Weak Ties, Strong Signals: Efficient Training Data Detection in Diffusion LLMs via Independent Token Sampling

Anonymous ACL submission

Abstract

Diffusion large language models (dLLMs) offer a compelling alternative to autoregressive models, yet they may expose sensitive training data during denoising. Detecting such usage is challenging because dLLMs lack the efficient one-pass probability decomposition of causal architectures. Existing methods rely on random masking to obtain tractable token-wise detection signals under limited query budgets, but fail to control dependencies among masked tokens. We demonstrate that this token-wise approximation introduces a non-negative structural estimation error, which is theoretically characterized by the cumulative conditional mutual information (CMI) among masked tokens and can obscure subtle memorization signals. This insight suggests that reliable detection requires masked token sets with weak internal dependency. To avoid the prohibitive cost of directly estimating CMI over token combinations, we propose *Independent Token Sampling* (ITS), a query-efficient framework that uses an attention-derived pairwise dependency proxy to approximate the CMI-aware selection criterion. ITS further incorporates a diversity-promoting strategy to improve token coverage across sampling rounds, yielding aggregated token-wise signals that are less affected by dependency-induced approximation error. Experiments on multiple datasets show that ITS consistently outperforms state-of-the-art baselines across different models and datasets, achieving an AUC improvement of 0.18 on the ArXiv dataset while maintaining strong performance under limited query budgets.

1 Introduction

Are all masked token combinations equally reliable for exposing memorization on diffusion large language models? Is the dependency among them merely a sampling artifact, or a decisive factor in whether the probability estimate remains faithful? We argue the latter: for diffusion Large Language Models (dLLMs), memorization is more reliably exposed by token combinations that provide weakly dependent evidence. This question is central to training data detection for modern diffusion language models. Training data detection is becoming increasingly important for modern dLLMs, as their growing capabilities in parallel generation and bidirectional reasoning raise privacy and copyright concerns about memorized fine-tuning data. However, detecting such memorization is fundamentally more challenging than in autoregressive LLMs. AR-LLMs admit a causal factorization of sequence likelihood into token-level conditional probabilities, which can be efficiently computed in a single forward pass (Shi et al., 2024; Zhang et al., 2025; Huang et al., 2025). In contrast, dLLMs estimate token probabilities under masked contexts, making the resulting detection signal depend on the queried masking configurations.

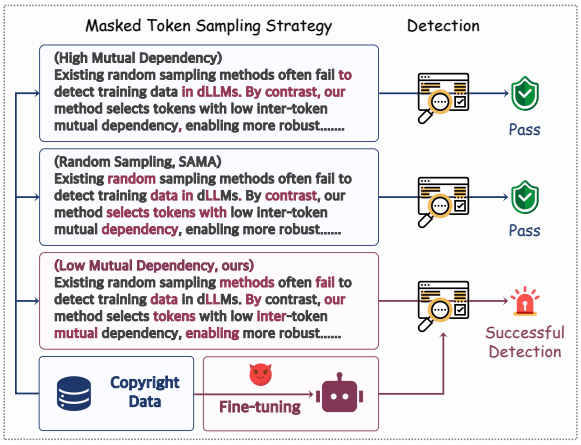
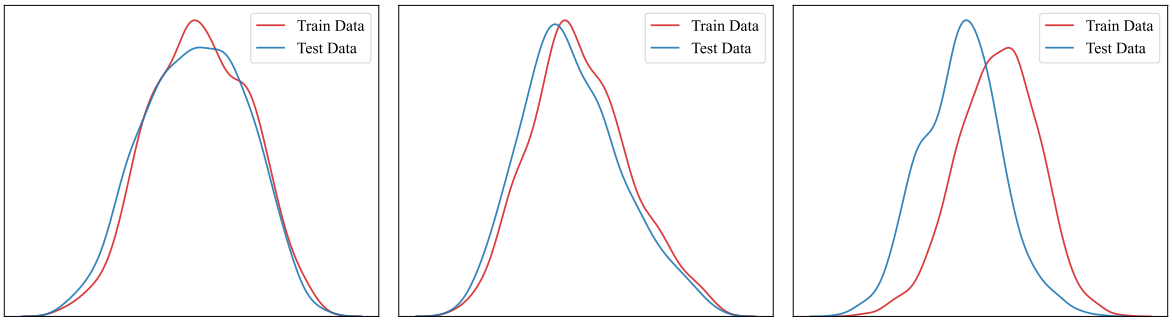


Figure 1: Illustration of token sampling strategies. Existing random sampling methods often fail to detect training data in dLLMs. By contrast, our method selects weakly dependent token subsets, enabling more robust detection.

