Characterizing Memorization in Diffusion Language Models: Generalized Extraction and Sampling Effects

Anonymous Authors[1]

Abstract

Autoregressive language models (ARMs) have been shown to memorize and occasionally reproduce training data verbatim, raising concerns about privacy and copyright liability. Diffusion language models (DLMs) have recently emerged as a competitive alternative, yet their memorization behavior remains largely unexplored due to fundamental differences in generation dynamics. To address this gap, we present a systematic theoretical and empirical characterization of memorization in DLMs. We propose a generalized probabilistic extraction framework that unifies prefix-conditioned decoding and diffusion-based generation under arbitrary masking patterns and stochastic sampling trajectories. Theorem 4.3 establishes a monotonic relationship between sampling resolution and memorization: increasing resolution strictly increases the probability of exact training data extraction, implying that autoregressive decoding corresponds to a limiting case of diffusion-based generation by setting the sampling resolution maximal. Extensive experiments across model scales and sampling strategies validate our theoretical predictions. Under aligned prefix-conditioned evaluations, we further demonstrate that DLMs exhibit substantially lower memorization-based leakage of personally identifiable information (PII) compared to ARMs.

1. Introduction

Large Language Models (LLMs) have achieved remarkable success and have been integrated into a wide range of everyday applications. Despite this progress, the field has long been dominated by the decoder-only architecture under the ARMs paradigm (Radford et al., 2018; Brown et al., 2020; Radford et al., 2019). More recently, diffusion principles (Ho et al., 2020) have been successfully incorporated into masked DLMs (Lou et al., 2023; Shi et al., 2024; Ou et al., 2024), leading to the emergence of Large DLMs (Nie et al., 2025a;b). DLMs have demonstrated competitive scalability and performance with fewer training tokens. Since the training paradigm of these models fundamentally shifts from unidirectional next-token prediction to a bidirectional masking and reverse denoising process, their memorization behavior remains a critical yet unexplored frontier. Unlike ARMs that minimize the negative log-likelihood of sequential conditionals, DLMs optimizes a variational lower bound through a non-causal masking objective (Ou et al., 2024). This fundamental difference in how models internalize data necessitates a formal definition and systematic investigation of their memorization behavior, as it remains unclear how bidirectional denoising affects exposure to training data.

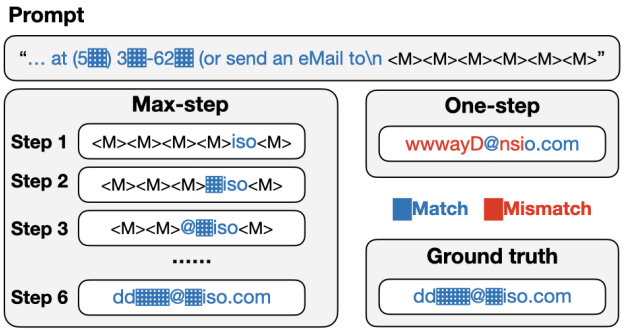


Figure 1. A randomly sampled verbatim PII memorization example from diffusion language model (LLaDA-8B). This case shows that finer-grained generation resolution (i.e., more steps) is more prone to verbatim recovery of training data. ⟨M⟩ denotes a mask token; blue highlights matched strings, and red highlights mismatches.

In the context of ARMs, a substantial body of work has shown that models can memorize and reproduce training samples verbatim (Carlini et al., 2021; 2022; Kiyomaru et al., 2024; Huang et al., 2024; Li et al., 2024; Zhang et al., 2025; Luo et al., 2025). Accordingly, studying memorization is a practical necessity, as it is closely tied to privacy leakage (Huang et al., 2022; Kim et al., 2023) and copyright infringement (Cooper et al., 2025; Ahmed et al., 2026).

Standard discoverable extraction procedures typically rely on prefix–suffix prompting (Carlini et al., 2022; Hayes et al., 2025). In contrast, DLMs differ in both conditioning and sampling, so we introduce an adapted evaluation framework

to measure discoverable extraction under their bidirectional denoising paradigm.

In this work, we present the first systematic investigation into the memorization behavior of diffusion language models (DLMs). Our contributions are three-fold:

- **A generalized formulation of memorization for diffusion language models.** We develop a generalized probabilistic formulation of discoverable memorization that applies to DLMs with arbitrary masking patterns and stochastic sampling trajectories, under which the prefix-conditioned (n, p)-discoverable extraction for ARMs arises as a special case.
- **Sampling resolution provably controls memorization.** We prove a monotonicity result showing that, under a natural recovery assumption, increasing the sampling resolution (i.e., recovering masked tokens in finer-grained denoising steps) monotonically increases the probability of exact memorization, and we demonstrate this effect empirically across sampling strategies and model scales.
- **Aligned analysis of PII memorization across model scales and architectures.** We align autoregressive and diffusion language models under the same prefix-conditioned PII completion setting and quantify memorization across multiple model scales and architectures, finding that autoregressive models exhibit higher PII memorization and leakage risk.

2. Related Work

2.1. Language Model Architecture

The Autoregressive Paradigm. The contemporary landscape of LLMs is predominantly defined by the ARMs paradigm, which factorizes the joint probability of sequences into a product of sequential conditionals (Radford et al., 2018; 2019; Brown et al., 2020). Leveraging the Transformer architecture, ARMs have demonstrated exceptional scalability and the emergence of complex capabilities (Wei et al., 2022) under empirical scaling laws (Kaplan et al., 2020; Hoffmann et al., 2022). Despite their success, ARMs exhibit inherent limitations directly tied to their unidirectional nature, most notably the "reversal curse"—a systemic failure to generalize bidirectional relationships (Berglund et al., 2023).

Masked Diffusion Paradigm. Diffusion models have seen success in the computer vision domain, typically formulated as a continuous generative task (Ho et al., 2020; Song et al., 2020). For the text-based task, one line of research explores continuous diffusion by embedding tokens into continuous spaces (Li et al., 2022; Gong et al., 2022; Zhang et al., 2023). In another discrete line, masked diffusion paradigm has emerged as a principled generative alternative for language modeling, leveraging a unique probabilistic framework that enables flexible bidirectional context modeling (Austin et al., 2021; Sahoo et al., 2024; Shi et al., 2024). This architecture is characterized by a forward process that either preserves tokens or replaces them with a unique mask token, and is trained with a reconstruction objective that approximates the reverse generative process. Foundational theoretical frameworks for Masked DLMs have established their connection to maximum likelihood estimation and variational lower bounds (Ou et al., 2024). Empirically, Masked DLMs have demonstrated promising results in smaller-scale settings (Lou et al., 2023).

Large Diffusion Language Models. Large DLMs are rapidly advancing diffusion-based language modeling into a regime that is competitive with mainstream ARMs at a comparable scale. Nie et al. (2025a) introduces the first formal scaling law for Masked DLMs, showing an ARM-like power-law improvement with scale. Building on this line of work, LLaDA (Nie et al., 2025b) follows the standard modern LLM training recipe with large-scale pretraining followed by supervised fine-tuning (SFT), and scales large DLMs to 8B parameters. It delivers broadly competitive performance against the same scale ARMs baseline on a suite of zero-/few-shot benchmarks, and shows in-context learning comparable to strong 8B-class models such as LLaMA3 (AI@Meta, 2024). Beyond accuracy, large DLMs' bidirectional conditioning improves robustness to directionality, outperforming GPT-4o (OpenAI, 2024) in the reversal settings completion task. Most recently, LLaDA2.0 (Bie et al., 2025) further scales diffusion language models to 16B/100B and shows that performance continues to improve at scale, achieving strong results on broad evaluations, while retaining the efficiency benefits of parallel generation.

2.2. Language Model Memorization

Prior work has established that LLMs can memorize and regurgitate portions of their training data (Carlini et al., 2021; 2022; Huang et al., 2024; Li et al., 2024; Zhang et al., 2025). Early evidence shows that black-box adversaries can extract hundreds of verbatim training sequences, including rare and sensitive content, from generative models (Carlini et al., 2021). Subsequent work quantifies how extractable memorization, showing consistent increases with both model capacity and training-data repetition (Carlini et al., 2022).

Beyond a strong verbatim-controlled setting, a growing body of research investigates memorization risks in realistic scenarios involving personally identifiable information (PII). Studies on PII leakage indicate that LLMs may retain sensitive personal information, which can be reconstructed through association-based attacks (Huang et al., 2022; Kim

et al., 2023). From a complementary angle, Lukas et al. (2023) formalize and evaluate PII leakage under black-box access, showing that semantic associations can enable adversaries to recover private attributes via indirect prompting. More recently, Luo et al. (2026) argue that memorization estimates may be inflated when prompts contain implicit cues, leading to an overestimation of a model's true memorization behavior.

Memorization and data extraction have also been examined in the context of production and aligned language models. Nasr et al. (2025) show that alignment safeguards can be circumvented at scale, enabling the extraction of memorized training samples from deployed proprietary systems. In parallel, analyses of open-weight models reveal that certain copyrighted books are memorized to a much greater extent than previously understood, in some cases allowing near-complete reconstruction via probabilistic extraction (Cooper et al., 2025). Extending these observations to deployed systems, Ahmed et al. (2026) demonstrates that, under specific prompting and jailbreak conditions, substantial portions of in-copyright books can be extracted near-verbatim, highlighting persistent memorization risks despite system-level safeguards.

