# Membership Inference in Fine-tuned Diffusion Language Models via Token-level Memorization Asymmetry

**Anonymous ACL submission**

## Abstract

Diffusion language models (DLMs) have recently emerged as an alternative modeling paradigm to autoregressive LMs, offering advantages such as parallel generation and bidirectional context modeling. Despite growing interest in their generative capabilities, the privacy risks of DLMs remain underexplored. We identify a phenomenon termed *token-level memorization asymmetry* through theoretical analysis of diffusion training dynamics. Building on this finding, we propose *Q-Skewness*, a skewness-based metric for membership inference on finetuned diffusion language models. Experiments across multiple fine-tuning datasets and models show that our membership inference attack outperforms existing baselines. Moreover, we show that Q-Skewness improves the effectiveness of PII extraction attacks.

## 1 Introduction

Diffusion language models (DLMs) (Nie et al., 2025; Bie et al., 2025; Ye et al., 2025) are emerging as a compelling alternative to autoregressive (AR) generation. Recent advances demonstrate that diffusion can be successfully adapted from continuous domains to text, leading to practical systems such as Gemini Diffusion (Google DeepMind, 2025), SEED Diffusion (Song et al., 2025), and a growing body of work from both academia (Nie et al., 2025) and industry (Inception Labs, 2025). Unlike AR models, which strictly generate tokens sequentially, DLMs generate text through iterative denoising processes that refine entire sequences over multiple steps. This paradigm offers capabilities that AR modeling does not naturally support, including flexible decoding (De Bortoli et al., 2025), global sequence refinement (Nie et al., 2025), and improved controllability over generation dynamics (Li et al., 2022), making diffusion an attractive generation paradigm other than next-token generation.

Driven by ongoing research efforts, state-of-the-art DLMs gradually achieve generative capabilities comparable to those of AR models, both in open-source models (Bie et al., 2025) and in commercial SaaS services (Inception Labs, 2025). However, few studies investigate the privacy issues of DLMs as a new language model architecture. In this paper, we mainly focus on Membership Inference (MI), which aims to determine whether a given data point is used to train the target model. On one hand, a malicious adversary can leverage this technique to violate the model privacy (Shokri et al., 2017). On the other hand, membership inference can also be used for unauthorized data auditing (DeAlcala et al., 2024).

A series of studies have developed membership inference methods against common (autoregressive) language models (Yeom et al., 2018), or the (vision) diffusion models. However, applying these methods to diffusion language models poses significant challenges. First, DLMs adopt an any-order denoising training objective rather than a single fixed left-to-right decomposition, which acts as an implicit Monte Carlo data augmentation mechanism and reduces exact sample memorization (Ni et al., 2025). Second, the diffusion training process introduces randomness for each token, since different tokens are masked with different probabilities. This fundamentally breaks methods based on lower-tail token scores, such as MinK (Shi et al., 2023) and Min-K++ (Zhang et al., 2024). Third, membership inference methods on vision diffusion models (Ho et al., 2020) show degraded performance on DLMs due to the transition of the state space from continuous to discrete.

To address this challenge, we conduct an in-depth analysis of diffusion training dynamics and find that the model's single-step memorization gain for each token is inversely proportional to the mask ratio during diffusion training process. Based on this, we theoretically derive the property of token-

level memorization asymmetry and accordingly propose a skewness-based MI indicator: **Q**uantile-weighted Skewness (**Q-Skewness**) of the sample tokens. Our method can effectively distinguish member sets from non-member sets from a data distribution perspective, beyond the memorization at the individual sample level. We conduct experiments across different finetuning objectives, dataset domains, and types of diffusion language models, including both base and instruction-tuned diffusion language models. **Q-Skewness** consistently outperforms baselines, and this superiority remains consistent across different training epochs, with an average AUC improvement of more than 10%. Furthermore, we show that the skewness indicator can also enhance the effectiveness of PII (Personally Identifiable Information) extraction.

Notably, there is a concurrent work SAMA (Chen et al., 2026) which also focuses on membership inference. Our work differs in two aspects: (1) we propose a new indicator beyond memorization at the individual sample level, and (2) our evaluation covers both instruction fine-tuning and domain fine-tuning (same training objective with pretraining), rather than only domain fine-tuning as in prior works (Fu et al., 2024a; Chen et al., 2026). Our method consistently outperforms theirs in both settings. In summary, our contributions are:

- We are the first to identify an inverse relationship between mask ratio and token memory gain, and theoretically characterize Token-level Memorization Asymmetry.
- We propose **Q-Skewness**, a novel membership inference metric that consistently outperforms baselines across datasets and training settings.
- We provide a new perspective on privacy risks in DLMs, showing that skewness also improves PII extraction and highlighting fundamental differences from AR models.

# 2 Background and Related Work

## 2.1 Diffusion Language Models

Unlike traditional autoregressive generating, diffusion language models (DLMs) recover data from a fully masked state through iterative denoising. Currently, successful large-scale DLMs are based on the discrete masked diffusion model architecture. Taking this as an example, we formalize the training and inference processes of DLMs as follows. Let $\mathbf{x}_0 = (x_0^1, x_0^2, \ldots, x_0^L)$ denotes a sample of the training set $\mathcal{D}_{\text{train}}$, consisting of $L$ discrete tokens, where $x_0^i \in \mathcal{V}$. The diffusion process is defined on an extended vocabulary $\mathcal{V}^+ = \mathcal{V} \cup \{[\texttt{MASK}]\}$ that includes a mask token. The forward diffusion process in DLMs is defined as a Markov chain that gradually replaces tokens with mask tokens. Then we can derive a closed-form solution to sample the state $\mathbf{x}_t$ directly at timestep $t$:

$$q(x_t^i|x_0^i) = \begin{cases} 1-\bar{\alpha}_t & \text{if } x_t^i = [\texttt{MASK}] \\ \bar{\alpha}_t & \text{if } x_t^i = x_0^i \\ 0 & \text{otherwise,} \end{cases}$$

where $\bar{\alpha}_t$ denotes the cumulative probability that a token remains unmasked at timestep $t$. Then the parameterized $p_\theta(\mathbf{x}_0|\mathbf{x}_t)$ is trained to maximize the Evidence Lower Bound (ELBO), equivalent to minimizing the negative log-likelihood on the masked positions:

$$\mathcal{L}_{\text{DLM}}(\theta) = \mathbb{E}_{t,\mathbf{x}_0,\mathbf{x}_t}\left[\frac{1}{|\mathcal{M}_t|}\sum_{i\in\mathcal{M}_t} -\log p_\theta(x_0^i|\mathbf{x}_t)\right], \quad (1)$$

where $t \sim \mathcal{U}(1,T)$, $\mathbf{x}_0 \sim \mathcal{D}_{\text{train}}$, and $\mathcal{I}_t = \{i \mid x_t^i = [\texttt{MASK}]\}$ denotes indices of masked tokens in timestamp $t$. During the inference process, the model $\theta$ starts from a fully masked sequence $\mathbf{x}_T$, and performs the following update steps through iterative sampling until $t = 0$:

$$\mathbf{x}_{t-1} \sim p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t).$$

## 2.2 Membership Inference

Membership inference aims to determine whether a given data sample belongs to the member set (training data) of a target model (Shokri et al., 2017). This task is widely regarded as a core method for quantifying privacy risks or auditing unauthorized data usage (DeAlcala et al., 2024). For generative models, the high number of parameters results in significant fine-tuning overhead. Hence, existing effective membership inference mainly falls under query-based methods, which typically require a gray-box setting, i.e., obtaining output logits.

**Membership inference on DMs.** Previous membership inference methods on diffusion models mainly focus on visual (continuous) diffusion models (Ho et al., 2020; Song et al., 2020). In the grey-box setting, these methods examine the characteristics of diffusion models and design various metrics to determine member and non-member sets, such as: (1) estimation errors of deterministic forward process (Duan et al., 2023; Kong et al., 2024), (2)

probability fluctuations in the presence of perturbations (Fu et al., 2024b); (3) likelihood discrepancy with different inputs (Zhai et al., 2024), et al.