We view this challenge through the lens of probability approximation. Ideally, training data detection should rely on a faithful estimate of the



(a) High Dependency (b) Random Sampling (c) Low Dependency

Figure 2: Distribution of sample-level detection scores under different token sampling strategies on LLaDA fine-tuned with ArXiv dataset. Each score is computed for one sample as the proportion of selected tokens where the target model assigns a higher probability than the reference model, and the distribution is obtained by aggregating scores over many samples. Compared with (a) high-dependency sampling and (b) random sampling, our (c) low-dependency strategy yields a clearer separation between training data (red) and test data (blue).

target sequence probability, but exact marginalization over all masking patterns is computationally infeasible because the number of possible configurations grows exponentially with sequence length. Practical detection must therefore approximate the desired joint probability using only tractable token-wise probabilities. The key question is therefore: *How can we obtain a faithful detection signal from tractable token-wise probabilities without exhaustive computation?*

A common solution is random token sampling (Chen et al., 2026), where a subset of tokens is masked and the model is scored by aggregating token-wise reconstruction signals. This makes probability estimation computationally tractable, but leaves the dependency structure of the sampled token set uncontrolled. The resulting surrogate implicitly treats sampled tokens as if their probability contributions can be accumulated independently, introducing a structural estimation error. As illustrated in Figure 1, this error can smear discriminative training signals and reduce detection reliability. We further prove that this structural error is characterized by the cumulative conditional mutual information (CMI) among sampled tokens.

This analysis suggests that the failure of random sampling is not merely due to insufficient samples, but to a dependency mismatch in probability approximation. A principled solution should therefore prefer token subsets with weak internal dependency. To this end, we propose *Independent Token Sampling* (ITS), an attention-guided framework for training data detection in dLLMs. Since directly estimating CMI for all candidate token combinations is computationally expensive, ITS leverages the model's internal attention structure to build a pairwise dependency proxy and selects token subsets with low estimated dependency. To improve the use of the sampling budget, we further introduce a diversity-promoting penalty, which encourages complementary token coverage across different sampling rounds.

Empirical evaluations across different model and data settings demonstrate the effectiveness of ITS. Our method consistently outperforms state-of-the-art (SOTA) baselines across multiple datasets and remains robust under challenging scenarios such as sparse model updates with LoRA and partial observations. These results show that modeling dependency structure is essential for faithful probability approximation in dLLMs. Our contributions are summarized as follows:

- **Theoretical Insight:** We identify the structural estimation error introduced by approximating joint probability with token-wise probability products. This error is characterized by the cumulative conditional dependency among masked tokens.

- **Independent Token Sampling:** We propose ITS, an attention-guided token sampling framework that selects weakly dependent token subsets. A diversity-promoting penalty is further introduced to improve token coverage across sampling rounds.

- **Empirical Validation:** We demonstrate that ITS significantly outperforms state-of-the-art baselines across different scenarios.

2 Related Work

2.1 Diffusion Large Language Models

Diffusion large language models have emerged as a scalable alternative to the autoregressive paradigm. Recent breakthroughs such as LLaDA (Nie et al., 2025) and Dream (Ye et al., 2025) demonstrate that dLLMs can match similarly sized autoregressive models in performance while excelling in parallel decoding (Wu et al., 2025) and mitigating the reversal curse (Berglund et al., 2023). Architectural scaling continues to advance through Mixture-of-Experts (Zhu et al., 2025) and extremely large-scale models like LLaDA 2.0 (Bie et al., 2025), which extends dLLMs to the 100B parameter regime. These efforts collectively transition dLLMs from theoretical interest to practically viable alternatives for large-scale deployment (Song et al., 2025).