Most prior evaluations of memorization rely on greedy or single-query extraction settings, which may not reflect realistic deployment conditions. To address this limitation, Hayes et al. (2025) introduces probabilistic extractable memorization metrics that measure the expected query effort required to recover a memorized sequence under non-deterministic decoding and repeated querying, providing a more deployment-relevant assessment of memorization risk.

Finally, memorization has also been studied in multilingual settings. Luo et al. (2025) extend memorization measurements to multilingual LLMs and identify a long-tail (Li et al., 2025) effect mediated by language similarity: within similar language groups, lower-resource languages tend to exhibit higher levels of memorization.

3. Preliminaries

In this section, we briefly review DLMs and ARMs architectures, introduce their commonly used sampling paradigms, and finally present the existing (n, p)-discoverable extraction framework defined for ARMs.

3.1. Masked Diffusion Model

We review the framework of the masked diffusion paradigm (Ou et al., 2024; Nie et al., 2025b), which serves as the fundamental architecture for Large DLMs.

Forward Process. Let $\mathcal{V} = \{0, 1, \ldots, K-1\}$ be the vocabulary set. Given a sequence $\boldsymbol{z}_0 \in \mathcal{V}^L$ of length L, the forward process at noise level $t \in [0, 1]$ gradually replaces tokens. The transition probability is defined as

$$q_{t|0}(\boldsymbol{z}_t|\boldsymbol{z}_0) = \prod_{l=1}^{L} q_{t|0}(z_t^l|z_0^l), \tag{1}$$

where $q_{t|0}(z_t^l|z_0^l) = \begin{cases} \alpha_t, & z_t^l = z_0^l \\ 1-\alpha_t, & z_t^l = m \end{cases}$. $\boldsymbol{z}_t$ denotes the noisy data at time t, z_t^l denote the l-th element of $\boldsymbol{z}_t$, m denotes a special [MASK] token (Devlin et al., 2019) and $\alpha_t = 1 - t$.

Reverse Process. The reverse process will recover the original tokens $\boldsymbol{z}_0$ from a fully masked sequence $\boldsymbol{z}_1$. For any $0 \le s < t \le 1$, the transition $q_{s|t}(\boldsymbol{z}_s|\boldsymbol{z}_t)$ recovers a portion of the masked tokens, which can be formulated as

$$q_{s|t}(z_s^l|\boldsymbol{z}_t) = \begin{cases} 1, & z_t^l \neq m, z_s^l = z_t^l \\ \frac{s}{t}, & z_t^l = m, z_s^l = m \\ \frac{t-s}{t} q_{0|t}(z_s^l|\boldsymbol{z}_t), & z_t^l = m, z_s^l \neq m \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

Training Objective. A model $p_\theta(z_0^l|\boldsymbol{z}_t)$ is trained to approximate the true data distribution by minimizing the following negative log-likelihood upper bound as

$$\mathcal{L} = \int_0^1 \frac{1}{t} \mathbb{E}_{q(\boldsymbol{z}_t|\boldsymbol{z}_0)} \left[\sum_{l:z_t^l=m} -\log p_\theta(z_0^l|\boldsymbol{z}_t) \right] dt. \tag{3}$$

Sampling Strategy. Sampling iteratively updates the currently masked positions while keeping unmasked tokens unchanged. A *linearly spaced* time grid $\{t_i\}_{i=0}^N$ is used, with $t_0 = 1$ and $t_{i+1} < t_i$, N is the sampling resolution. The linear schedule determines the fraction of tokens to be recovered at each step $t \to s$ by $1 - \frac{s}{t}$. At step $t \to s$, for each masked position l ($z_t^l = m$), the model predicts a token $\hat{z}_0^l = \arg\max_{v \in \mathcal{V}} p_\theta(v \mid \boldsymbol{z}_t)$ with confidence $p_\theta(\hat{z}_0^l \mid \boldsymbol{z}_t)$. *Greedy sampling* selects the masked positions with the highest confidence and recovers $k = \lfloor |\mathcal{M}|(1-\frac{s}{t}) \rfloor$ tokens, while *random sampling* recovers the same number of tokens by uniformly sampling positions from $\mathcal{M}$.

3.2. Autoregressive Model

Most existing LLMs follow the ARM formulation (Radford et al., 2018). Given a sequence $\boldsymbol{z} \in \mathcal{V}^L$, an ARMs factorizes

$$p_\theta(\boldsymbol{z}) = \prod_{l=1}^{L} p_\theta\left(z^l \mid \boldsymbol{z}^{1:l-1}\right), \tag{4}$$

where $\boldsymbol{z}^{1:l-1} = [z^1, \ldots, z^{l-1}]$ is the prefix; we use this as the canonical baseline paradigm throughout.

3.3. Sampling Paradigm

We then briefly introduce the sampling paradigms that are commonly used in ARMs and DLMs. Among these paradigms, temperature and top-k sampling are widely adopted in ARMs (Fan et al., 2018), while DLMs use Gumbel noise perturbation (Nie et al., 2025a).

Temperature Sampling. Given a prefix $\boldsymbol{z}^{1:l-1}$, let ℓ_v denote the logit for token $v \in \mathcal{V}$. Temperature sampling parameterizes the predictive categorical distribution as

$$\Pr(z^l = v \mid \boldsymbol{z}^{1:l-1}) = \frac{\exp(\ell_v/T)}{\sum_{u\in\mathcal{V}} \exp(\ell_u/T)}, \quad T > 0.$$

Here, T serves as a scaling factor that controls the sharpness of the distribution: a lower T concentrates the probability mass on high-logit tokens, thereby reducing stochasticity, while a higher T flattens the distribution, promoting diversity in generation. The standard softmax distribution is recovered as a special case when $T = 1$.

Top-k sampling. Let $p(v) = \mathrm{softmax}(\ell_v)$ and let $\mathcal{K}$ be the set of k tokens with largest $p(v)$. Top-k sampling restricts the distribution to $\mathcal{K}$ and renormalizes it as

$$Pr(z^l = v \mid \boldsymbol{z}^{1:l-1}) \;=\; \begin{cases} \frac{p(v)}{\sum_{u\in\mathcal{K}} p(u)}, & v \in \mathcal{K}, \\ 0, & \text{otherwise.} \end{cases}$$

This procedure collapses to deterministic greedy decoding when $k = 1$, where only the most probable token is selected.

Gumbel noise perturbation. In diffusion models, a commonly used sampling paradigm is Gumbel noise perturbation. To perform stochastic token sampling, we draw i.i.d. random variables $U_v \sim \mathrm{Uniform}(0, 1)$ for each vocabulary item $v \in \mathcal{V}$, and construct the perturbed scores as

$$S_v \;=\; \frac{\exp(\ell_v)}{\big(-\log U_v\big)^{T_{\text{Gumbel}}}}, \tag{5}$$

where T_{Gumbel} is a temperature parameter that controls the level of stochasticity. The sampled token is then selected as the most confident one, i.e.,

$$\arg\max_{v\in\mathcal{V}} S_v.$$

When $T_{\text{Gumbel}} = 0$, $S_v = \exp(\ell_v)$. In this case, the procedure reduces to greedy decoding, which can be naturally paired with deterministic sampling in diffusion models.

3.4. (n, p)-discoverable Extraction Framework

We review two definitions of memorization for ARMs as introduced by Hayes et al. (2025).

Let $[L] = \{1, 2, \ldots, L\}$ be the set of all token indices and $\boldsymbol{z} = [z_1, \ldots, z_L] \in \mathcal{V}^L$ be a training example.

Definition 3.1 ((n, p)-discoverable Extraction)**.** A sequence is (n, p)-discoverably extractable if, given a prefix, the model generates the exact suffix within n independent queries with probability at least p.

Let $p_{\boldsymbol{z}}$ denote the probability of generating the exact suffix in a single trial. Consequently, the probability of failing to recover the correct suffix in n independent trials is $(1 - p_{\boldsymbol{z}})^n$. In order to ensure that the probability of at least one successful extraction is at least p, we require $1-(1-p_{\boldsymbol{z}})^n \geq p$. Solving for n, we obtain the estimation for the required number of queries as

$$n \geq \frac{\log(1-p)}{\log(1-p_{\boldsymbol{z}})}. \tag{6}$$

This relationship implies that the extractability parameters (n, p) can be determined once $p_{\boldsymbol{z}}$ is known. In practice, $p_{\boldsymbol{z}}$ is computed as the cumulative product of the conditional probabilities associated with recovering each token at its corresponding sampling step.

Beside the exact extraction, Hayes et al. (2025) also proposed a relaxed version for approximate recovery.

Definition 3.2 ((ϵ, n, p)-discoverable Extraction)**.** A suffix of an example is (ϵ, n, p)-discoverably extractable if the probability that at least one of n independent queries results in a generated suffix with a distance $\leq \epsilon$ from the ground truth is at least p.

4. A Generalized Discoverable Memorization Framework for Diffusion Language Models

Existing definitions of discoverable memorization (Hayes et al., 2025) are inherently tied to the autoregressive paradigm, where memorization is evaluated through a fixed prefix–suffix structure under left-to-right decoding. While this formulation closely mirrors the unidirectional generation process of autoregressive language models, it does not naturally extend to diffusion language models, which recover masked tokens through stochastic, non-sequential denoising trajectories.

To address this mismatch, we develop a generalized formulation of discoverable memorization that removes the prefix-based decoding assumption and explicitly accounts for arbitrary masking patterns and diffusion sampling trajectories. Moreover, within our proposed framework, prefix-conditioned extraction can be recovered as a special case (see Proposition 4.4), allowing memorization in diffusion language models to be analyzed under their native generation dynamics.