**Membership inference on LLMs.** Many studies explore membership inference for LLMs (Yeom et al., 2018; Shi et al., 2023; Zhang et al., 2024). In the fine-tuning scenario, due to the open availability of base models, methods for LLMs typically involve an additional assumption: the attacker has access to reference models. Watson et al. (2022) firstly introduces this assumption and proposes *Calibration Attack*. SPV-MIA (Fu et al., 2024a) leverages self-prompting to construct reference models, leading to more precise inference. Similarly, the concurrent work, SAMA (Chen et al., 2026), conducts membership inference on diffusion language models (DLMs) following the same setting. In contrast to our work, their empirical approach lacks theoretical analysis of the difference between DLMs and traditional AR language models, leading to less effectiveness in practice (Tab. 1, Tab. 2).

# 3 MIA on Fine-tuned DLMs

## 3.1 Threat Model

In this work, we consider an adversary that aims to infer whether a given data record was used to train the target model $\theta_{\text{tar}}$. Since the pre-training process primarily utilizes crawled public data and is implemented by only a few third-party institutions, we focus on the fine-tuning stage, which is generally recognized as the stage that is most susceptible to privacy or copyright risks (Yu et al.; Fu et al., 2024a; Chen et al., 2026).

We strictly follow the mainstream setting of query-based membership interference (Shi et al., 2023; Zhang et al., 2024; Duan et al., 2023; Watson et al., 2022), where the adversary can only access the output logits of the target model $\theta_{\text{tar}}$, without having access to the model weights or gradients. The formalization is:

$$\mathcal{A}(\mathbf{x}, \theta_{\text{tar}}) = \mathbb{1}\left[\mathcal{A}'(\mathbf{x}, \theta_{\text{tar}}) > \tau\right],$$

where $\mathcal{A}'$ denotes an indicator function that reflects membership information, and $\tau$ denotes a tunable decision threshold. Following the state-of-the-art reference-based membership inference methods (Meng et al., 2025; Watson et al., 2022; Fu et al., 2024a; Huang et al., 2025), we further assume that the adversary has access to an unfine-tuned reference model $\theta_{\text{ref}}$. Then, a more precise

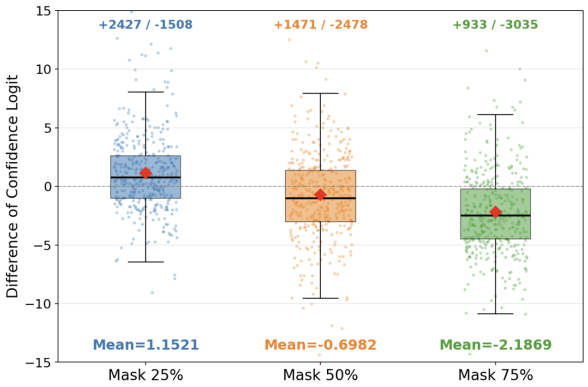


Figure 1: Violin and jitter plots of confidence logit differences before and after training. A smaller mask ratio results in a higher memorization gain for the token post-training.

strategy of membership inference can be formalized as:

$$\mathcal{A}(\mathbf{x}, \theta_{\text{tar}}) = \mathbb{1}\left[\mathcal{A}'(\mathbf{x}, \theta_{\text{tar}}, \theta_{\text{ref}}) > \tau\right].$$

Note that this setup is easily achievable for membership inference on fine-tuning scenarios due to the availability of open-sourced base models. And we further consider a weaker assumption involving misaligned reference models in Section 4.3.

**Fine-tuning setups.** Notably, previous membership inference methods on fine-tuned language models (Fu et al., 2024a; Chen et al., 2026) consider only domain fine-tuning, which shares the same training objective as the pretraining stage and utilizes a fixed training block size, such as 64, 128, or 256 tokens. However, we emphasize that these evaluation setups are limited: ❶ It does not represent the fine-tuning scenarios for instruction-tuning, which constitute the primary fine-tuning paradigm. ❷ may lead to an inflated illusion of membership inference success by failing to account for the variable-length nature of instruction-tuning datasets. Therefore, in our experiments, we evaluate both *domain fine-tuning (DFT)* and *instruction fine-tuning (IFT)* scenarios to align with real-world scenarios (Section 4).

## 3.2 Token-level Memorization Asymmetry

The training process of DLMs distinguishes them from AR language models and thereby introduces unique characteristics. In this part, we provide a formal proof characterizing the distributional differences of token-level memorization between member (training) data $\mathbf{x} \in \mathcal{D}_{train}$ and non-member (test) data $\mathbf{x} \in \mathcal{D}_{test}$.

Let $\mathbf{x}$ be a sequence of tokens $\mathbf{x} = (w_1, \ldots, w_L)$ in the training set $\mathcal{D}_{train}$. We consider a diffusion language model $\theta_K$ trained for $K$ epochs on $\mathcal{D}_{train}$, and $\theta_0$ denotes the original version that has not yet been trained. Since the model's memorization of training data gradually accumulates as the training progresses, we define the memorization gain $\mathcal{G}_{w_i}$ during the training process as the cumulative reduction in loss. For the $i$-th token $w_i$ in sequence $\mathbf{x}$, we have:

$$\mathcal{G}_{w_i} := \mathcal{L}(w_i; \theta_0) - \mathcal{L}(w_i; \theta_K) = \sum_{k=1}^{K} \delta_{w_i,k}, \quad (2)$$

where the $\delta_{w_i,k}$ denotes the memory gain (i.e. marginal loss reduction) for token $w_i$ at epoch $k$.

**Member set analysis.** For the member sample $\mathbf{x} \in \mathcal{D}_{train}$, a global mask rate $\beta_k \sim U(0,1)$ is sampled at the training epoch $k$. Let $m_{w_i,k} \in \{0,1\}$ denotes the binary indicator whether token $w_i$ is masked (and thus trained) at epoch $k$ [1]:

$$P(m_{w_i,k} = 1 \mid \beta_k) = \beta_k.$$

In the diffusion training progress, we find the increase in token memorization is associated with the masking ratio. Insightfully, low-$\beta$ steps may lead to more concentrated gradient updates, resulting in stronger single-step memory gain for $w_i$. Hence, we propose the following assumption:

**Assumption 3.1.** (Inverse Mask-Ratio Scaling) We posit that the magnitude of the memory gain, when a token is selected, is a function of the mask rate:

$$\delta_{w_i,k} = m_{w_i,k} \cdot \phi(w_i, \beta_k),$$

where $\phi : (0,1) \to \mathbb{R}^+$ denotes the update strength function reflected by the loss and is strictly monotonically decreasing.

To validate Assumption 3.1, we fine-tuned LLaDA (Nie et al., 2025) on the XSUM (Narayan et al., 2018) dataset for four epochs. During training, we set the mask ratio to three fixed values, 0.25, 0.50, and 0.75, with a fixed mask template for each data sample, instead of using a random mask ratio. After training, we utilize the prediction confidence to estimate the memorization of individual tokens. It is evident from Fig. 1 that a lower mask ratio leads to more significant token memorization on average under the same number of training epochs. We leave more experiments in Appendix B. Therefore, for a member datapoint $x$, the single-step gain $\delta_{w_i,k}$ obviously follows a zero-inflated mixture distribution. We then derive the shape properties of the cumulative gain $\mathcal{G}$.

**Theorem 3.2.** *(Proof in Appendix A.1) Let $\mathcal{G}$ be the memorization gain for token-level in the training set. For any finite number of epochs $K \geq 1$, the distribution of $\mathcal{G}_{w_i}$ is strictly right-skewed:*

$$\mathit{Skew}(\mathcal{G}_{w_i}) = \frac{\mathbb{E}\left[(\mathcal{G}_{w_i} - \mathbb{E}[\mathcal{G}_{w_i}])^3\right]}{\sigma(\mathcal{G}_{w_i})^3} > 0 \quad (3)$$

We further analyze and demonstrate in Appendix A.2 that this skewness specifically represents a long-tail distribution.