2.2 Training Data Detection on (d)LLMs

Training data detection aims to identify whether a specific sample was included in a model's training set, a task crucial for copyright auditing. In the context of autoregressive (AR) models, early methods utilized loss-based signals (Yeom et al., 2018) or likelihood ratios with reference models (Watson et al., 2021). Recent advances focus on token-level statistics like Min-k% (Shi et al., 2024; Zhang et al., 2025), input perturbations (Mattern et al., 2023), or contrastive conditional likelihoods (Wang et al., 2025; Huang et al., 2025).

However, these methods struggle with the unique inference mechanisms of dLLMs, where training signals are highly dependent on specific masking patterns. To address this, SAMA (Chen et al., 2026) introduces a progressive mask-density sampling and sign-based aggregation strategy to mitigate signal sparsity. While SAMA improves detection for dLLMs, the base question for probability approximation is not resolved.

3 Preliminaries

3.1 Diffusion Large Language Models

Diffusion large language models (dLLMs) generate text through masked denoising rather than left-to-right prediction. Given an input sequence x of length $|x|$, a subset of positions $S \subseteq [|x|]$, where $[|x|] = \{1, \ldots, |x|\}$, is sampled and replaced with a special [MASK] token. We use x_S to denote the original tokens at masked positions and $x_{[|x|]\setminus S}$ to denote the partially observed sequence. The core module of dLLMs is a mask predictor $\mathcal{P}(\cdot \mid x_{[|x|]\setminus S})$, which predicts all masked tokens in parallel. At inference time, generation starts from a fully or partially masked sequence, and the model iteratively predicts and reveals tokens until a complete sequence is obtained. Since masked positions can be predicted in parallel, dLLMs avoid the strict left-to-right constraint of autoregressive models and are naturally suited for bidirectional reasoning, infilling, and controllable generation.

3.2 Problem Formulation

We study the problem of *training data detection* (Shi et al., 2024) for diffusion large language models, focusing on the fine-tuning stage (Chen et al., 2026). Specifically, given a pre-trained reference model $\mathcal{P}_{\text{ref}}$ and a training dataset $\mathcal{D}$, the model is fine-tuned to obtain $\mathcal{P}_\theta$. The goal of an auditor is to determine whether a specific data sample x is used to train the model, *i.e.*, $x \in \mathcal{D}$ or $x \notin \mathcal{D}$.

Adopting the grey-box framework from Chen et al. (2026), we assume the auditor possesses the capability to query the target model with arbitrary, partially masked sequences. Under this setting, the auditor observes the resulting output distributions (e.g., logits) but lacks access to the model's internal weights or gradients. Furthermore, we grant the auditor full access to publicly available, pre-trained reference models to facilitate the auditing process.

4 Method

4.1 Full-Sequence Signal for Detection

Let $\mathcal{P}_\theta(x)$ and $\mathcal{P}_{\text{ref}}(x)$ be the generative distributions of the target and reference dLLMs, respectively. A common observation in detection is that a fine-tuned model tends to assign higher probability to its training samples than a comparable reference model (Huang et al., 2025):

$$x \in \mathcal{D} \implies \mathcal{P}_\theta(x) > \mathcal{P}_{\text{ref}}(x). \tag{1}$$

Thus a natural signal for detection is

$$\Delta(x) := \frac{\mathcal{P}_\theta(x)}{\mathcal{P}_{\text{ref}}(x)}. \tag{2}$$

Higher $\Delta(x)$ indicates greater confidence of $x \in \mathcal{D}$.

However, unlike AR-LLMs (Shi et al., 2024; Watson et al., 2021), dLLMs do not provide an explicit tractable factorization of the full-sequence probability $\mathcal{P}(x)$. A common approximation is to