4.1. Generalized Exact Memorization in DLMs

We first focus on the case of exact memorization.

4.1.1. Definition and Approximation

Definition 4.1 (Generalized (n, p)-discoverable Extraction)**.** A sequence $\boldsymbol{z}$ is (n, p)-discoverably extractable under a mask $\mathcal{M} \subset \{1, \ldots, L\}$ if, given the observed tokens $\boldsymbol{z}_{\bar{\mathcal{M}}}$, the model recovers the exact original tokens at all masked positions $\mathcal{M}$ within n independent queries with probability at least p.

Formally, let $\hat{\boldsymbol{z}}_{\mathcal{M}}^{(i)}$ be the output of the i-th independent sampling. The sequence $\boldsymbol{z}$ is (n, p)-extractable if:

$$\Pr\left(\exists i \in \{1, \ldots, n\} : \hat{\boldsymbol{z}}_{\mathcal{M}}^{(i)} = \boldsymbol{z}_{\mathcal{M}} \mid \boldsymbol{z}_{\bar{\mathcal{M}}}\right) \geq p, \quad (7)$$

where $\bar{\mathcal{M}} = [L] \backslash \mathcal{M}$ denotes the complement set of indices (the observed positions).

Following Hayes et al. (2025), the required number of independent queries n can be estimated given a target extraction probability p and an approximation of the single-trial recovery probability $p_{\boldsymbol{z}}$. In the context of diffusion-based LLMs, the approximation of $p_{\boldsymbol{z}}$ is a bit different and it is inherently dependent on the sampling resolution N.

For a sampling process with N steps in the i-th independent query, we partition $\mathcal{M}$ into N disjoint chunks $\{\Delta_1^{(i)}, \Delta_2^{(i)}, \ldots, \Delta_N^{(i)}\}$ such that $\Delta_k^{(i)}$ represents the set of token indices recovered at step k. Let $\mathcal{U}_k^{(i)}$ be the set of observed indices after step k, with $\mathcal{U}_0^{(i)} = [L] \backslash \mathcal{M}$ and $\mathcal{U}_k^{(i)} = \mathcal{U}_{k-1}^{(i)} \cup \Delta_k^{(i)}$. Because tokens are recovered independently at each step (given the unmasked context), the total probability $p_{\boldsymbol{z}}^{(i)}$ is defined by the multiplicative product of their respective conditional probabilities

$$\begin{aligned} p_{\boldsymbol{z}}^{(i)} &= \Pr(\hat{\boldsymbol{z}}_{\mathcal{M}} = \boldsymbol{z}_{\mathcal{M}} \mid \boldsymbol{z}_{\bar{\mathcal{M}}}, N) \\ &= \prod_{k=1}^{N} \prod_{\pi \in \Delta_k^{(i)}} \Pr(\hat{z}_\pi = z_\pi \mid \boldsymbol{z}_{\mathcal{U}_{k-1}^{(i)}}). \end{aligned} \quad (8)$$

Unlike ARMs, where the decoding order is typically fixed, diffusion-based models exhibit stochasticity in the recovery path (i.e., the sequence of token subsets selected at each step). Consequently, the single-trial probability $p_{\boldsymbol{z}}$ cannot be reliably estimated from a single query, as is possible in the deterministic setting of autoregressive. To account for this randomness, we estimate the expected $\hat{p}_{\boldsymbol{z}}$ by averaging results over a set of R trials, i.e. $\hat{p}_{\boldsymbol{z}} = \frac{1}{R}\sum_{i=1}^{R} p_{\boldsymbol{z}}^{(i)}$. The validation results are in Sec 6.1.

4.1.2. The Impact of Sampling Resolution

Assumption 4.2 (Monotonicity of Recovery Probability)**.** We assume that for any inference step, the probability of correctly recovering a subset of masked tokens increases as the set of observed tokens expands.

Formally, for any two observation index sets $\mathcal{U}, \mathcal{W} \subseteq [L]$ such that $\mathcal{U} \subset \mathcal{W}$, and for any target mask set $\mathcal{M} \subseteq [L] \backslash \mathcal{W}$, we have

$$\Pr(\hat{\boldsymbol{z}}_{\mathcal{M}} = \boldsymbol{z}_{\mathcal{M}} \mid \boldsymbol{z}_{\mathcal{W}}) \geq \Pr(\hat{\boldsymbol{z}}_{\mathcal{M}} = \boldsymbol{z}_{\mathcal{M}} \mid \boldsymbol{z}_{\mathcal{U}}) \quad (9)$$

where $\boldsymbol{z}_{\mathcal{U}}$ and $\boldsymbol{z}_{\mathcal{W}}$ denotes the sub-vector of tokens at indices in $\mathcal{U}$ and $\mathcal{W}$ respectively, and $\hat{\boldsymbol{z}}_{\mathcal{M}}$ is the reconstruction of tokens at indices $\mathcal{M}$.

This assumption formalizes the intuitive property that additional *correct context* should not reduce recovery likelihood, which is also validated empirically in Sec 6.2. It is important to note that we specifically consider context consisting of correctly recovered tokens, rather than potentially erroneous predictions. This focus is consistent with our objective of analyzing *exact memorization*, where the recovery of the entire sequence is only successful if every intermediate step along the sampling trajectory remains faithful to the training instance.

Hereby we have the following Theorem.

Theorem 4.3 (Impact of Sampling Resolution on Memorization)**.** *Under Assumption 4.2, for diffusion-based LLMs, the probability of generating a sequence that exactly matches a training instance generally increases with the number of sampling steps N.*

In certain cases, this relationship is theoretically monotonic under a fixed recovery sequence. Let $\sigma = (\pi_1, \pi_2, \ldots, \pi_M)$ be a fixed permutation of indices in $\mathcal{M}$ representing the recovery order. For any N, we partition σ into N disjoint and ordered chunks $\mathcal{P}_N = \{\Delta_1, \Delta_2, \ldots, \Delta_N\}$. If $N_1 > N_2$ and the partition $\mathcal{P}_{N_1}$ is a refinement of $\mathcal{P}_{N_2}$ (i.e., each chunk in $\mathcal{P}_{N_2}$ is the union of one or more consecutive chunks in $\mathcal{P}_{N_1}$), then:

$$Pr(\hat{\boldsymbol{z}}_{\mathcal{M}} = \boldsymbol{z}_{\mathcal{M}} \mid \boldsymbol{z}_{\bar{\mathcal{M}}}, N_1) \geq Pr(\hat{\boldsymbol{z}}_{\mathcal{M}} = \boldsymbol{z}_{\mathcal{M}} \mid \boldsymbol{z}_{\bar{\mathcal{M}}}, N_2). \quad (10)$$

Proof. See Appendix A. □

This theorem establishes that increasing the sampling resolution consistently enhances the probability of exact memorization. Since the number of sampling steps N is upper-bounded by the total number of masked tokens $|\mathcal{M}|$, the most granular refinement occurs when $N = |\mathcal{M}|$. In this limiting case, the diffusion model recovers tokens in one-by-one fashion, mirroring the fundamental mechanism of ARMs. This leads to the following proposition.

Proposition 4.4. *Conceptually, ARMs can be viewed as a special case of a diffusion-based LLM where the sampling resolution N reaches its maximum value ($N = |\mathcal{M}|$). In this regime, the model recovers only one token per step in a fixed sequential order. Following the logic of Theorem 4.3,*

as the sampling process transitions from parallel refinement to fully sequential generation, the increasing availability of step-wise context suggests that the ARM-style decoding represents an upper-limit behavior for exact memorization probability.

4.2. Generalized Relaxed Memorization in DLMs

Definition 4.5 (Generalized (ϵ, n, p)-discoverable Extraction)**.** We generalize the notion of discoverable extraction to arbitrary masking patterns. An example $\boldsymbol{z}$ is (ϵ, n, p)-discoverably extractable under a mask $\mathcal{M}$ if, given the observed tokens $\boldsymbol{z}_{\bar{\mathcal{M}}}$, the model generates a reconstructed sequence $\hat{\boldsymbol{z}}_{\mathcal{M}}$ such that the distance to the original tokens $d(\hat{\boldsymbol{z}}_{\mathcal{M}}, \boldsymbol{z}_{\mathcal{M}}) \leq \epsilon$ holds for at least one of n independent queries with probability at least p, i.e.,

$$\Pr\left(\exists i \in \{1, \ldots, n\} : d(\hat{\boldsymbol{z}}^{(i)}_{\mathcal{M}}, \boldsymbol{z}_{\mathcal{M}}) \leq \epsilon \mid \boldsymbol{z}_{\bar{\mathcal{M}}}\right) \geq p, \quad (11)$$

where $d(\cdot, \cdot)$ is a distance metric. When $\epsilon = 0$, this definition reduces to the case of exact extraction.