**Non-member set analysis.** For the non-member sample $\mathbf{x} \notin \mathcal{D}_{train}$, $\mathcal{G}_{w_i} = \sum_{k=1}^{K} \epsilon_{i,k}$, where $\epsilon_{w_i,k}$ represents stochastic generalization noise. Without the zero-inflated selection mechanism and the monotonic inverse mask-ratio amplification assumed for members, the resulting distribution typically exhibits much weaker asymmetry. In practice, this value is often close to zero when the aggregated fluctuations are approximately symmetric: $\text{Skew}(\mathcal{G}_{w_i}) \approx 0$. Recalling Eq. (1), we define the following for diffusion language models:

$$\mathcal{L}(w_i; \theta) := \mathbb{E}_{t,\mathbf{x}_0,\mathbf{x}_t}\left[-\log p_\theta(w_i \mid \mathbf{x}_t)\right],$$

where $\mathbf{x}_t$ denote the noisy sequence of $\mathbf{x}$ at timestep $t$. We use $\Delta_{w_i}$ as the empirical estimator of the memorization gain $\mathcal{G}_{w_i}$:

$$\Delta_{w_i}(\theta_{\text{tar}}, \theta_{\text{ref}}) := \underbrace{\mathbb{E}_{t,\mathbf{x}_t}\left[-\log p_{\theta_{\text{ref}}}(w_i \mid \mathbf{x}_t)\right]}_{\text{Memorization of Reference Model}} - \underbrace{\mathbb{E}_{t,\mathbf{x}_t}\left[-\log p_{\theta_{\text{tar}}}(w_i \mid \mathbf{x}_t)\right]}_{\text{Memorization of Target Model}}. \quad (4)$$

Assuming the reference model approximates the untrained state, i.e. $\theta_{\text{ref}} \approx \theta_0$, and the target model is the trained state, i.e. $\theta_{\text{tar}} = \theta_K$ after $K$ epoch training, the $\Delta_{w_i}(\theta_{\text{tar}}, \theta_{\text{ref}})$ shares the same distribution properties (Theorem 3.2). Hence, the following equation holds:

$$\begin{aligned}&\text{Skew}(\{\Delta_{w_{\text{member},i}}(\theta_{\text{tar}}, \theta_{\text{ref}}) \mid i \in \{1, \ldots, L\}\}) \\ &> 0 \approx \text{Skew}(\{\Delta_{w_{\text{non-member},i}}(\theta_{\text{tar}}, \theta_{\text{ref}}) \mid i \in \{1, \ldots, L\}\})\end{aligned}$$

where $w_{\text{member},i}$ and $w_{\text{non-member},i}$ denote the tokens of the member set and non-member set samples, respectively. We define:

$$\mathbb{I}(\mathbf{x}, \theta_{\text{tar}}, \theta_{\text{ref}}) = \text{Skew}(\{\Delta_{w_i}(\theta_{\text{tar}}, \theta_{\text{ref}}) | i = 0, 1, 2 \ldots\}),$$

[1] For clarity, we do not consider the impact of parameter updates during training on other masked tokens. In practice, parameter changes from training on other masked tokens may introduce noise that is unbiased on average, and thus does not affect our subsequent conclusions.

where $\mathbf{x} = (w_1, \ldots, w_L)$. For a given data sample $\mathbf{x}$, a reference model $\theta_{\text{ref}}$, and a target model $\theta_{\text{tar}}$, the indicator function $\mathbb{I}(\mathbf{x}, \theta_{\text{tar}}, \theta_{\text{ref}})$ serves as a metric for membership inference. Intuitively, if $\mathbb{I}(\mathbf{x}, \theta_{\text{tar}}, \theta_{\text{ref}})$ exceeds a threshold $\tau$, the sample $\mathbf{x}$ is likely from the member set; otherwise, it belongs to the non-member set.

### 3.3 Quantile-weighted Skewness with Cyclic Sampling

In practice, we (1) calculate the loss difference for every individual mask sampling instead of the mathematical expectation, increasing the set size for distribution estimation; (2) design a cyclic sampling method to replace the random Monte Carlo sampling of $t$, eliminating the randomness in scenarios with a low number of samples, ensuring coverage uniformity; (3) design a more robust metric based on Eq. (3), to distinguish between the member set and the non-member set.

**Token-level Memorization Score.** Let $\mathbf{x} = (w_1, \ldots, w_L)$ be an input sequence. For any mask subset $\mathcal{M} \subset \{1, \ldots, L\}$ sampled at a time step $t \sim \mathcal{T}$, we define the instance loss difference $\Delta_{w_i,\mathcal{M}}$ for a masked token $w_i$ ($i \in \mathcal{M}$) as:

$$\begin{aligned}\Delta_{i,\mathcal{M}} := &[-\log p_{\theta_{\text{ref}}}(w_i \mid \mathbf{x}_{\mathcal{M}^c}, t)] \\ &- [-\log p_{\theta_{\text{tar}}}(w_i \mid \mathbf{x}_{\mathcal{M}^c}, t)]\,,\end{aligned} \tag{5}$$

where $\mathbf{x}_{\mathcal{M}^c}$ denotes the context provided by unmasked tokens.

**Cyclic Sampling.** To mitigate the stochastic variance inherent in Monte Carlo sampling and ensure uniform token coverage, we employ a Cyclic Sampling strategy. Instead of random replacement, we perform $R$ full-coverage round sampling. In the $r$-th rounds, we generate a random permutation of indices $\pi^{(r)} = \text{Permutation}(\{1, \ldots, L\})$. This permutation is partitioned into a sequence of disjoint batches $\mathbf{B}^{(r)} = \{\mathcal{M}_{r,1}, \ldots, \mathcal{M}_{r,m}\}$, such that $\bigcup_j \mathcal{M}_{r,j} = \{1, \ldots, L\}$ and $\mathcal{M}_{r,j} \cap \mathcal{M}_{r,j'} = \emptyset$, where each batch has size $|\mathcal{M}| \approx \rho L$. The final memorization set aggregates the scores from all disjoint batches across all cycles:

$$\mathcal{S}_{\mathbf{x}} = \bigcup_{r=1}^{R} \bigcup_{j=1}^{m} \left\{\Delta_{i,\mathcal{M}_{r,j}} \mid i \in \mathcal{M}_{r,j}\right\}. \tag{6}$$

**Quantile-weighted Skewness.** In practice, extreme outliers can significantly inflate the mean of the entire set, making it difficult to distinguish between member and non-member samples. Therefore, we consider the median-based metric following Pearson's Second Skewness Coefficient:

$$\frac{\mu(\mathcal{S}_{\mathbf{x}}) - \tilde{\mathcal{S}}_{\mathbf{x}}}{\sigma(\mathcal{S}_{\mathbf{x}})}, \tag{7}$$

where $\tilde{\mathcal{S}}_{\mathbf{x}}$ denotes the median of $\mathcal{S}_{\mathbf{x}}$. Since different quantiles exhibit varying impacts on membership inference accuracy, we further define the *Quantile-weighted Skewness* (Q-Skewness):

$$\begin{aligned}&\mathcal{A}'(x, \theta_{\text{tar}}, \theta_{\text{ref}}) := \\ &\quad f_{\text{QSkew}}(\mathbf{x}) = \frac{\frac{1}{|\mathcal{S}_{\mathbf{x}}|}\sum_{s \in \mathcal{S}_{\mathbf{x}}} W(s) \cdot (s - \tilde{\mathcal{S}}_{\mathbf{x}})}{\sigma(\mathcal{S}_{\mathbf{x}}) \cdot \bar{W}},\end{aligned} \tag{8}$$

where $\bar{W} = \frac{1}{|\mathcal{S}_{\mathbf{x}}|}\sum_{s \in \mathcal{S}_{\mathbf{x}}} W(s)$ is a normalization factor. To mitigate the influence of outliers and focus on informative regions, we formulate the weight function $W(s)$ using the empirical cumulative distribution function (ECDF):

$$W(s) = \exp\left(-\frac{(F(s)-0.15)^2}{2\tau^2}\right) + \exp\left(-\frac{(F(s)-0.85)^2}{2\tau^2}\right), \tag{9}$$

where $F(s) \in [0, 1]$ represents the quantile position of $s$ within $\mathcal{S}_{\mathbf{x}}$, and $\tau$ is a bandwidth hyperparameter (set to 0.1 in experiment). By anchoring weights at the 15th and 85th percentiles, this metric captures distributional asymmetry while remaining robust to long-tailed distributions.

## 4 Experiments

**Datasets and Models.** According to Section 3.1, We consider both domain finetuning and instruction finetuning stages on language diffusion models across different dataset domains. Specifically, for (1) domain finetuning (DFT), we finetune LLaDA-Base-8B (GSAI-ML, 2025a) with ArXiv (RealTimeData, 2025a), WikiText (RealTimeData, 2025b), and XSUM (Narayan et al., 2018) datasets. and for (2) instruction finetuning (IFT), we finetune LLaDA-Instruct-8B (GSAI-ML, 2025b) with MedQA (Li et al., 2023), Alpaca (Taori et al., 2023), and Tulu-3 (Lambert et al., 2024) datasets.