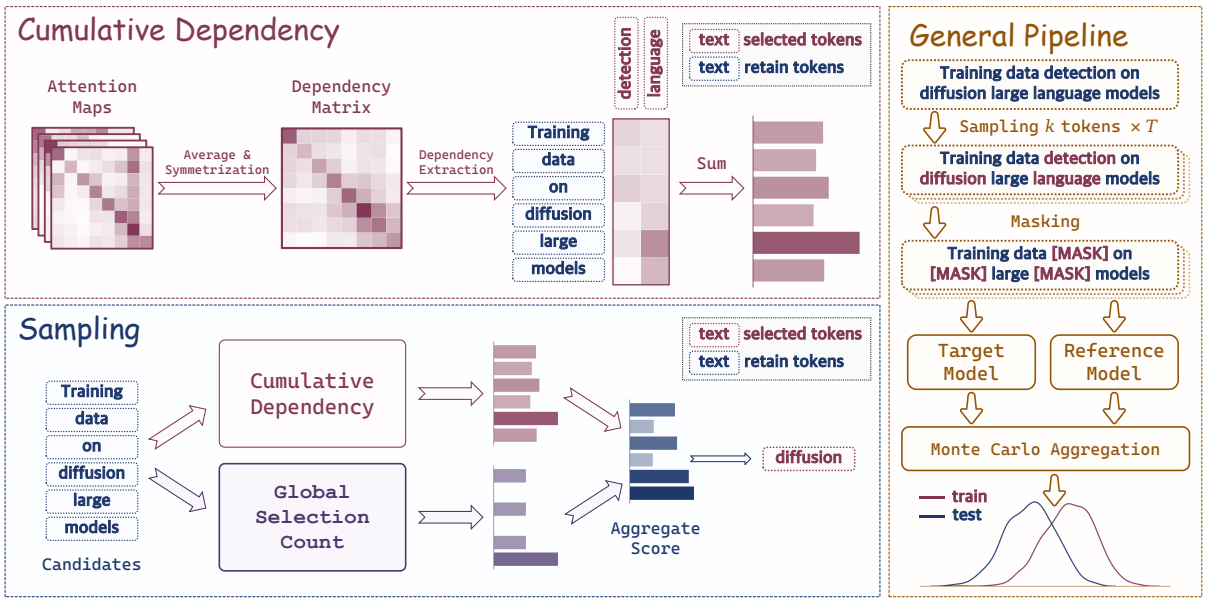


Figure 3: Overview of our proposed ITS detection method

replace the intractable joint probability with a product of token-wise probabilities (Nie et al., 2025; Chen et al., 2026):

$$\hat{\mathcal{P}}(x) := \prod_{i=1}^{|x|} \mathcal{P}(x_i). \tag{3}$$

But $\hat{\mathcal{P}}(x)$ is inaccurate due to the dependence among $x_1, x_2, \ldots, x_{|x|}$.

KL divergence measures the distance between two probability distributions. With this metric, we define the expected error of $\hat{\mathcal{P}}(x)$ as

$$\text{error}\Big[\hat{\mathcal{P}}(x)\Big] := \mathcal{KL}\Big(\mathcal{P}(x) \Big\| \hat{\mathcal{P}}(x)\Big). \tag{4}$$

This error can be determined by the mutual information $\mathcal{I}$ among all tokens, which is revealed by our Theorem 1.

Theorem 1 (Full-Sequence Error Decomposition)**.**

$$\text{error}\Big[\hat{\mathcal{P}}(x)\Big] = \sum_{i=2}^{|x|} \mathcal{I}\big(X_{<i}; X_i\big). \tag{5}$$

The proof is provided in Appendix A. Mutual information $\mathcal{I}$ measures the dependence between two random variables. In practice, the total mutual information among all tokens is substantial, leading to an intolerable error when estimating $\mathcal{P}(x)$. This error renders $\Delta(x)$ unreliable for the detection. This motivates the use of selected token subsets, which can reduce the dependency-induced approximation error, while preserving useful detection signals.

4.2 Subset Signal for Error Reduction

Motivated by the above analysis, we turn from the full sequence to selected token subsets. Let $S \subseteq [|x|]$ be the set of masked positions, and let $c := x_{[|x|]\setminus S}$ denote the unmasked context. We define the selected-token signal as

$$\Delta(x_S|c) := \frac{\mathcal{P}_\theta(x_S|c)}{\mathcal{P}_{\text{ref}}(x_S|c)}, \quad S \subseteq [|x|]. \tag{6}$$

By masking the positions in S and querying the dLLM with the visible context c, we obtain token-wise conditional probabilities $\mathcal{P}(x_{S_i}|c)$ for $i = 1, 2, \ldots, |S|$. Similarly, the conditional probability $\mathcal{P}(x_S|c)$ is estimated by:

$$\hat{\mathcal{P}}(x_S|c) := \prod_{i=1}^{|S|} \mathcal{P}(x_{S_i}|c). \tag{7}$$