In Hayes et al. (2025), the error-tolerant extraction probability $p_{\boldsymbol{z},\epsilon}$ can be empirically estimated by calculating the proportion of successful reconstructions (the distance is within the threshold ϵ) across R independent trials, i.e.,

$$\hat{p}_{\boldsymbol{z},\epsilon} = \frac{1}{R} \sum_{i=1}^{R} \mathbb{I}\left(d(\hat{\boldsymbol{z}}^{(i)}_{\mathcal{M}}, \boldsymbol{z}_{\mathcal{M}}) \leq \epsilon\right). \quad (12)$$

This estimation method naturally extends to the generalized definition. Specifically, when the error tolerance is set to $\epsilon = 0$, the estimator $\hat{p}_{\boldsymbol{z},0}$ provides a direct empirical approximation for the exact memorization probability $\Pr(\hat{\boldsymbol{z}}_{\mathcal{M}} = \boldsymbol{z}_{\mathcal{M}} \mid \boldsymbol{z}_{\bar{\mathcal{M}}}, N)$ with a large R.

In addition, we investigate the statistical behavior of the error distribution under large-scale trials. We refer the reader to Appendix B for the related experimental analysis.

5. Experimental Setup

5.1. Model

Foundation model (DLM and ARM). To isolate scale and architectural effects on memorization, we pretrain our DLMs and an ARM baseline under an identical setup. Specifically, we reproduce the training recipe of Nie et al. (2025a). under a fixed compute budget of 10^{21} FLOPs, and train three DLMs with 170M, 690M, and 1.1B parameters on SlimPajama. For a one-to-one paradigm comparison, we also train a 1.1B ARM baseline with the *exact same* setup (and tokenizer) as the 1.1B DLM. It ensures that any differences in memorization behavior are attributable primarily to the scale and generative architecture, rather than to data, tokenization, or training procedure.

Finetuning (LLaDA-8B alignment). Since LLaDA-8B's original pretraining data are not publicly available, we evaluate memorization after finetuning on the Enron email dataset. To ensure comparability across model families, we use the same finetuning protocol for LLaDA-8B and all SlimPajama-pretrained DLMs and ARMs, continuing training for one epoch on Enron with each model's end-of-pretraining learning rate. This shared adaptation setting ensures memorization differences are not driven by mismatched finetuning procedures.

5.2. Dataset Collection

Validating Data Collection. For the validation experiments (Sec 6.1, 6.2), we draw evaluation samples from the pretrain dataset by uniformly sampling **100-token windows** at random from the corpus. For each window, we apply random masking with mask ratio $r \in \{0.20, 0.25, 0.30\}$, and evaluate reconstruction/extraction under these different masking levels.

PII Collection. For the alignment experiment (Sec. 6.3), we construct a PII-focused evaluation set from the Enron email dataset, extracting emails and phone numbers via regular expressions, following (Kim et al., 2023) procedure (see Appendix F for details). We sample **3,000** email-containing texts and **3,000** phone-number-containing texts, and for each PII occurrence use the 100 preceding tokens as the prefix to test **verbatim memorization** under prefix-conditioned generation. To avoid spurious pattern-based cues, email prompts are further filtered to contain no additional email addresses.

6. Analysis and Results

This section empirically evaluates our memorization framework. We first validate that generalized discoverable extraction is measurable in DLMs despite stochastic masking and iterative denoising, and then confirm that finer sampling resolution increases verbatim memorization. We further align DLMs and ARMs in a PII completion setting to compare memorization-based leakage under identical prefix-conditioned reconstruction, and finally show the metric reflects memorization rather than generalization via train vs. disjoint same-domain test comparisons.

6.1. Rationality of generalized discoverable extraction in DLMs

To validate that our proposed notion of memorization is operational and measurable in practice (Definition 4.1), we quantify the memorization probability at the level of individual examples.

Since DLMs decoding depends on stochastic masking and

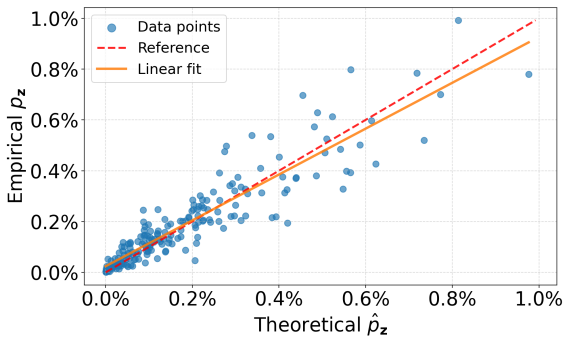


Figure 2. Empirical vs. theoretical memorization probability measured by DLM-1.1B on 200 SlimPajama examples. Empirical $p_{\boldsymbol{z}}$ is estimated from 100,000 random-decoding generations; theoretical $\hat{p}_{\boldsymbol{z}}$ is estimated from 1,024 mask-patterns at a fixed mask ratio ($[0.2, 0.3]$).

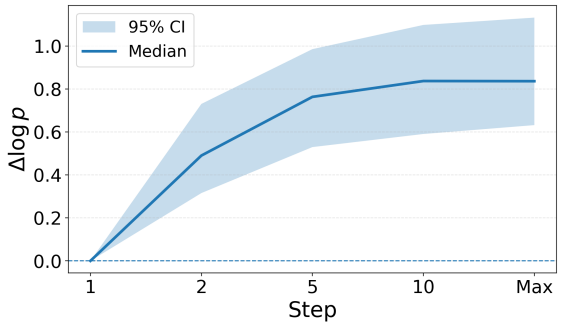


Figure 3. Empirical log-probability ratio across decoding steps (DLM-1.1B). We sample 200 mask-patterns and, for each, run 10,000 random decoding generations to estimate success probabilities at steps $\{1, 2, 5, 10, \text{per-token}\}$. We plot the median log-probability difference to one-step (baseline), with 95% bootstrap Confidence Intervals (CIs) across generation-trajectories.

iterative denoising, we fix the mask ratio and sample multiple mask-patterns (i.e., different masked positions); we then aggregate the one-step recovery outcomes across these mask-patterns to estimate the probability of recovering the target content.

Figure 2 compares the theoretical $\hat{p}_{\boldsymbol{z}}$ with the empirical $p_{\boldsymbol{z}}$. For computational feasibility, we evaluate on 200 examples, uniformly sampled from those with a preliminary estimated recovery probability above 10^{-3} (screened using 1,024 mask-patterns at a fixed mask ratio in $[0.2, 0.3]$). For these selected examples, we then re-estimate $\hat{p}_{\boldsymbol{z}}$ from scratch using a fresh, independent set of 1,024 mask-patterns, and compute $p_{\boldsymbol{z}}$ from 100,000 randomly decoded generations. The two distributions exhibit a high level of agreement throughout the range, suggesting that (8) reliably approximates the underlying recovery probability with relatively few mask-patterns. Additional results under different numbers of mask-patterns are provided in Appendix C.

6.2. Impact of Sampling Resolution on Verbatim Memorization

In this section, we provide empirical evidence supporting Assumption 4.2. Building on Sec 6.1, we select 200 mask-patterns from the previous section whose theoretical one-step recovery rate exceeds 0.01, a subset chosen for computational feasibility. For each selected trajectory, we run 10,000 unconstrained generation trials under five generation resolutions: 1-step ($N = 1$), 2-step ($N = 2$), 5-step ($N = 5$), 10-step ($N = 10$), and Max ($N = |\mathcal{M}|$, where the number of masked tokens is used as the generation step.)

Figure 3 summarizes these *empirical* exact-recovery success rates by plotting, for each resolution, the cross-trajectory median of the log success-rate ratio relative to the one-step baseline. We report $\Delta \log p := \text{median}\big(\log(\hat{p}_N/\hat{p}_1)\big)$,

Model	Step	Email		Phone	
		$p = 50\%$	$p = 99\%$	$p = 50\%$	$p = 99\%$
DLM-1.1B	One	0	0	0	0
	Max	16	7	0	0
ARM-1.1B	N/A	213	139	5	3
LLaDA-8B	One	9	8	7	3
	Max	179	122	23	17

Table 1. Generalized (n, p)-discoverable Extraction. Number of memorized samples at target probabilities p under a query budget of $n = 10{,}000$ on 3,000 completion prompts, for different PII types.

where $\hat{p}_N$ is the empirical exact-recovery hit rate at resolution N and the median is taken over the mask-patterns samples. The resulting curve exhibits a clear upward trend as the number of generation steps increases, suggesting an approximately monotonic relationship between generation resolution and exact-recovery success. This implies that allocating more denoising steps consistently increases the likelihood of reproducing the target content exactly, providing empirical support for Assumption 4.2. Additional memorization examples are provided in Appendix E.

6.3. Empirical Analysis of PII Leakage via Aligned ARM–LDM Measurement

Under the aligned PII prefix-suffix verbatim completion task, DLMs exhibit substantially lower (n, p)-discoverable PII memorization than scale-matched ARMs. Table 1 reports the number of memorized email and phone samples at target probabilities $p \in \{0.5, 0.99\}$ under a query budget of $n = 10{,}000$ over 3,000 prompts. This gap may partly reflect differences between training objectives: random masking in DLMs disrupts learning to preserve long contiguous context,

while the ARMs training paradigm naturally matches the sequential prefix-suffix decoding.

At a larger scale, the LLaDA-8B model exhibits a clear increase in memorization when moving from single-step reconstruction to per-token reconstruction, providing practical task evidence for Assumption 4.2 that finer-grained denoising increases exact recovery probability. However, even under per-token reconstruction, the 8B diffusion model is only comparable to the 1.1B ARM. Importantly, this comparison should be interpreted in light of training cost: LLaDA-8B is trained with roughly two orders of magnitude more FLOPs and about $150\times$ more training token exposure than the 1.1B ARM baseline (15.6B vs. 2300B) (Nie et al., 2025b). Overall, in this aligned prefix-conditioned PII setting, DLMs exhibit lower memorization behavior.