We do not consider MIMIR (Duan et al., 2024b) as (Chen et al., 2026), because the training (member) and test (non-member) sets in MIMIR already exhibit a prior discrepancy within the model, making it unsuitable for evaluating fine-tuning scenarios. We detail the validation in Appendix C.1.

**Baselines.** We consider a broad range of membership inference baselines from the following sources:

Table 1: The main experiment results of membership inference on diffusion language models on domain fine-tuning tasks (DFT). We bold the best results, and our method is marked in grey.

| | Arxiv (DFT) | | | WikiText (DFT) | | | XSUM (DFT) | | |
|---|---|---|---|---|---|---|---|---|---|
| | AUC | TPR@10%FPR | TPR@1%FPR | AUC | TPR@10%FPR | TPR@1%FPR | AUC | TPR@10%FPR | TPR@1%FPR |
| Loss | 61.3 | 16.4 | 5.0 | 60.3 | 17.1 | 2.0 | 61.1 | 15.8 | 1.7 |
| MinK | 64.5 | 24.1 | 2.7 | 62.5 | 24.0 | 8.7 | 65.8 | 29.0 | 4.8 |
| MinK++ | 66.4 | 29.0 | 1.3 | 56.9 | 16.2 | 2.3 | 61.2 | 27.3 | 12.0 |
| Calibration | 59.1 | 22.7 | 5.3 | 56.9 | 19.7 | 2.0 | 56.4 | 16.8 | 2.0 |
| $\text{SecMI}_{\text{DLM}}$ | 58.0 | 18.8 | 2.0 | 62.4 | 26.0 | 1.7 | 64.8 | 31.7 | 17.3 |
| SAMA | 70.2 | 34.5 | 17.6 | 70.3 | 36.4 | 17.2 | 71.0 | 40.5 | 18.2 |
| **Q-Skew** | **83.4** | **66.1** | **50.6** | **80.3** | **57.7** | **39.7** | **80.7** | **63.6** | **47.2** |

Table 2: The main experiment results of membership inference on diffusion language models on instruction fine-tuning (IFT) tasks. We highlight the important values as Tab. 1.

| | MedQA (IFT) | | | Alpaca (IFT) | | | Tulu-3 (IFT) | | |
|---|---|---|---|---|---|---|---|---|---|
| | AUC | TPR@10%FPR | TPR@1%FPR | AUC | TPR@10%FPR | TPR@1%FPR | AUC | TPR@10%FPR | TPR@1%FPR |
| Loss | 57.3 | 18.8 | 2.7 | 57.4 | 14.7 | 1.0 | 54.9 | 12.0 | 2.3 |
| MinK | 60.7 | 16.2 | 3.0 | 67.3 | 27.6 | 3.0 | 57.2 | 11.3 | 3.3 |
| MinK++ | 49.7 | 8.7 | 2.3 | 59.6 | 21.2 | 1.7 | 55.1 | 13.0 | 2.7 |
| Calibration | 55.6 | 17.1 | 1.0 | 61.7 | 20.3 | 5.0 | 60.1 | **15.9** | 2.6 |
| $\text{SecMI}_{\text{DLM}}$ | 54.0 | 11.7 | 2.3 | 51.2 | 15.0 | 4.7 | 56.8 | 12.3 | 3.3 |
| SAMA | 65.1 | 26.0 | 3.2 | 69.1 | 31.0 | 11.2 | 59.7 | 12.1 | 3.7 |
| **Q-Skew** | **72.8** | **31.6** | **6.7** | **81.4** | **61.4** | **41.6** | **67.2** | 13.7 | **4.3** |

❶ methods targeting traditional auto-regressive language models: Loss (Yeom et al., 2018), MinK (Shi et al., 2023), Mink++ (Zhang et al., 2024), Calibration (Watson et al., 2021); ❷ adaptions of methods designed for (vision) diffusion models: SecMI (Duan et al., 2023); and ❸ one concurrent work: SAMA (Chen et al., 2026).

**Metrics.** We consider the widely used evaluation metrics, including the area under the receiver operating characteristic curve (AUC), the True Positive Rate when the False Positive Rate of 10% and 1% (TPR@10%FPR, TPR@1%FPR), following previous works (Carlini et al., 2022; Duan et al., 2023; Fu et al., 2024a; Chen et al., 2026).

**Implementation.** We conduct three independent runs and report the average of each metric to reduce random error. We uniformly train for 6 epochs with a batch size of 16. We also consider other training intensities (Section 4.2). For all methods, we apply 16 denoising steps per sample to ensure consistent computational complexity. When calculating the token-level memorization score, we set $t = 0.35$ in (5), as this noise level maximizes the distinction between the member and non-member sets. We also conduct ablation studies to show the effectiveness of each component of our method in Appendix E.

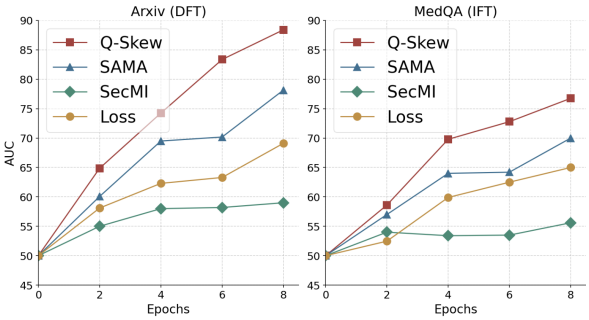


Figure 2: Effectiveness under different training epochs.

## 4.1 Main Results

As reported in Tab. 1 and Tab. 2, the **Q-Skewness** (our method) indicator consistently outperforms the baselines across various dataset distributions, different language diffusion models, and multiple training stages, including domain fine-tuning and instruction fine-tuning.

## 4.2 Performance on Various Training Epochs

As the fine-tuning processes, the model's memorization of the training data gradually increases. In practical scenarios, the number of fine-tuning steps by the model trainer varies. Membership inference methods that more effectively reveal membership information across different steps are considered superior (Zhai et al., 2024; Lian et al., 2025). We evaluate the effectiveness of the membership infer-

Table 3: Performance of membership inference methods under a weaker assumption: using misaligned reference models

| | Base DLM | | |
|---|---|---|---|
| | AUC | TPR@10 | TPR@1 |
| Golden Reference | 83.4 | 66.1 | 50.6 |
| Instruction Reference | **78.4** (-5.0) | 55.5 (-10.7) | **45.5** (-5.1) |
| Para-modification | 78.4 (-5.0) | **60.0** (-6.2) | 45.0 (-5.6) |
| SAMA (w/ Ref) | 70.2 | 34.4 | 17.6 |

| | Instruction DLM | | |
|---|---|---|---|
| | AUC | TPR@10 | TPR@1 |
| Golden Reference | 72.8 | 31.6 | 6.7 |
| Base Reference | **71.5** (-1.3) | **32.0** (+0.4) | **6.0** (-0.7) |
| Para-modification | 70.0 (-2.8) | 25.0 (-6.6) | 2.5 (-4.2) |
| SAMA (w/ Ref) | 65.1 | 26.0 | 3.2 |

ence method across different training epochs Fig. 2. Experimental results demonstrate that our method consistently outperforms the baselines across different training steps.

### 4.3 Weaker Assumption of Reference Models

Although it is reasonable to assume that the MI adversary can access an opensource unfinetuned reference model in the finetuning setting, we additionally consider the effectiveness of our method when the used reference model is misaligned. We consider the following three alternative strategies when the unfinetuned model corresponding to the target model is unavailable: (1) Use an unfinetuned base model as a substitute for the unfinetuned instruction model. (2) Use an unfinetuned instruction model as a substitute for the unfinetuned base model. (3) Use a model that has been fine-tuned on other perturbed data as a substitute for the unfinetuned model.

We test the situation of domain fine-tuning on the ArXiv dataset and instruction fine-tuning on the MedQA dataset. In Tab. 3, while the membership inference performance of our method exhibits a slight decline when the reference model is misaligned, this decrease is minor, and our approach still outperforms the best baseline.

