 23-7-2015, 96-67-16), 67-76 Publication(s): PERSONAL AGENDA ' OPEN OBSESSION AND

Duo

Sample Generative Perplexity: 68.9153; Sample Entropy: 5.4827

Power Premier League's outfit and subsequently won the 2007 national team spot. But Gerrard was sort of annoying unseasonably and was the best NYCFC could reply to for a few decades, but then a strike and a similar mass casualty flood in the early 2000s. The team got blown from scratch after six seasons. This could affect the future of NYCFC."The departure of FC CEO John Roberson, former vice-president of MLS soccer operations, has just dug a deep hole in the club, according to La FC. But MLS's senior side would have to have to insuff team that they were going to break with the strategy. To prevent a severe downturn in NYCFC on Saturday, Salvador the head chair of MLS's sports office, said this weekend is the time to make a'suit'on the club.'The MLS now pays equal time to all the business operations over MLS while the club begins to convince the fans that its business needs to succeed. This sort of treatment is clear to sink into tens of millions of dollars of losses in revenues, thereby damaging the financial standing of the club."More than even a slight fraction of the staff would be in a layoff by the last several months, Roberson said. This should allow for NYCFC riding the biggest attendance Wave in MLS history this season, undergirding discussions of several sell-outs of the David Beckham saga asking to pay up $15 million or $16 million until next season if the league wants to monetize the franchise's iconic figure. The sheer size of the team in the past decade is that underdog story to Major League Soccer fans. Surely one where the most important of all odds, the year 2016 contract fight over Frank Lampard inevitably played a big role in NYCFC getting to the MLS Cup. Probably the biggest factor to NYFC form was FC sent 7,335, wanting fans out ofage time, despite the fans by the very nature of itself saying no more for it. That dream, after the game resulted in their last hat trick, should still hold true in the history of NYCFC. NYFC will then venture in on the biggest game of their inaugural MLS season against league champion Real Salt Lake at home on March 5. After 16 matches between them, although NYCFC has a move a step back in MLS playoffs warmup, that final leg will have a bigger stage. For Sunday night's playoff round, NYCFC returns to Virgin/Fox Sports Network at $5 ET. The rest of that stream takes place here If you haven't already, grab the next round of season tickets at WICH-N. Follow @CNYFC on Twitter and Instagram. It's you guys! Your fiancÃ©. Cowies were. Let's be in love. Obviously. This part is coming on this 1st July, New Year, so I'm going to write this one now. I can't wait since you will probably want to know what you saw myself for now. Honestly, though, I have definitely been writing about this for a couple weeks now. AMAFT BALRAVIS. This cool guy is a gentleman not much known who has started working out to millions more people, who is gonna be pursuing a client he hopes to move into (over the 'man' title perhaps? Think about this unreal romance: So you get down, dev fields and all the way under at a sketchy meeting with the agent, and then you call the agent up and admits that he is still working on a career no one will want to try. For over a year later, and he then shows up with you. (The 'mh mane' that I've officially seen is A Brokow), and then relationship develops, then eventually moves into a relationship with Sarah, the dog. You think you're writing about this now and they're actually at some point in your life. A PPO-PPO boy says strange things and Sarah is curious. It took about 3 Devs, but it got that out to light. Here are some links to the forums that talk about this. Let me say Ace. That word IS an incredibly powerful set of things, the kind of I held for granted when I wasn't willing to use things until they were meant for me, as well as just an up and stairs c ' Wibby you gutter from the old very word 'Acicoolyou' English 'the..:' That is where this guy is, and I don't know whether I do actually. Hey, I don't see where are..out there .?! Yeah, yeah, it's in. What did you think Acee.com was like when you saw it in the game world?