Similar to $\hat{\mathcal{P}}(x)$, we define the expected error of the conditional probability $\hat{\mathcal{P}}(x_S|c)$ as:

$$\text{error}\Big[\hat{\mathcal{P}}(x_S|c)\Big] := \mathcal{KL}\Big(\mathcal{P}(x_S|c) \Big\| \hat{\mathcal{P}}(x_S|c)\Big). \tag{8}$$

This error can be determined by the conditional mutual information (CMI) among the selected tokens, which is revealed by our Theorem 2.

Theorem 2 (Subset Error Decomposition)**.**

$$\text{error}\Big[\hat{\mathcal{P}}(x_S|c)\Big] = \sum_{i=2}^{|S|} \mathcal{I}\big(X_{S_{<i}}; X_{S_i}|c\big). \tag{9}$$

Algorithm 1 Independent Token Sampling (ITS)

Input: Dependency matrix $\mathbf{A} \in \mathbb{R}^{L \times L}$, sampling rounds T, selection sizes $\{k_1, \ldots, k_T\}$, diversity penalty λ

Output: Selected token sets $S^1, \ldots, S^T$

Initialize global selection counts $\mathbf{C} \leftarrow \mathbf{0}$
for $t = 1$ **to** T **do**
  Select seed token $v^* \leftarrow \arg\min_{v \in [L]} C_v$
  Initialize $S^t \leftarrow \{v^*\}$
  Update $C_{v^*} \leftarrow C_{v^*} + 1$
  Initialize cumulative dependency $\mathbf{h} \leftarrow \mathbf{A}_{v^*,:}$
  while $|S^t| < k_t$ **do**
    Compute candidate scores:
    $u_v \leftarrow h_v + \lambda C_v, \quad \forall v \notin S^{(t)}$
    Select $v_{\text{next}} \leftarrow \arg\min_{v \notin S^{(t)}} u_v$
    Update $S^t \leftarrow S^t \cup \{v_{\text{next}}\}$
    Update $C_{v_{\text{next}}} \leftarrow C_{v_{\text{next}}} + 1$
    Update $\mathbf{h} \leftarrow \mathbf{h} + \mathbf{A}_{v_{\text{next}},:}$
  end
end
return $S^1, \ldots, S^T$

Theorem 2 reveals that the selected-token signal becomes more reliable when the masked tokens are weakly dependent given the context. Therefore, we formulate a greedy dependency-aware selection principle. Each token is selected by

$$S_i \leftarrow \underset{j \in [|x|] \setminus S_{<i}}{\arg\min} \; \mathcal{I}\big(X_{S_{<i}}; X_j \big| c'\big), \tag{10}$$

where $c' := x_{[|x|] \setminus (S_{<i} \cup \{j\})}$ denotes the visible context when the previously selected tokens and the candidate token are masked, for $i = 2, 3, \ldots, |S|$. This criterion selects the token with the weakest dependency on the previously selected tokens.

4.3 Independent Token Sampling

Motivated by the greedy CMI-aware principle for constructing masked token sets described above, we introduce Independent Token Sampling (ITS). However, directly estimating the conditional mutual information for all candidates is computationally expensive. To make this principle practical, we leverage the internal attention structure of dLLMs to build a pairwise dependency proxy for the CMI-based selection criterion. While this attention-derived proxy does not directly measure conditional mutual information, it provides an efficient model-internal estimate of relative dependency between token positions (Wen et al., 2022). Therefore, ITS selects token combinations with low estimated dependency to reduce the error characterized by Theorem 2.

Specifically, let $\mathbf{A}_{l,h}$ be the attention map of the l-th layer and h-th head,

$$\overline{\mathbf{A}} := \underset{l,h}{\mathbf{mean}} \, \mathbf{A}_{l,h}, \tag{11}$$

$$\mathbf{A} := \overline{\mathbf{A}} + \overline{\mathbf{A}}^T. \tag{12}$$

We obtain $\mathbf{A}_{l,h}$ from the available reference model. $\mathbf{A}$ represents pairwise token dependency, where a lower value A_{ij} indicates weaker estimated dependency between token positions i and j.