## 5 Token-Level Asymmetry Signal Increases PII Reconstruction Risks

Previously, we showed that our skewness metric improves membership inference attacks. We now consider a complementary privacy risk in generative language models: *data reconstruction attacks* (Carlini et al., 2018, 2023; Hayes et al., 2024; Biderman et al., 2023; Lee et al., 2022; Ippolito et al., 2023). Specifically, we investigate whether Q-skewness—designed to capture distributional difference between seen and unseen tokens—can be effective in extracting personally identifiable information (PII).

### 5.1 Threat Model

We consider a PII reconstruction adversary (Lukas et al., 2023; Meng et al., 2025; Nakka et al., 2024; Kim et al., 2023) who is given a training record with PII redacted and aims to recover the redacted tokens via queries to the target model. For example, given the PII-redacted record "`Insurance Member ID: [REDACTED]`", the objective is to reconstruct the true value "`XJ45-7831-92`". The adversary can generate *candidate PII values* and *rank* them to identify the true PII.

### 5.2 PII Reconstruction for DLMs

PII reconstruction attacks have been widely studied for autoregressive language models, where attackers query the model and identify memorized sequences via high conditional probabilities. However, these approaches do *not* directly apply to DLMs. (1) DLMs generate tokens through iterative denoising rather than left-to-right prediction, and thus do not expose conditional probabilities of the form $P(x_t \mid x_{<t})$. (2) Prior attacks rely on directly prompting models to reproduce memorized sequences, whereas DLMs reconstruct tokens within partially noisy contexts, making such generation-based extraction less reliable.

**Candidate-based PII extraction.** To adapt existing approaches, we formulate PII extraction as a candidate ranking problem. Given a context containing a masked PII record and candidate PII items $\{z_1, z_2, \ldots, z_k\}$, the attacker aims to identify the most likely memorized candidate. Candidates are generated via random remasking across $N = 64$ independent denoising process. Each candidate $z_i$ is inserted into the masked position, and its plausibility is estimated by computing the cross-entropy loss $l_{CE}$ under stochastic masking. This likelihood-based ranking serves as a baseline.

**Incorporating our skewness score.** Previous likelihood-based ranking provides a strong baseline for PII extraction by measuring the reconstruction loss of candidate completions. However, this approach primarily reflects the overall sequence likelihood under the model's *language prior*, which can cause different candi-

dates to appear similarly plausible. For example, PII records such as "`john.doe@gmail.com`" and "`alice.bob@gmail.com`" may both receive similar $l_{CE}$ values because common tokens like "`gmail.com`" are strongly supported by the language prior. As a result, the ranking may be dominated by generic token probabilities rather than signals of memorization. Our skewness score (token-level asymmetry) in contrast, captures localized reconstruction behavior and can highlight asymmetries that arise from memorized sequences. This makes it a *complementary* signal for candidate ranking. To combine the strengths of both signals, we compute a weighted score:

$$\begin{aligned}\text{Score}(z_i) =&(1-\alpha)\cdot Z(l_{CE}(z_i))\\ &+(\alpha)\cdot Z(\text{Skew}_{\text{Pearson2}}(z_i))\end{aligned}$$

where $Z$ denotes z-score normalization. In practice, an adversary may have access to a few training records and use them to estimate an optimal $\alpha$. We demonstrate this in G. To better capture localized memorization effects, we compute skewness over a contextual span of length $X$ centered on the PII record. For example, when $X = 20$, the span includes 8 preceding context tokens, the 5 PII tokens, and 7 following tokens. For each candidate sequence, we generate masked variants by randomly masking 35% of tokens within this span. The process is repeated for 4–16 batches. Token-level loss differences are computed across these masked variants and aggregated to estimate the final score.

### 5.3 Evaluation

We evaluate our attacks on two types of PII: *phone* numbers and *email* addresses. To this end, we use the TREC dataset (Wu et al., 2006), which contains 174,299 records, split into 78k, 78k, and 17k examples for training, validation, and testing, respectively. Consistent with our membership inference setup, we fine-tune the LLaDA-8B-Base model on the training set. We identify PII instances applying regular expressions that match email addresses and phone numbers, and randomly sample 50 unique email addresses and 50 unique phone numbers as targets. The attacker generates candidate PII by inserting 20 mask tokens at the target location within a record and running 64 independent diffusion processes to fill them, then selects the most likely candidate using the attack signals defined in §5.2. We report attack success rate (ASR) as the proportion of cases in which the true PII appears as the top-1 candidate among the 50 targets of each class. Additional details are provided in Appendix D.

Table 4: ASR (%) of our attack methods on LLaDA-8B-Base for phone numbers and email addresses. Higher is better. The best results on `Combine-0.1` are in bold.

| Our Attacks | Signal | ASR (%) | |
|---|---|---|---|
| | | Phone | Email |
| `Ranking` | CE | 30 [15/50] | 20 [10/50] |
| `Mask only PII` | Q-Skewness | 22 [11/50] | 16 [ 8/50] |
| `Mask only context` | Q-Skewness | 10 [ 5/50] | 8 [ 4/50] |
| `Combine-0.1` | Both | **34** [17/50] | **24** [12/50] |

**Results.** Table 4 summarizes our results. We compare four attack methods: cross-entropy (CE)-based ranking (`Ranking`), two skewness-based variants that mask either only PII tokens (`Mask only PII`) or only context tokens (`Mask only context`), and a combined method (`Combine-0.1`) that integrates CE and skewness signals. The skewness-based method names reflect whether the randomly masked tokens are applied to PII or context tokens. The value 0.1 denotes the $\alpha$ value used to combine the two (see Appendix G for details on how we select the optimal $\alpha$).

We show that CE-based ranking serves as a strong baseline, achieving 30% and 20% ASR on phone numbers and emails. Skewness-based methods alone performs worse (10–22% ASR on phone numbers and 8–16% on emails), reflecting that PII reconstruction is primarily a ranking task rather than pure inference. However, combining CE with skewness produces the best results, improving ASR to 34% for phone numbers and 24% for emails. This suggests that the skewness helps mitigate the influence of strong language priors acquired during pre-training, which can bias CE-based ranking. Additional qualitative analysis supporting this claim are in Appendix H, and ablation studies are deferred to Appendix F due to space limits.

## 6 Conclusion

In this paper, we first define the *token-level memorization asymmetry*. Based on this, we theoretically derive an MI method for fine-tuned DLMs that outperforms existing baselines across datasets and setups. Furthermore, we demonstrate that our method facilitates PII extraction. Our work indicates that, due to differences in the training process, DLMs exhibit distinct characteristics from autoregressive (AR) LLMs, encouraging the community to focus on the unique privacy risks of DLMs.

## Limitations

While our method advances the privacy-violating attacks such as membership inference and PII extraction in diffusion language models, several limitations remain. First, because publicly available open source diffusion language models that are suitable for training are still limited, we mainly evaluate our method on current mainstream models. Moreover, diffusion language models are developing rapidly. More diverse diffusion architectures and training mechanisms may emerge in the future, and the generalization ability of our method to such future models remains unknown.

## Impact Statement

This work advances our understanding of privacy risks in DLMs, an emerging class of generative models. The proposed attack, **Q-Skew**, is intended as auditing tools to assess and mitigate privacy risks, rather than to facilitate misuse. We believe that exposing this vulnerability is a necessary step toward developing effective privacy defenses and informing the responsible deployment of DLMs. More broadly, this work underscores the importance of evaluating privacy guarantees in emerging model architectures and encourages the development of privacy-preserving training and inference techniques for diffusion-based models.

## References


Stella Biderman, USVSN Sai Prashanth, Lintang Sutawika, Hailey Schoelkopf, Quentin Gregory Anthony, Shivanshu Purohit, and Edward Raff. 2023. Emergent and predictable memorization in large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Tiwei Bie, Maosong Cao, Kun Chen, Lun Du, Mingliang Gong, Zhuochen Gong, Yanmei Gu, Jiaqi Hu, Zenan Huang, Zhenzhong Lan, and 1 others. 2025. Llada2. 0: Scaling up diffusion language models to 100b. *arXiv preprint arXiv:2512.15745*.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. 2022. Membership inference attacks from first principles. In *2022 IEEE symposium on security and privacy (SP)*, pages 1897–1914. IEEE.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Xiaodong Song. 2018. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *USENIX Security Symposium*.