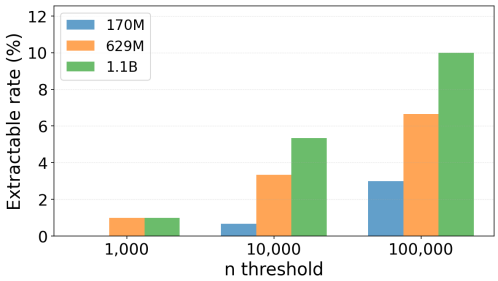


Figure 4. Per-token-step email hit rates under different model scales (170M/629M/1.1B). Each bar reports the Extractable rate over 3,000 samples, measured as whether an email can be recovered within a querying budget n, at fixed $p = 0.5$.

Figure 4 reports (n, p)-discoverable email memorization across aligned DLMs scales, with extractable rates increasing with model size. For each model, gains from larger query budgets are sublinear, suggesting most practically discoverable memorization is recovered with modest n and further increases yield little additional privacy leakage. Detailed PII leakage for other DLMs scales is provided in Appendix D.

6.4. Validating Memorization Beyond Generalization

A key question is whether our extraction metric reflects true training-data memorization rather than generalization-based reconstruction, especially since DLMs-1.1B shows a low level of verbatim memorization.

To validate the metric, we follow prior work (Hayes et al., 2025) and use the TREC 2007 Spam email dataset(Bratko et al., 2006). Like the Enron dataset, it consists of email text and therefore lies in the same broad domain, while remaining disjoint from the training data. From each corpus, we sample $10,000$ contents of length 100 tokens, and evaluate extraction using a mask ratio of 0.25 with 512 sampling queries per example.

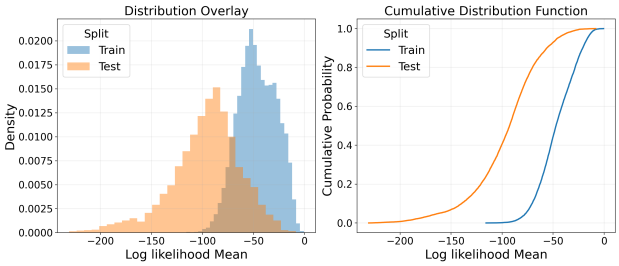


Figure 5. Validating memorization of training data. We compare the reconstruction likelihood from the Enron (train) against the unseen from the TREC 2007 Spam (test), which is from the same domain.

Figure 5 compares training and test data under our extraction metric using a distribution overlay and the corresponding cumulative distribution function. In both views, the training data distribution is shifted toward higher log-likelihood values than the unseen test data, indicating uniformly higher reconstruction likelihood for training data. This clear separation suggests that the metric captures memorization rather than reconstruction driven by generalization. Accordingly, even though DLM-1.1B shows low overall verbatim memorization, it exhibits a systematic advantage on Enron, implying that the extracted content reflects memorized training examples rather than coincidental matches. The consistent gap between train and test further supports the robustness of the previous section's findings. Additional results for the other model scales are provided in Appendix D, Figure 9.

7. Conclusion and Future Work

This paper introduces a generalized probabilistic extraction framework for measuring memorization in diffusion language models, allowing discoverable extraction beyond prefix-based decoding to arbitrary masking patterns and stochastic diffusion sampling trajectories. It further establishes a principled link between sampling design and leakage risk by proving that finer-grained sampling resolution (more denoising steps) increases the probability of memorization, with autoregressive decoding recoverable as a special case of maximal resolution. Empirically, across model scales and sampling strategies, and under an aligned prefix-conditioned PII completion task, DLMs exhibit lower memorization than ARM counterparts. Building on these results, future work will measure how post-training (SFT and preference optimization) changes memorization in diffusion language models using evaluation protocols that preserve aligned behavior and match real-deployment prompting. In particular, it is important to test whether post-training drives DLMs toward more prefix-suffix-style generation, and whether this shift increases verbatim memorization and extractability.

Impact Statements

This paper presents a theoretical and empirical analysis of memorization in diffusion-based Large Language Models. Our findings contribute to the broader field of trustworthy AI by elucidating how sampling strategies influence the risk of training data leakage. While our analysis reveals potential vulnerabilities in how these models retain training data, this understanding is a prerequisite for building robust defenses and ensuring the responsible deployment of generative models in privacy-sensitive domains. There are many potential societal consequences of our work, and we believe our contribution to data transparency and model auditing is the most significant one.

References

Ahmed, A., Cooper, A. F., Koyejo, S., and Liang, P. Extracting books from production language models. *arXiv preprint arXiv:2601.02671*, 2026.

AI@Meta. Llama 3 model card. 2024. URL https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

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

Berglund, L., Tong, M., Kaufmann, M., Balesni, M., Stickland, A. C., Korbak, T., and Evans, O. The reversal curse: Llms trained on" a is b" fail to learn" b is a". *arXiv preprint arXiv:2309.12288*, 2023.

Bie, T., Cao, M., Chen, K., Du, L., Gong, M., Gong, Z., Gu, Y., Hu, J., Huang, Z., Lan, Z., et al. Llada2.0: Scaling up diffusion language models to 100b. *arXiv preprint arXiv:2512.15745*, 2025.

Bratko, A., Filipic, B., Cormack, G. V., Lynam, T. R., and Zupan, B. Spam filtering using statistical data compression models. *The Journal of Machine Learning Research*, 7:2673–2698, 2006.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., et al. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., and Zhang, C. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*, 2022.

Cooper, A. F., Gokaslan, A., Ahmed, A., Cyphert, A. B., De Sa, C., Lemley, M. A., Ho, D. E., and Liang, P. Extracting memorized pieces of (copyrighted) books from open-weight language models. *arXiv preprint arXiv:2505.12546*, 2025.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. 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)*, pp. 4171–4186, 2019.

Fan, A., Lewis, M., and Dauphin, Y. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*, 2018.

Gong, S., Li, M., Feng, J., Wu, Z., and Kong, L. Diffuseq: Sequence to sequence text generation with diffusion models. *arXiv preprint arXiv:2210.08933*, 2022.

Hayes, J., Swanberg, M., Chaudhari, H., Yona, I., Shumailov, I., Nasr, M., Choquette-Choo, C. A., Lee, K., and Cooper, A. F. Measuring memorization in language models via probabilistic extraction. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 9266–9291, 2025.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Huang, J., Shao, H., and Chang, K. C.-C. Are large pre-trained language models leaking your personal information? *arXiv preprint arXiv:2205.12628*, 2022.

Huang, J., Yang, D., and Potts, C. Demystifying verbatim memorization in large language models. *arXiv preprint arXiv:2407.17817*, 2024.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Kim, S., Yun, S., Lee, H., Gubri, M., Yoon, S., and Oh, S. J. Propile: Probing privacy leakage in large language models. *Advances in Neural Information Processing Systems*, 36:20750–20762, 2023.

Kiyomaru, H., Sugiura, I., Kawahara, D., and Kurohashi, S. A comprehensive analysis of memorization in large language models. In Mahamood, S., Minh, N. L., and Ippolito, D. (eds.), *Proceedings of the 17th International Natural Language Generation Conference*, pp. 584–596, Tokyo, Japan, September 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.inlg-main.45.

Li, B., Zhao, Q., and Wen, L. Rome: Memorization insights from text, logits and representation. *arXiv preprint arXiv:2403.00510*, 2024.

Li, Q., Luo, X., Chen, Y., and Bjerva, J. Trustworthy machine learning via memorization and the granular long-tail: A survey on interactions, tradeoffs, and beyond. *arXiv preprint arXiv:2503.07501*, 2025.

Li, X., Thickstun, J., Gulrajani, I., Liang, P. S., and Hashimoto, T. B. Diffusion-lm improves controllable text generation. *Advances in neural information processing systems*, 35:4328–4343, 2022.

Lou, A., Meng, C., and Ermon, S. Discrete diffusion language modeling by estimating the ratios of the data distribution. 2023.

Lukas, N., Salem, A., Sim, R., Tople, S., Wutschitz, L., and Zanella-Béguelin, S. Analyzing leakage of personally identifiable information in language models. In *2023 IEEE Symposium on Security and Privacy (SP)*, pp. 346–363. IEEE, 2023.

Luo, X., Chen, Y., Bjerva, J., and Li, Q. Shared path: Unraveling memorization in multilingual llms through language similarities. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Suzhou, China, 2025.

Luo, X., Chen, Y., Li, Q., and Bjerva, J. Do llms really memorize personally identifiable information? revisiting pii leakage with a cue-controlled memorization framework. *arXiv preprint arXiv:2601.03791*, 2026.

Nasr, M., Rando, J., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., Ippolito, D., Choquette-Choo, C. A., Tramèr, F., and Lee, K. Scalable extraction of training data from aligned, production language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Nie, S., Zhu, F., Du, C., Pang, T., Liu, Q., Zeng, G., Lin, M., and Li, C. Scaling up masked diffusion models on text. In *ICLR*, 2025a.

Nie, S., Zhu, F., You, Z., Zhang, X., Ou, J., Hu, J., ZHOU, J., Lin, Y., Wen, J.-R., and Li, C. Large language diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b.

OpenAI. Hello GPT-4o. https://openai.com/index/hello-gpt-4o/, 2024.