Given $\mathbf{A}$, ITS constructs masked sets through the greedy procedure in Algorithm 1. For each sampling round, the algorithm starts from a seed token and then progressively adds the token with the lowest cumulative dependency to the already selected set. For a candidate token v, this dependency score is computed by accumulating its dependency with all previously selected tokens. This encourages the masked set to contain weakly connected tokens, making the token-wise signal approximation less affected by internal dependency.

To avoid repeatedly selecting the same isolated positions across different rounds, ITS further introduces a diversity-promoting penalty λC_v into the candidate score, where C_v records the number of selections for position v. This penalty encourages complementary low-dependency masked sets while preserving the dependency-reduction principle.

Overall, ITS implements the CMI-aware selection principle through an attention-derived dependency proxy. By reducing dependency within each masked set and improving coverage diversity across rounds, ITS provides more reliable token-wise signals for training data detection in dLLMs.

4.4 General Pipeline

Following the hierarchical masking framework proposed in Chen et al. (2026), we compute the score by querying the target model for T rounds. At each query step t, a masking ratio ρ_t is specified, with values linearly spaced between $\rho_{\min}$ and $\rho_{\max}$, and the masked subset $S^{(t)}$ is selected using ITS. The point-wise log-likelihood ratio is then defined as follows:

$$\phi(i, S^t) := \frac{\mathcal{P}_\theta(x_i | x_{[|x|] \setminus S^t})}{\mathcal{P}_{\text{ref}}(x_i | x_{[|x|] \setminus S^t})}. \tag{13}$$

Within this framework, we further employ a Monte Carlo aggregation strategy by randomly

drawing N subsets $\{S^t_{(j)}\}_{j=1}^N$ of size M from S^t, and computing the proportion of subsets with a positive cumulative ratio (Chen et al., 2026):

$$\phi(S^t) = \frac{1}{N}\sum_{j=1}^{N} \mathbb{1}\Big[\prod_{i \in S^t_{(j)}} \phi(i, S^t) > 1\Big]. \quad (14)$$

The final score is the average proportion across all T rounds: $\text{Score}(x) = \frac{1}{T}\sum_{t=1}^{T}\phi(S^t)$. The higher score indicates that the sample x is more likely to be contained in the training dataset. Unlike standard random masking, our ITS ensures that tokens within S^t are sampled from non-redundant regions, maximizing the signal for detection.

5 Experiment

5.1 Experimental Setup

Target Models and Datasets. Following Chen et al. (2026), we evaluate our method on two types of diffusion large language models: LLaDA (Nie et al., 2025) and Dream (Ye et al., 2025). The former is trained from scratch, while the latter is adapted from autoregressive language models (AR-LLMs). We conduct experiments on six datasets, including AG News (Zhang et al., 2015) and five domains from the MIMIR benchmark (Duan et al., 2024), namely ArXiv, Pile-CC, Wikipedia (en), PubMed Central, and GitHub. We follow Chen et al. (2026) for the construction of the training and test splits. The training details are provided in Appendix B.1.

Evaluation Metrics. We evaluate the effectiveness of the detection method using standard metrics (Carlini et al., 2022), including the area under the ROC curve (AUC) and the true positive rate (TPR) at 10% and 1% false positive rates (FPR), denoted as T@10% and T@1%, respectively. Full numerical results will be provided in Appendix C.

Baselines. We compare our method with several representative baselines, including Loss (Yeom et al., 2018), Min-k% (Shi et al., 2024), Min-k%++ (Zhang et al., 2025), Ratio (Watson et al., 2021), DF-MIA (Huang et al., 2025), and SAMA (Chen et al., 2026). Among them, only SAMA is specifically designed for diffusion large language models, while the other baselines were originally proposed for AR-LLMs and are adapted following the protocol in Chen et al. (2026).

Implementation Details. We follow the implementation of SAMA (Chen et al., 2026) for all experimental settings, except for the sampling budget. Specifically, we uses a small budget of $T = 4$ throughout the experiments. Additional implementation details and hyperparameters are provided in Appendix B.2.