Yuetian Chen, Kaiyuan Zhang, Yuntao Du, Edoardo Stoppa, Charles Fleming, Ashish Kundu, Bruno Ribeiro, and Ninghui Li. 2026. Membership inference attacks against fine-tuned diffusion language models. In *The Fourteenth International Conference on Learning Representations*.

Valentin De Bortoli, Alexandre Galashov, Arthur Gretton, and Arnaud Doucet. 2025. Accelerated diffusion models via speculative sampling. In *International Conference on Machine Learning*, pages 12590–12631. PMLR.

Daniel DeAlcala, Aythami Morales, Julian Fierrez, Gonzalo Mancera, Ruben Tolosana, and Javier Ortega-Garcia. 2024. Is my data in your ai model? membership inference test with application to face images. *arXiv preprint arXiv:2402.09225*.

Jinhao Duan, Fei Kong, Shiqi Wang, Xiaoshuang Shi, and Kaidi Xu. 2023. Are diffusion models vulnerable to membership inference attacks? In *International Conference on Machine Learning*, pages 8717–8730. PMLR.

Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. 2024a. Do membership inference attacks work on large language models? *arXiv preprint arXiv:2402.07841*.

Michael Duan and 1 others. 2024b. Mimir dataset. https://huggingface.co/datasets/iamgroot42/mimir. Accessed: 2026-03-17.

Wenjie Fu, Huandong Wang, Chen Gao, Guanghua Liu, Yong Li, and Tao Jiang. 2024a. Membership inference attacks against fine-tuned large language models via self-prompt calibration. *Advances in Neural Information Processing Systems*, 37:134981–135010.

Wenjie Fu, Huandong Wang, Liyuan Zhang, Chen Gao, Yong Li, and Tao Jiang. 2024b. A probabilistic fluctuation based membership inference attack for diffusion models. *Preprint*, arXiv:2308.12143.

Google DeepMind. 2025. Gemini diffusion. https://deepmind.google/models/gemini-diffusion/. Accessed: 2026-03-16.

GSAI-ML. 2025a. Llada-8b-base. https://huggingface.co/GSAI-ML/LLaDA-8B-Base. Accessed: 2026-03-17.

GSAI-ML. 2025b. Llada-8b-instruct. https://huggingface.co/GSAI-ML/LLaDA-8B-Instruct. Accessed: 2026-03-17.

Jamie Hayes, Marika Swanberg, Harsh Chaudhari, Itay Yona, and Ilia Shumailov. 2024. Measuring memorization in language models via probabilistic extraction. In *North American Chapter of the Association for Computational Linguistics*.

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

Zhiheng Huang, Yannan Liu, Daojing He, and Yu Li. 2025. Df-mia: A distribution-free membership inference attack on fine-tuned large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 343–351.

Inception Labs. 2025. Introducing mercury, our general chat diffusion large language model. https://www.inceptionlabs.ai/blog/introducing-mercury-our-general-chat-model. Accessed: 2026-03-16.

Daphne Ippolito, Florian Tramer, Milad Nasr, Chiyuan Zhang, Matthew Jagielski, Katherine Lee, Christopher Choquette Choo, and Nicholas Carlini. 2023. Preventing generation of verbatim memorization in language models gives a false sense of privacy. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 28–53, Prague, Czechia. Association for Computational Linguistics.

Siwon Kim, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sungroh Yoon, and Seong Joon Oh. 2023. ProPILE: Probing privacy leakage in large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Fei Kong, Jinhao Duan, RuiPeng Ma, Heng Tao Shen, Xiaoshuang Shi, Xiaofeng Zhu, and Kaidi Xu. 2024. An efficient membership inference attack for the diffusion model by proximal initialization. In *The Twelfth International Conference on Learning Representations*.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, and 1 others. 2024. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, Dublin, Ireland. Association for Computational Linguistics.

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

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6).

Puwei Lian, Yujun Cai, Songze Li, and Bingkun Bao. 2025. Unveiling impact of frequency components on membership inference attacks for diffusion models. *arXiv preprint arXiv:2505.20955*.

Nils Lukas, A. Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-B'eguelin. 2023. Analyzing leakage of personally identifiable information in language models. *2023 IEEE Symposium on Security and Privacy (SP)*, pages 346–363.

Wenlong Meng, Guo Zhenyuan, Lenan Wu, Chen Gong, Wenyan Liu, Weixian Li, Chengkun Wei, and Wenzhi Chen. 2025. Rr: Unveiling llm training privacy through recollection and ranking. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17383–17397.

Krishna Kanth Nakka, Ahmed Frikha, Ricardo Mendes, Xue Jiang, and Xuebing Zhou. 2024. PII-compass: Guiding LLM training data extraction prompts towards the target PII via grounding. In *Proceedings of the Fifth Workshop on Privacy in Natural Language Processing*, pages 63–73, Bangkok, Thailand. Association for Computational Linguistics.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 1797–1807.

Jinjie Ni, Qian Liu, Longxu Dou, Chao Du, Zili Wang, Hang Yan, Tianyu Pang, and Michael Qizhe Shieh. 2025. Diffusion language models are super data learners. *arXiv preprint arXiv:2511.03276*.

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

RealTimeData. 2025a. Realtimedata arxiv alltime dataset. https://huggingface.co/datasets/RealTimeData/arxiv_alltime. Accessed: 2026-03-17.

RealTimeData. 2025b. Realtimedata wikitext alltime dataset. https://huggingface.co/datasets/RealTimeData/wikitext_alltime. Accessed: 2026-03-17.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2023. Detecting pretraining data from large language models. *arXiv preprint arXiv:2310.16789*.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE.

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Yuxuan Song, Zheng Zhang, Cheng Luo, Pengyang Gao, Fan Xia, Hao Luo, Zheng Li, Yuehang Yang, Hongli Yu, Xingwei Qu, and 1 others. 2025. Seed diffusion: A large-scale diffusion language model with high-speed inference. *arXiv preprint arXiv:2508.02193*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Lauren Watson, Chuan Guo, Graham Cormode, and Alex Sablayrolles. 2021. On the importance of difficulty calibration in membership inference attacks. *arXiv preprint arXiv:2111.08440*.

Lauren Watson, Chuan Guo, Graham Cormode, and Alexandre Sablayrolles. 2022. On the importance of difficulty calibration in membership inference attacks. In *International Conference on Learning Representations*.

Yejun Wu, Douglas W. Oard, and Ian Soboroff. 2006. An exploratory study of the w3c mailing list test collection for retrieval of emails with pro/con argument. In *International Conference on Email and Anti-Spam*.

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

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pages 268–282. IEEE.

Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, and 1 others. Differentially private fine-tuning of language models. In *International Conference on Learning Representations*.

Shengfang Zhai, Huanran Chen, Yinpeng Dong, Jiajun Li, Qingni Shen, Yansong Gao, Hang Su, and Yang Liu. 2024. Membership inference on text-to-image diffusion models via conditional likelihood discrepancy. *Advances in Neural Information Processing Systems*, 37:74122–74146.

Jingyang Zhang, Jingwei Sun, Eric Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Frank Yang, and Hai Li. 2024. Min-k%++: Improved baseline for detecting pre-training data from large language models. *arXiv preprint arXiv:2404.02936*.

# A Theoretical Analysis

## A.1 Proof of Theorem 3.2

*Proof of Theorem 3.2.* For brevity, we omit the token index $w_i$ for the single-epoch memory gain and denote it as $\delta$ Consider the random variable for a single epoch $\delta = m \cdot \phi(\beta)$ (subscripts omitted for brevity). The marginal probability of a token being selected is $P(m = 1) = \mathbb{E}[\beta] = 0.5$. Conditioned on selection ($m = 1$), the probability density of the mask rate becomes:

$$f_{\beta|m=1}(\beta) = 2\beta.$$

Let $V$ be a random variable drawn from this conditional distribution, representing the update magnitude when a token is selected, i.e., $V = \phi(\beta)|_{m=1}$. We can thus rewrite the unconditional update as $\delta = J \cdot V$, where the indicator $J \sim \text{Bernoulli}(0.5)$ is independent of $V$.