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

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. Improving language understanding by generative pre-training. 2018.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

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

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

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

Zhang, J., Zhao, Q., Li, L., and Lin, C.-h. Extending memorization dynamics in pythia models from instance-level insights. *arXiv preprint arXiv:2506.12321*, 2025.

Zhang, Y., Gu, J., Wu, Z., Zhai, S., Susskind, J., and Jaitly, N. Planner: Generating diversified paragraph via latent language diffusion model. *Advances in Neural Information Processing Systems*, 36:80178–80190, 2023.

A. Proof of Theorem 4.3

Proof. Consider two sampling step counts $N_1 > N_2$, where the partition $\mathcal{P}_{N_1}$ is a refinement of $\mathcal{P}_{N_2}$. For any token $\pi \in \mathcal{M}$, let $\mathcal{U}^{(prev)}$ and $\mathcal{W}^{(prev)}$ be the observed index sets at the step immediately preceding the recovery of π in the respective N_1 and N_2 steps sampling processes.

Since $\mathcal{P}_{N_1}$ is a refinement of $\mathcal{P}_{N_2}$ and the recovery order σ is fixed, the N_1 process updates the observed set more frequently. Specifically, for any token π, the set of tokens already recovered in the N_1 process is always a superset of those recovered in the N_2 process at the time π is sampled:

$$\mathcal{W}^{(prev)} \subseteq \mathcal{U}^{(prev)}. \tag{13}$$

By Assumption 4.2, the probability of correctly recovering a token π increases with the size of the observed set. Therefore, for every $\pi \in \mathcal{M}$:

$$\Pr(\hat{z}_\pi = z_\pi \mid \boldsymbol{z}_{\mathcal{U}^{(prev)}}) \geq \Pr(\hat{z}_\pi = z_\pi \mid \boldsymbol{z}_{\mathcal{W}^{(prev)}}). \tag{14}$$

Since we assume the same recovery order of tokens in two cases, taking the product over all $\pi \in \mathcal{M}$, we can conclude that:

$$\Pr(\hat{\boldsymbol{z}}_{\mathcal{M}} = \boldsymbol{z}_{\mathcal{M}} \mid \boldsymbol{z}_{\bar{\mathcal{M}}}, N_1) \geq \Pr(\hat{\boldsymbol{z}}_{\mathcal{M}} = \boldsymbol{z}_{\mathcal{M}} \mid \boldsymbol{z}_{\bar{\mathcal{M}}}, N_2). \tag{15}$$

This completes the proof. □

B. Validation of generalized (ϵ, n, p)-discoverable extraction

We initiate our empirical analysis by performing $R = 10,000$ stochastic generation trials on a series of infilling tasks, where 20 tokens are randomly masked within a 100-token window, randomly selecting from Sec 6.2 masked-pattern. For each trial, we record the exact number of wrongly reconstructed tokens. Based on the empirical observations of the resulting distribution, we postulate the following assumptions.

Assumption B.1. For a diffusion-based LLM, let $d(\hat{\boldsymbol{z}}^{(i)}_{\mathcal{M}}, \boldsymbol{z}_{\mathcal{M}})$ denote the Hamming distance between the i-th independently sampled reconstruction and the ground-truth suffix. We observe that for sufficiently large $|\mathcal{M}|$, the distribution of d over multiple trials exhibits a strong central tendency. Specifically, within its central regime (excluding the tails), the distribution can be approximated by a normal distribution

$$d(\hat{\boldsymbol{z}}^{(i)}_{\mathcal{M}}, \boldsymbol{z}_{\mathcal{M}}) \approx \mathcal{N}(\mu, \sigma^2), \tag{16}$$

where the mean μ and variance σ^2 are parameters governed by the model's intrinsic extraction capability.

This Gaussian behavior is motivated by the fact that the Hamming distance can be decomposed into a sum of M indicator random variables

$$d = \sum_{j=1}^{|\mathcal{M}|} \mathbb{I}(\hat{z}_{\pi_j} \neq z_{\pi_j}), \tag{17}$$

where each indicator represents a token-level recovery failure. Owing to the inherent stochasticity of the sampling process in large DLMs, the total error count can be modeled as a sum of weakly dependent Bernoulli random variables. In accordance with the Central Limit Theorem (CLT) for weakly dependent sequences, the aggregate sum converges toward a normal distribution as the number of masked tokens $|\mathcal{M}|$ increases. The Quantile–Quantile (Q-Q) plots in Figure 6 (left) provide strong empirical evidence for the normality of the recovered token counts. Specifically, the strict adherence of the sample quantiles to the diagonal reference line suggests that the error distribution across $10,000$ trials is well-characterized by a normal distribution. While minor deviations appear in the tail portions, the empirical distribution aligns well with the theoretical normal curve in the central regime.

In contrast, Figure 7 reveals a markedly different trend for ARMs. Due to the sequential nature of the ARM sampling process, there exist significantly stronger inter-token dependencies, which prevent the distribution from converging to a normal form, leading to a distinct departure from the normality assumption.

The right column of Figure 6 compares the empirical CDF obtained via (12) with the Gaussian CDFs fitted using $10,000$ and 128 trials, respectively. As illustrated, the Gaussian approximation aligns well with the empirical CDF, effectively capturing the behavior of $\hat{p}_{\mathbf{z},\epsilon}$. While minor discrepancies still appear in the extreme regions, where the Gaussian model

tends to underestimate the probability of tail events far from the mean, the overall fit remains robust. Notably, the curves fitted with 128 and $10,000$ trials are remarkably similar. This suggests that a significantly smaller number of trials can be employed to obtain a rough estimation of $\hat{p}_{\mathbf{z},\epsilon}$ within the central regime, offering a more query-efficient alternative to the empirical counting method

$$\hat{p}_{\mathbf{z},\epsilon} = \Pr(d(\hat{\mathbf{z}}^{(i)}_{\mathcal{M}}, \mathbf{z}_{\mathcal{M}}) \leq \epsilon) \approx \Phi\left(\frac{\epsilon - \mu}{\sigma}\right), \tag{18}$$

where $\Phi(\cdot)$ denotes the CDF of the standard normal distribution $\mathcal{N}(0,1)$, while μ and σ represent the mean and standard deviation estimated from a few trials.

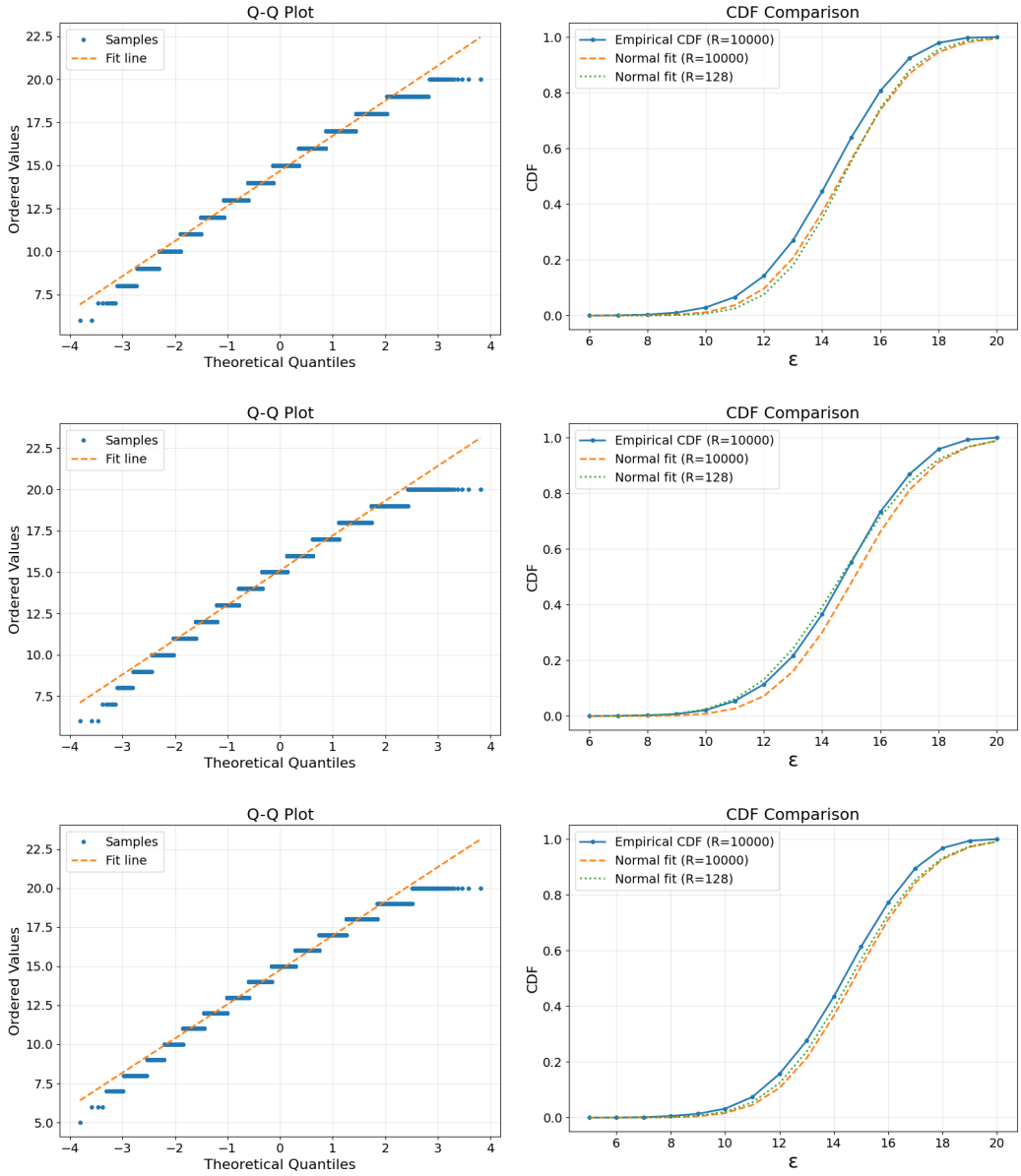


Figure 6. Paired Q-Q plots (left) and CDF comparisons (right) for three cases for DLM.