5.2 Comparison to Baselines

Table 1 reports the detection performance of ITS and existing baselines on six datasets using LLaDA. Overall, ITS achieves the strongest performance across datasets and metrics, showing that selecting weakly dependent masked token sets provides a more reliable signal than random masking or autoregressive-style scoring heuristics.

ITS obtains an average AUC of approximately 0.87, improving over the strongest baseline SAMA (0.77) by about 13% relatively. The improvement is especially clear compared with traditional detection method for AR-LLMs such as Loss, Zlib, Min-k%, Min-k%++, DF-MIA, and Ratio, which mostly remain close to random guessing. This suggests that methods designed for autoregressive do not directly transfer to diffusion paradigm.

Across datasets, ITS remains consistently effective under different levels of detection difficulty. On high-signal datasets such as Pile CC and Wikipedia, ITS achieves near-saturated AUC scores of 0.99 and 0.95, respectively, while also providing large gains at strict FPR thresholds. On more challenging datasets such as ArXiv, PubMed Central, and AG News, where most baselines are substantially weaker, ITS still maintains clear AUC improvements. This robustness supports our hypothesis that controlling dependency within the masked token set reduces likelihood smearing and preserves subtle memorization signals.

We further evaluate ITS on the Dream model (Ye et al., 2025), with results reported in Appendix C.1. ITS maintains clear advantages over all baselines and achieves similar AUC improvements to those observed on LLaDA. This suggests that the effectiveness of ITS is not tied to a specific dLLM architecture, but generalizes across different diffusion language models.

5.3 Detection on LoRA: Localization of Training Signals

To investigate the architectural localization of training signatures, we evaluate ITS on models fine-tuned via Low-Rank Adaptation (LoRA). We vary

Method	ArXiv			Pile CC			AG News		
	AUC	T@10%	T@1%	AUC	T@10%	T@1%	AUC	T@10%	T@1%
Loss	$0.52_{\pm 0.01}$	$0.13_{\pm 0.01}$	$0.01_{\pm 0.01}$	$0.52_{\pm 0.01}$	$0.13_{\pm 0.01}$	$0.02_{\pm 0.01}$	$0.52_{\pm 0.00}$	$0.10_{\pm 0.01}$	$0.01_{\pm 0.01}$
Zlib	$0.49_{\pm 0.01}$	$0.11_{\pm 0.02}$	$0.01_{\pm 0.01}$	$0.50_{\pm 0.01}$	$0.11_{\pm 0.02}$	$0.01_{\pm 0.00}$	$0.53_{\pm 0.01}$	$0.12_{\pm 0.01}$	$\mathbf{0.02}_{\pm 0.01}$
Min-k%	$0.51_{\pm 0.01}$	$0.13_{\pm 0.01}$	$0.01_{\pm 0.01}$	$0.52_{\pm 0.01}$	$0.13_{\pm 0.01}$	$0.02_{\pm 0.01}$	$0.52_{\pm 0.00}$	$0.10_{\pm 0.01}$	$0.01_{\pm 0.01}$
Min-k%++	$0.49_{\pm 0.01}$	$0.08_{\pm 0.01}$	$0.01_{\pm 0.00}$	$0.50_{\pm 0.02}$	$0.08_{\pm 0.02}$	$0.01_{\pm 0.00}$	$0.51_{\pm 0.01}$	$0.11_{\pm 0.01}$	$0.01_{\pm 0.00}$
DF-MIA	$0.52_{\pm 0.01}$	$0.14_{\pm 0.01}$	$0.01_{\pm 0.00}$	$0.54_{\pm 0.02}$	$0.11_{\pm 0.01}$	$0.01_{\pm 0.00}$	$0.61_{\pm 0.02}$	$0.18_{\pm 0.02}$	$0.01_{\pm 0.00}$
Ratio	$0.52_{\pm 0.01}$	$0.11_{\pm 0.01}$	$0.01_{\pm 0.01}$	$0.54_{\pm 0.01}$	$0.13_{\pm 0.02}$	$0.02_{\pm 0.00}$	$0.55_{\pm 0.01}$	$0.11_{\pm 0.02}$	$0.01_{\pm 0.00}$
SAMA	$0.70_{\pm 0.01}$	$0.34_{\pm 0.03}$	$0.06_{\pm 0.01}$	$0.92_{\pm 0.01}$	$0.78_{\pm 0.02}$	$0.40_{\pm 0.04}$	$0.67_{\pm 0.01}$	$0.25_{\pm 0.02}$	$0.00_{\pm 0.00}$
ITS (Ours)	$\mathbf{0.88}_{\pm 0.01}$	$\mathbf{0.64}_{\pm 0.02}$	$\mathbf{0.26}_{\pm 0.06}$	$\mathbf{0.99}_{\pm 0.00}$	$\mathbf{0.97}_{\pm 0.00}$	$\mathbf{0.88}_{\pm 0.02}$	$\mathbf{0.77}_{\pm 0.01}$	$\mathbf{0.37}_{\pm 0.01}$	$0.01_{\pm 0.00}$