To determine skewness, we analyze the third central moment $\mu_3(\delta)$. Using the definition of moments for this mixture:

$$\mu_3(\delta) = \frac{1}{4}\left(2\mathbb{E}[V^3] - 3\mathbb{E}[V]\mathbb{E}[V^2] + (\mathbb{E}[V])^3\right)$$

Applying the AM-GM Inequality and Lyapunov's Inequality:

$$(\mathbb{E}[V^3])^{1/3} \geq (\mathbb{E}[V^2])^{1/2},$$

and given that $\phi(\beta)$ is non-constant (implying $\text{Var}(V) > 0$), we obtain the strict inequality:

$$\begin{aligned} &2\mathbb{E}[V^3] + (\mathbb{E}[V])^3 > 3\mathbb{E}[V]\mathbb{E}[V^2] \\ &\implies \mu_3(\delta) > 0 \end{aligned}$$

To extend this to the cumulative gain $\mathcal{G}_{w_i} = \sum_{k=1}^{K} \delta_k$, we assume for analytical tractability that the token-level updates $\delta_k$ are approximately independent and identically distributed (i.i.d.) across epochs within a local fine-tuning window. Then the skewness of the sum scales by $K^{-1/2}$, preserving the positive asymmetry:

$$Skew(\mathcal{G}_{w_i}) = \frac{1}{\sqrt{K}} Skew(\delta) > 0$$

□

## A.2 Long Tail Analysis

*Remark* A.1 (Mechanism of Tail Formation). While Theorem 3.2 proves asymmetry, the specific "heavy-tailed" shape arises from the interaction between the sampling probability and the update strength. Consider the probability density function (PDF) $f_V(v)$ of the update magnitude $V$. By the change of variables, $f_V(v) = f_{\beta|m=1}(\phi^{-1}(v)) \cdot |\frac{d\beta}{dv}|$. Since $f_{\beta|m=1}(\beta) = 2\beta$, we have:

$$f_V(v) \propto \beta(v) \cdot \left|\frac{d\beta}{dv}\right|$$

Since $\phi(\beta)$ is monotonically decreasing, high memory gains (large $v$) correspond to low mask rates (small $\beta$). Crucially, as $v \to \infty$, $\beta \to 0$. This implies that the probability weight term $\beta(v)$ vanishes for large memory updates. Physically, this means extreme memorization events are structurally rare: they require a low $\beta$ to generate a large gradient, but low $\beta$ inherently reduces the probability of the token being sampled. This probabilistic suppression forces the density to decay, preventing the formation of a high-value mode (bimodality) and instead forming a monotonically decaying long tail.

# B Inverse Mask-Ratio Scaling with Diverse Metrics

We use different metrics to estimate the memorization of individual tokens, including prediction confidence, with or without softmax, and token-level loss. We present the experimental results using violin plots with jitter plots (Fig. 3). It is evident that a lower mask ratio leads to more significant token memorization on average under the same number of training epochs, across different metrics.

# C Dataset Usage

## C.1 Hallucination of Membership Inference Success on the MIMIR Dataset.

Note that the evaluation of membership inference (MI) on fine-tuned models differs fundamentally from that on pre-trained models (Duan et al., 2024a). For MI on pre-trained language models, researchers perform no additional training and utilize data from distinct release periods (Duan et al., 2024b) to detect whether the model memorized specific samples. In contrast, evaluation data for MI during the fine-tuning stage must ensure that the model possesses equivalent prior knowledge for both member and non-member samples. This

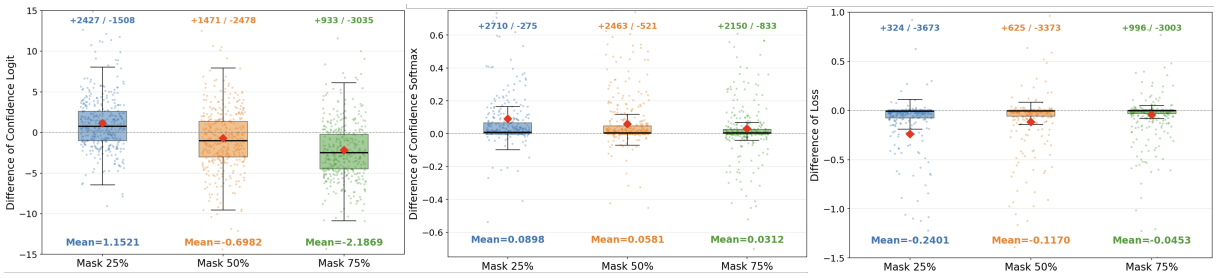


Figure 3: Inverse Mask-Ratio Scaling phenomenon with different metrics.

Table 5: Performance of Loss MI on unfinetuned model with MIMIR datasets. The high AUC values indicate that the **LLaDA model possesses a strong prior on the MIMIR training (member) set**, which leads to a hallucination of membership inference success.

| Dataset | AUC | TPR@10%FPR | TPR@1%FPR |
|---|---|---|---|
| Arxiv | 0.70 | 0.32 | 0.05 |
| Github | 0.81 | 0.47 | 0.07 |
| Hackernews | 0.56 | 0.18 | 0.02 |
| Pile_cc | 0.53 | 0.13 | 0.02 |
| Pubmed_central | 0.65 | 0.22 | 0.01 |
| Wikipedia_(en) | 0.63 | 0.28 | 0.05 |
| **Average** | **0.64** | **0.27** | **0.04** |

consistency ensures that MI success stems from actual memory extraction rather than internal model priors. We note that a concurrent work (Chen et al., 2026) fine-tunes LLaDA-8B-Base on MIMIR (Duan et al., 2024b) to evaluate its MI performance. However, MIMIR is specifically designed for pre-training MI analysis. We evaluate the unfinetuned LLaDA on the MIMIR dataset using a loss-based MI method and report the results in Appendix C.1. We find that the model exhibits strong inherent priors on MIMIR even without any fine-tuning. This bias suggests that seemingly successful MI results may be independent of the inference method itself, making MIMIR unsuitable for evaluating MI during the fine-tuning phase.

### C.2 Dataset Construction

To ensure a fair evaluation, we utilize recently collected data (RealTimeData, 2025a,b) for the Arxiv and Wiki datasets to guarantee that the model has no prior knowledge of the training or test sets. For the other dataset in Section 4, we clean the samples to prevent leakage between member and non-member sets and to ensure that model internal priors remain consistent.

## D Detail Experimental Setups

### D.1 PII Reconstruction

**Implementation.** The PII reconstruction attack we perform is composed of two steps: candidate generation and candidate ranking. The attacker has access to a training record where the target PII is redacted, and therefore knows both the surrounding context and target PII location. For the DLM-aligned candidate generation method, the attacker first replaces the target PII location with 20 mask tokens, then forces the target model to fill them in over 128 diffusion steps 64 separate times, using random remasking. The attacker then uses a regular expression over the 64 generated texts to idenitfy generated candidate PII of the target class. Each candidate is then inserted into the target PII location. This new record where the candidate occupies the location of the target PII is then used to create a candidate score.

**Metrics and Baselines.** For our baseline, we use cross entropy loss as the candidate-scoring mechanism. We compute the cross-entropy loss over the candidate PII tokens for each candidate in a set, and then rank them. Our attack success metric, top-1 accuracy, measures the proportion of targets in which the ground-truth PII achieves the highest rank. Following prior work, we report top-1 accuracy as a value between 0 and 100. In the case of cross-entropy loss we expect the ground-truth PII to have a lower loss compared to the candidates. In contrast, we expect the ground-truth PII to have a higher skewness relative to other candidates.

## E Ablation Studies

We conduct an ablation study on the domain fine-tuning task using the WikiText (RealTimeData, 2025b) dataset. We consider the following settings: (1) without the **Cyclic Sampling** strategy. (2) replacing the skewness calculation with the mean value. (3) using the skewness metric without quan-

Table 6: Performance of ablation studies on the domain finetuning task.

| | AUC | TPR@10%FPR | TPR@1%FPR |
|---|---|---|---|
| Q-Skew | 80.3 | 57.7 | 39.7 |
| w/o Cyclic Sampling | 78.5 | 53.0 | 37.0 |
| w/o Skewness Information | 77.6 | 54.3 | 34.0 |
| w/o quantile weighting | 78.2 | 55.0 | 33.4 |

tile weighting. The experimental results in Tab. 6 show that each component in the method is effective.