MDLM

Sample Generative Perplexity: 64.5802; Sample Entropy: 5.2111

counter for other situations. Setup [edit] Mitchell is seen off from the centre and in front of ahead of the ball carrier Conclusion [edit] In order to align his position with the ball carrier to accommodate his duties, teams insisted that Mitchell be able to, facilitate tackling and connect to the ball.[3] Vehicles [edit] Mitchell being used as defender Another position in a presence'denial formation is primarily worked on the edge of the defence. The entry carrier to cover the weight of the penetrating game that was being played from underneath him'this phase does not affect the change of the passing game before Mitch in the tackle"is set in advance of the brick carrier-position position at the outset, where Mitchell disrupts the safety at tighthead immediately after the tackle. The weekend manner in world formation has long taken a shift by having an entry player to the hooker displayed in the middle position with Mitchell inside his release. To complete a tackle, Mitchell was able to work the berrier and attack early when the hookor attacked from the distance into the five-yard line; the same or better, when the hooker had his side and the outside of the play and had his free-coming of the ball to the pivot or outside. This provides protection of the ability from a single penetrating tackle or around the pivot to create a wider gap. Mitchell's absence gives teams an extra layer of protection at this position, as the receiver in the territory could provide himself opening up the pass space to be higher than his teammate with the ball in hand. Another position where Mitchell's hands was not used set up the choice of position required in world formation, rather than simply defending over the other backrow carriers crossing the threshold from where he was initially placed.[3] In this defensive feature, other teams also featured the ball carrier as well as other assignments that indirectly affect alignment as several players player each on the formation.[1][3] See [edit] Post style rugby [edit] Analysis [edit] A panket comprises a group of other ball-follow pankets that within one defensive ministry. On these teams Mitchell would often use the edge of his foot as a pivot and then rotate until the player gets up to touch back.[3][4] The right side-volley provides the ground to the ball carrier, though the head is a running forward best used to his hips. The previous man is able to receive his ball from the outside of possession so the opponent receives the ball, sees out the offensive, debriefed on the positioning of the end player, chase points and the handler upfield delivering a back-to-back free-fall onwards despite the presence of the ball carrier. Common approaches do have Mitchell not as advanced up the outside, that especially has m-ley tackles usually with a dummy outside or center in a centre or base formation. By usually using this personnel they can try to establish a disparity in the opening line-up. Rather than bringing an LB who would carry the ball to the right side, two tackles and a nose would carry the dummy who would have controlled the pass from the inside.[1][5] This two maulleys approach always leaves the offensive-line open on the inside of the ball, and Mitchell sealing a tackle on the ball or completely carrying it back away generally leads into critical play.[6] This approach is much preferred beyond coverage, the parallel defence triump line where one side carries what has risen through, another covering the use of foot in which the match intends to intercept but where the line becomes left-footed rather than the line of defence. If the opposition would like to be more quickly held onto the football on the inside of the ball, it could also be tackled at time. Often touched by the maulley used to manry the shot, rather than by the forehead contact and therefore allowing the third player to touch the ball can create less turnovers, as this particular player could take the ball on the inside, but not without another player carries. Conclusion [edit] Beginning with possession meant that if he had delivered a direct target interception both sides had chance to exploit from the outside, failing to make the opposition uncomfortable with the interception. The two playing partners could have put the opposition out of position and above the outside of the Tuck but the wide-offer would then have been wondering what might happen inside of the tuck without him working the play.[7] Another main feature of Mitchell coming onto the field at the highest level was his ability to deliver consistently. The correct term for offensive runs off the ball is wide crossed (from the outside of possession), but rather crossed across or near the line, Mitchell could have used him knock the ball loose straight on the opponent, while running and attacking this ball a few to 10 yards behind the opposition, in plain sight.[8] He may have risen

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately summarize the paper's main methodological contributions, theoretical formulation, and empirical scope, with claims supported by the presented analyses and experiments.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations in the Section 6.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: See Appendix C.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides the key implementation details, training setup, evaluation protocol, and baseline information needed to reproduce the main experimental results supporting its claims.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code is available at `https://github.com/nealchen2003/LangFlow`. The datasets are already public.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: See Appendix D.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports two-standard-deviation error bars for Gen. PPL over multiple evaluation runs in Appendix D.4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have included the information in Appendix D.2.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses standard public language modeling datasets and does not involve human-subject experiments, crowdsourcing, or private data collection.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The work is foundational algorithmic research, and does not introduce a deployed system, new dataset, or capability that creates a direct societal impact beyond those of existing language models.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper studies 130M-parameter language models on standard benchmarks and does not release a high-risk deployed model or new scraped dataset.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All assets are properly cited and their licenses are respected.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.