C. Effect of Masked-patterns on Memorization Estimation

In this section, we provide additional empirical evidence regarding the impact of the number of sampled masked-patterns R on the estimation of memorization probabilities. Figure 8 illustrates the correlation between the empirical memorization

Model	Step	Email				Phone			
		$p = 10\%$	$p = 50\%$	$p = 90\%$	$p = 99\%$	$p = 10\%$	$p = 50\%$	$p = 90\%$	$p = 99\%$
DLM-170M	One	0	0	0	0	0	0	0	0
	Max	7	2	0	0	0	0	0	0
DLM-629M	One	0	0	0	0	0	0	0	0
	Max	17	10	7	4	0	0	0	0
DLM-1.1B	One	0	0	0	0	0	0	0	0
	Max	26	16	10	7	0	0	0	0

Table 2. Generalized (n, p)-discoverable Extraction. Number of memorized samples at target probabilities p under a query budget of $n = 10{,}000$ on 3,000 completion prompts, for different PII types.

probability (p_z) and the theoretical prediction ($\hat{p}_z$) across different masked-patterns.

Convergence and Accuracy. Our results demonstrate that the estimation accuracy significantly improves as the number of sampled masked-patterns R increases from 128 to 512. We observe that sampling $R = 512$ times provides a high-fidelity fit to the theoretical line, suggesting that this amount of masked-patterns is sufficient to capture the underlying memorization behavior of DLMs without prohibitive computational overhead.

Analysis of Empirical Variance. We observe that the empirical points exhibit a degree of dispersion around the identity line across the entire probability spectrum . This variance is a natural consequence of the stochastic nature of the denoising process in DLMs:

- **Stochastic Sampling Fluctuations:** Since each empirical data point is calculated as the mean of R random masked, it is essentially a realization of a binomial-like estimator. Therefore, some level of statistical fluctuation is inevitable. Even for samples with high theoretical memorization probability, the specific set of R trajectories sampled may include more or fewer "successful" extraction paths than the expected average.

- **Natural Statistical Dispersion:** This variance does not represent a flaw in the theoretical derivation, but rather reflects the inherent randomness of the sampling-based estimation. Crucially, as R increases, the standard deviation of these fluctuations decreases, and the empirical values remain centered around the theoretical identity line within a consistent order of magnitude.

In summary, the empirical results in Figure 8 confirm that our theoretical framework provides an accurate "mean" behavior for data extraction. The observed variance is consistent with standard Monte Carlo estimation errors and does not impede the practical utility of our metric in identifying memorized content.

D. PII Memorization

In this section, we extend our evaluation of (n, p)-discoverable memorization to smaller model variants (DLM-170M and DLM-629M) across different types of Personally Identifiable Information (PII). Table 2 summarizes the number of memorized samples for email addresses and phone numbers under varying target probability thresholds p.

Low Leakage in Small-Scale DLMs. Our empirical results indicate that small-scale DLMs exhibit extremely low leakage probabilities for PII. This is particularly evident for *phone numbers*, where the number of memorized samples remains zero across all tested models and confidence levels $p \in \{10\%, 50\%, 90\%, 99\%\}$. Even for email addresses, which appear slightly more susceptible to extraction, the risk remains significantly lower compared to larger model scales.

Validation of Model Memorization. To ensure that these low extraction rates are a meaningful reflection of the model's privacy-preserving properties—rather than a byproduct of underfitting or a general failure to memorize the training set—we conducted auxiliary measurements. Specifically, we evaluated the DLM-170M and DLM-629M models using the same training data reconstruction methodology described in Sec 6.4. Figure 9 confirmed that both models show higher confidence in the Enron training corpus. Therefore, the low extraction rates reported in Table 2 serve as a valuable and reliable reference for PII risks in smaller diffusion language models.

E. Memorization examples

We include the following memorized examples from Section 6.2 in this appendix.

Original

```
1948 Summer Olympics.

References

External links


1927 births
2020 deaths
Athletes (track and field) at the 1948 Summer Olympics
Brazilian female sprinters
Olympic athletes of Brazil
Olympic female sprinters</s>  Why is Tommy Hunter making us all nervous?
One of the Phillies' offseason acquisitions
```

Recoverable Masked-Pattern

```
<MASK>  <MASK> 948 <MASK>  <MASK> . <MASK>
<MASK> ferences

External <MASK>


<MASK> 92 <MASK>  <MASK> s <MASK>  <MASK> 020 deaths <MASK> Athlet <MASK>  (track and <MASK> ) at the 19 <MASK>
<MASK>  Summer <MASK>  <MASK> Brazilian female sprinter <MASK>
Olympic <MASK> es <MASK>  Brazil <MASK> O <MASK> ic <MASK>  spr <MASK> s</s>  Why <MASK>  Tommy Hunter making us
all nervous?
One <MASK>  <MASK>  Phillies' offseason <MASK> quisitions
```

Original

```
te change is a natural phenomenon caused by several factors such as solar output, the planet's orbit, and ocean
circulation. However, human activity that releases greenhouse gases into the atmosphere has made the effects of
climate change more severe.
According to scientists, even if the world stopped emitting carbon dioxide today, global warming would persist
for decades to come because of the amount of carbon dioxide we've already pumped into the atmosphere.
```

Recoverable Masked-Pattern

```
te change is a natural <MASK>  <MASK>  caused <MASK>  several factors such <MASK>  solar output <MASK>  <MASK>
planet <MASK> s orbit, <MASK>  ocean circulation. However, human activity that releases green <MASK>  gases
<MASK>  the atmosphere <MASK>  made <MASK>  effects of climate change more severe.
According <MASK>  scientists, even <MASK>  the <MASK>  <MASK>  emitting carbon dioxide today, global war <MASK>
would persist for decades to come because of the amount of carbon diox <MASK>  we've already <MASK> ump <MASK>
into the atmosphere.
```

Original

```
control. Researchers believe that hereditary ataxia may supply clues to the genetic causes of RLS.
Attention-deficit hyperactivity disorder (ADHD).
Psychiatric disorders, such as depression.
Environmental and Dietary Factors
The following environmental and dietary factors can trigger or worsen RLS:
Iron deficiency. People who are deficient in iron are at risk for R
```

Recoverable Masked-Pattern

```
control. Researchers believe that her <MASK> ary ataxia may supply clues to the genetic causes of RLS.
Att <MASK>  <MASK>  <MASK> ic <MASK>  hyperactivity disorder <MASK> AD <MASK> ).
<MASK> sychiatric disorders, such as de <MASK>  <MASK>
<MASK> al and <MASK> ary Factors
The following environmental and <MASK> etary <MASK>  can trigger or <MASK> orsen RLS <MASK>
I <MASK>  deficiency. People who <MASK>  defic <MASK>  <MASK>  iron are at risk for R
```