# F PII Reconstruction: Ablation Study

### F.0.1 Hyperparameter Investigation

We investigate four hyperparameters influencing our combined loss-skewness score. We vary these values and plot the resulting ASR for LLaDA in Fig. 4. The first is the length of the token span X over which skewness is calculated. The second is the $\alpha$, controlling the strength of the loss and skewness signal in the final score. The third is the ratio of tokens masked in each masking batch used in the skewness calculation. The fourth is the number of masking batches used in the skewness calculation. Hyperparameter plots are organized from left to right in the order described.

For token span $X$, we do not observe a significant impact on ASR. For emails on LLaDA, however, there is a notable improvement between token spans of 20 and 30, indicating some potential variability. Overall, the relationship between token span and ASR remains weak. For $\alpha$, we observe a consistent negative relationship between increasing $\alpha$ and ASR. We attribute this to the importance of a strong cross-entropy loss signal in the weighted candidate score. This holds across both PII types. For the mask ratio $t$, we observe strong PII-specific effects. For phone numbers, higher mask ratios degrade ASR, whereas for emails, higher mask ratios lead to substantial improvements, particularly compared to the lower mask ratio of $0.1$. For the number of masking batches $B$, both excessively large or small values tend to reduce performance. When $B$ is too large, this may result from the distribution of loss differences becoming less skewed, thereby weakening the membership signal.

### F.0.2 Top-K Candidate Ranking

In Table 7, we show the top-k accuracy of different ranking methods where k is either 1, 3, or 5. Surprisingly for phone numbers, `Ranking` narrowly beats `Combine-0.1` for values of k larger than 1. Interestingly, the top-k accuracy advantage over `Ranking` in the case of emails improves drastically when k is equal to 5, with over 50% accuracy, the highest of any method and any value of k. Unsurprisingly, `Mask only PII` and `Mask only context` perform worse by comparison, similar to the top-1 accuracy.

Table 7: ASR@k (%) of different attack methods on LLaDA-8B-Base for phone numbers and email addresses. Higher is better. Best results per column are in bold.

| **Our Attacks** | **Signal** | **Phone** | | | **Email** | | |
|---|---|---|---|---|---|---|---|
| | | **@1** | **@3** | **@5** | **@1** | **@3** | **@5** |
| `Ranking` | CE | 30 | **42** | **44** | 20 | 36 | 42 |
| `Mask only PII` | Q-Skewness | 22 | 36 | 42 | 16 | 28 | 34 |
| `Mask only context` | Q-Skewness | 10 | 22 | 26 | 8 | 10 | 14 |
| `Combine-0.1` | Both | **34** | 38 | 42 | **24** | **44** | **52** |

# G PII Reconstruction: Setting $\alpha$

We assume that an attacker with partial knowledge of the fine-tuning dataset could reasonably approximate an effective $\alpha$ for the loss-skewness combined attack. In this scenario, we assume that the attacker has access to a small set of full, un-redacted records from the training dataset. In addition, we assume that they can interact with the target model the same way as when they target the unknown PII. To simulate this attacker, we select 16 records containing phone PII, and do a sweep of the $\alpha$ hyperparameter across `Combine` attacks. Critically, these 16 records form a hold-out set that does not overlap with the set of target records and PIIs used in the main evaluation. Fig. 5 shows that with only a small hold-out sample of target records, an $\alpha$ value of 0.1 leads to the highest ASR.

# H PII Reconstruction: Qualitative

Our `Combine-0.1` score achieves higher top-1 accuracy across multiple settings; however, skewness alone never outperforms `Ranking`. We attribute this to a mismatch between the sets of PII items correctly reconstructed by skewness and those recovered by `Ranking`.

For email targets reconstructed only by `Ranking`, we observe that true PIIs tend to be slightly shorter and closer to natural language. The fact that `Ranking` more easily extracts names of this form may point to the influence of pretraining priors—i.e., if similar structures were present during pretraining, the loss score may be biased toward

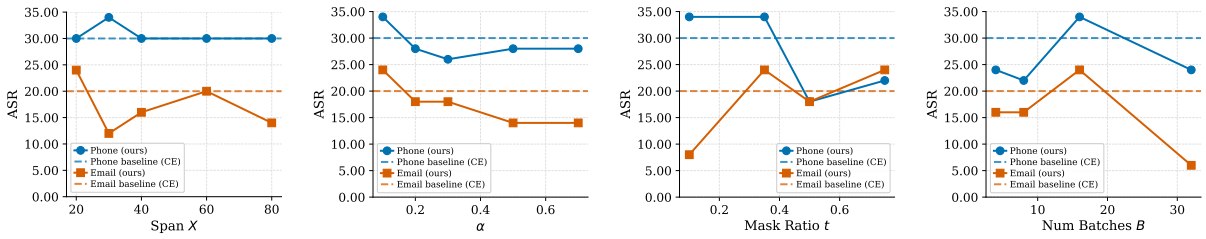


Figure 4: Ablation results of LLaDA for PII extraction. Dashed lines denote the CE-based (`Ranking`) method.

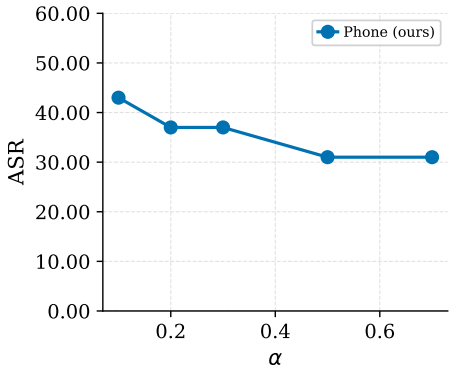


Figure 5: Impact of choosing $\alpha$ in (0, 1) on LLaDA for phone PII reconstruction, showing strong performance even with a small hold-out set. Higher is better.

**Extracted by Skewness**

```
w3c-sgml-wg
w3c-wai-ig
w3c-wai-er-ig
w3c-dist-auth
```

**Extracted by Ranking**

```
danny666
gclemm
jpalme
jdbryan
```

Figure 6: Examples of email IDs extracted exclusively by each candidate-scoring method.

**Extracted by Combine-0.1**

```
w3c-wai-ig
wai-report
ietf-discuss
w3c-wai-er-ig
w3c-dist-auth
w3c-wai-gl
w3c-wai-gl
w3c-sgml-wg
```

Figure 7: Examples of email IDs extracted under `Combine-0.1` but *not* `Ranking`.

them. We illustrate this difference in Fig. 6. For targets reconstructed only by skewness, we observe that true PIIs tend to be longer, aligning with observations from membership inference that longer member/non-member examples lead to improved skewness performance. These targets also tend to be more structured, which may help explain why the attack success rate is consistently higher for phone numbers, a more structured PII class.

For target emails that our `Combine-0.1` score successfully extracts but `Ranking` does not, we list the corresponding IDs in Fig. 7. It is clear that the influence of the skewness score enables the extraction of these more well-structured IDs, reinforcing the pattern observed in Fig. 6.

For phone targets reconstructed only by `Ranking`, we observe a larger candidate set size overall. This demonstrates the strength of loss as a metric for distinguishing correct in-context training data from a large non-training set. In contrast, targets reconstructed only by skewness tend to have slightly smaller candidate sets. The candidates within these sets also vary more at the token level. We show an example comparison in Fig. 8. Another characteristic of targets reconstructed only via skewness is their longer length, which aligns with prior work on memorization in autoregressive models, where increasing context length can increase the degree of memorization (Carlini et al., 2023).

**Extracted by Skewness**

```
617-983-4033
617-693-4603
617-693-4309
617-964-3073
617-693-4096
617-639-4036
617-693-4063
617-693-3034
617-693-4039
617-693-4030
617-693-4308
617-693-0433
617-693-6032
617-693-0133
617-693-4006
```

**Extracted by Ranking**

```
212 834-8114
212 648-1814
212 864-8811
212-484-1814
212 648-8114
212 684-1844
(212) 684-1844
(212) 684-0994
(212) 684-1414
(212) 844-1814
212 258-1814
212 658-1814
221-688-1487
122 684-9411
(212) 684-1841
212 254-1814
(212) 684-1814
212-648-1814
613 289-8124
212 683-2484
718 488-1813
212 864-1814
212 684-1417
212 683-1812
212 862-1814
212-684-1814
418-268-1814
212-864-9091
617 254-1818"
212 852-1814
```

Figure 8: Examples of candidate sets for targets extracted exclusively by a given scoring method.