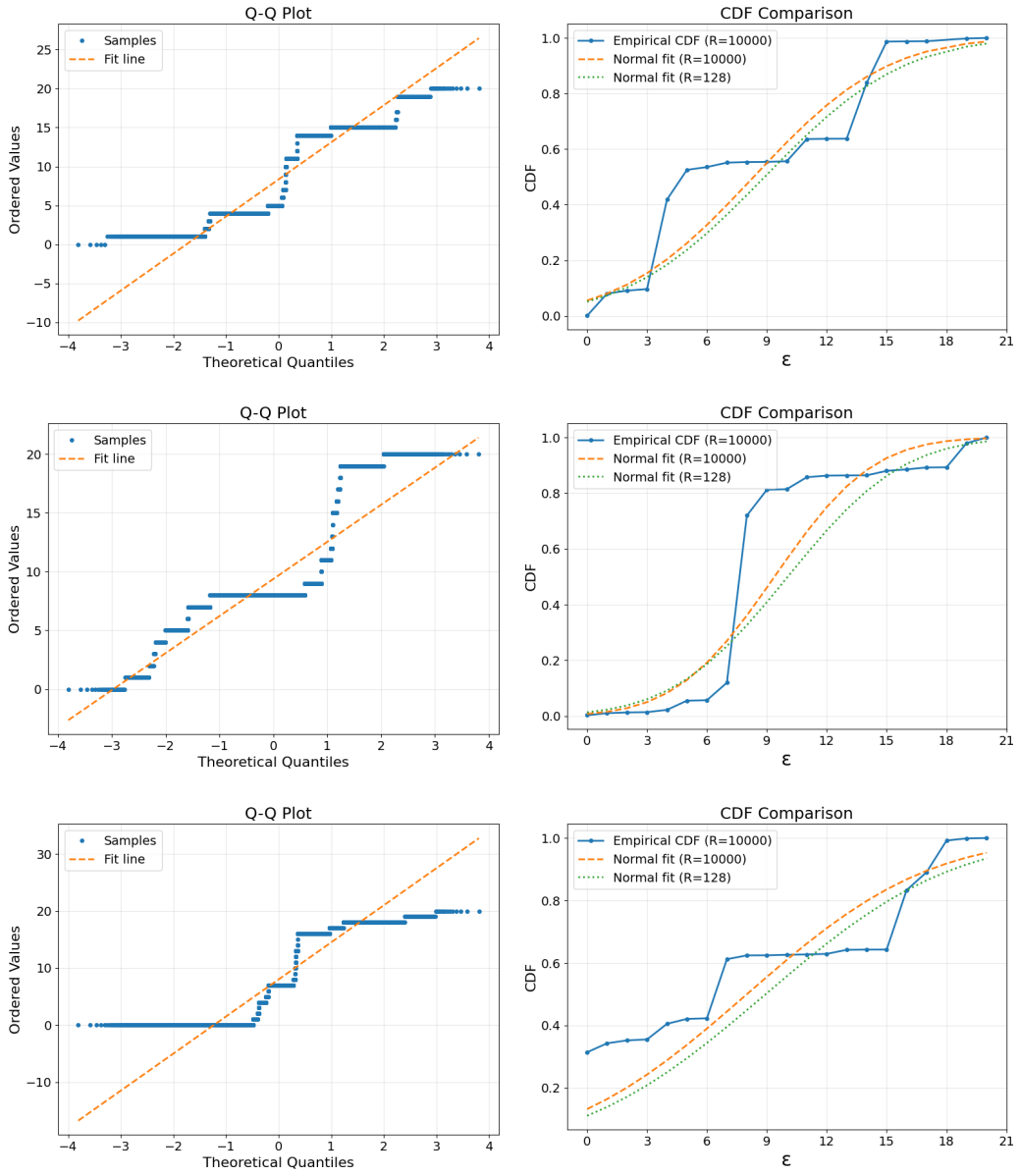
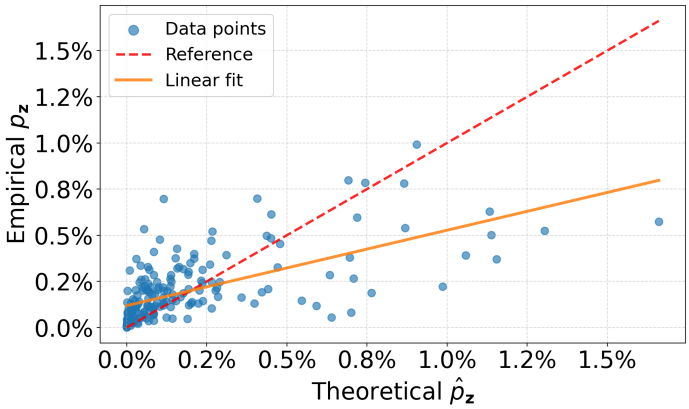
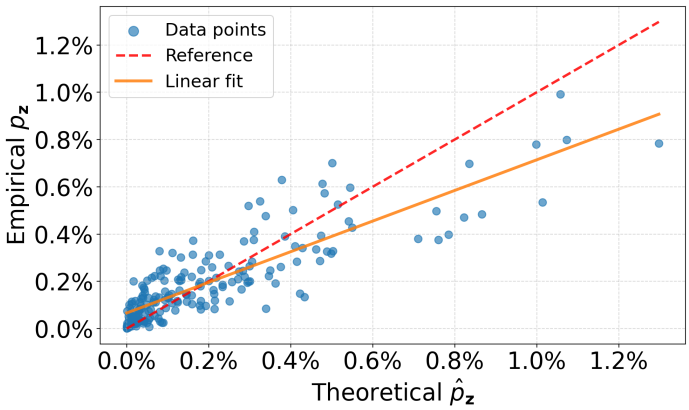


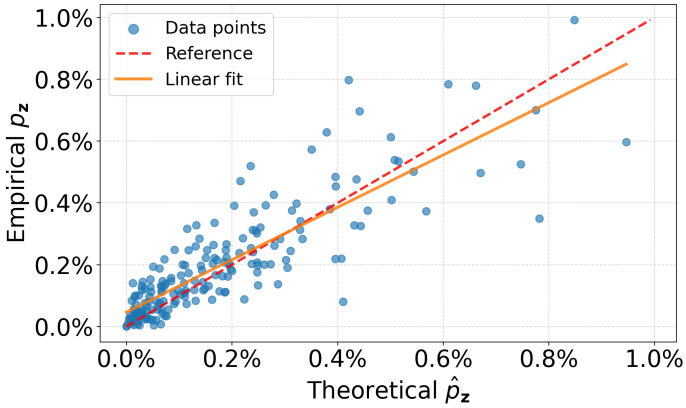
Figure 7. Paired Q-Q plots (left) and CDF comparisons (right) for three cases for ARM.



(a) masked-patterns = 128

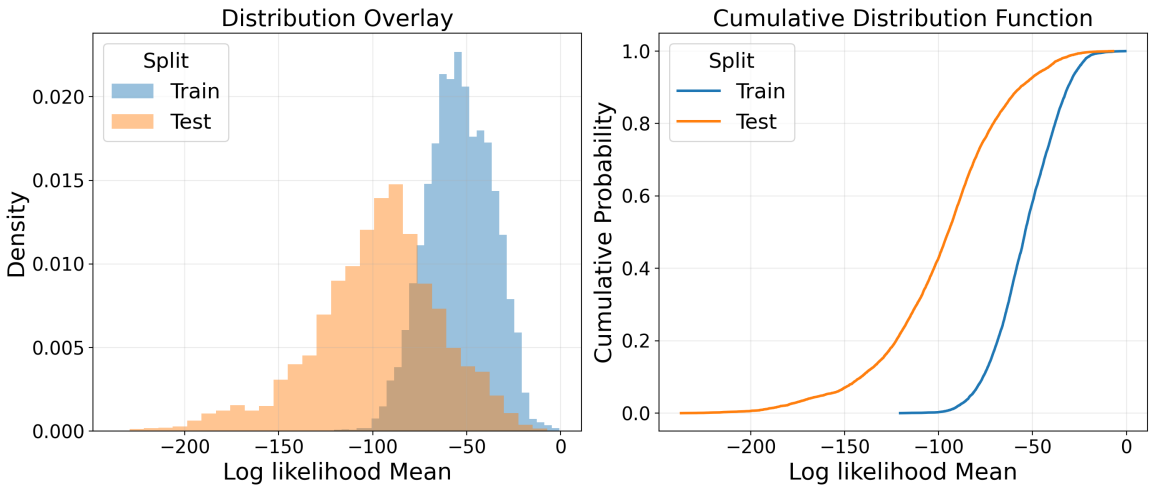


(b) masked-patterns = 256

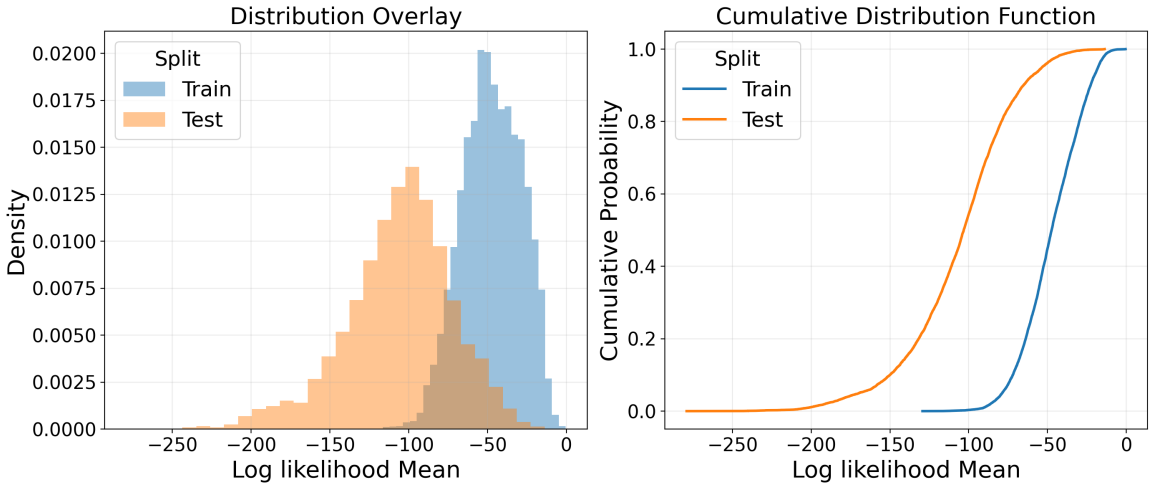


(c) masked-patterns = 512

Figure 8. Comparison of empirical and theoretical memorization probabilities under different sampling masked-patterns quantities. As the number of samples increases, the empirical values converge towards the theoretical identity line.



(a) DLM-170M



(b) DLM-629M

Figure 9. Validating memorization of training data under different DLM scales. We compare the reconstruction likelihood from the Enron (train) against the unseen from the TREC 2007 Spam (test), which is from the same domain.

F. PII extraction

We adopt the regular expressions in (Kim et al., 2023). to search for email addresses and U.S. phone numbers; the patterns are shown below.

```
EMAIL_RE = re.compile("^([a-zA-Z0-9_\-\.]+)@([a-zA-Z0-9_\-\.]+)\.([a-zA-Z]{2,5})$")
PHONE_RE = re.compile("[0-9][0-9][0-9][-.()][0-9][0-9][0-9][-.()][0-9][0-9][0-9][0-9]")
```