Method	PubMed Central			Wikipedia (en)			Github		
	AUC	T@10%	T@1%	AUC	T@10%	T@1%	AUC	T@10%	T@1%
Loss	$0.51_{\pm 0.01}$	$0.14_{\pm 0.01}$	$0.01_{\pm 0.01}$	$0.51_{\pm 0.01}$	$0.13_{\pm 0.01}$	$0.01_{\pm 0.00}$	$0.59_{\pm 0.01}$	$0.22_{\pm 0.02}$	$0.04_{\pm 0.01}$
Zlib	$0.48_{\pm 0.01}$	$0.09_{\pm 0.02}$	$0.00_{\pm 0.00}$	$0.49_{\pm 0.01}$	$0.09_{\pm 0.01}$	$0.01_{\pm 0.00}$	$0.58_{\pm 0.01}$	$0.22_{\pm 0.02}$	$0.05_{\pm 0.01}$
Min-k%	$0.51_{\pm 0.01}$	$0.14_{\pm 0.01}$	$0.01_{\pm 0.01}$	$0.51_{\pm 0.01}$	$0.13_{\pm 0.01}$	$0.01_{\pm 0.00}$	$0.59_{\pm 0.01}$	$0.22_{\pm 0.02}$	$0.04_{\pm 0.01}$
Min-k%++	$0.49_{\pm 0.01}$	$0.10_{\pm 0.01}$	$0.01_{\pm 0.01}$	$0.51_{\pm 0.01}$	$0.12_{\pm 0.02}$	$0.01_{\pm 0.00}$	$0.52_{\pm 0.01}$	$0.08_{\pm 0.02}$	$0.01_{\pm 0.00}$
DF-MIA	$0.50_{\pm 0.01}$	$0.11_{\pm 0.01}$	$0.01_{\pm 0.00}$	$0.51_{\pm 0.01}$	$0.10_{\pm 0.01}$	$0.01_{\pm 0.01}$	$0.52_{\pm 0.02}$	$0.09_{\pm 0.01}$	$0.01_{\pm 0.00}$
Ratio	$0.52_{\pm 0.01}$	$0.12_{\pm 0.01}$	$0.01_{\pm 0.00}$	$0.50_{\pm 0.01}$	$0.11_{\pm 0.01}$	$0.01_{\pm 0.01}$	$0.58_{\pm 0.01}$	$0.16_{\pm 0.01}$	$0.02_{\pm 0.00}$
SAMA	$0.70_{\pm 0.02}$	$0.30_{\pm 0.02}$	$0.04_{\pm 0.01}$	$0.82_{\pm 0.01}$	$0.54_{\pm 0.01}$	$0.20_{\pm 0.03}$	$0.79_{\pm 0.01}$	$0.44_{\pm 0.02}$	$0.08_{\pm 0.02}$
ITS (Ours)	$\mathbf{0.77}_{\pm 0.01}$	$\mathbf{0.39}_{\pm 0.03}$	$\mathbf{0.08}_{\pm 0.04}$	$\mathbf{0.95}_{\pm 0.00}$	$\mathbf{0.88}_{\pm 0.01}$	$\mathbf{0.55}_{\pm 0.03}$	$\mathbf{0.86}_{\pm 0.01}$	$\mathbf{0.59}_{\pm 0.03}$	$\mathbf{0.12}_{\pm 0.04}$

Table 1: Detection Performance on LLaDA.

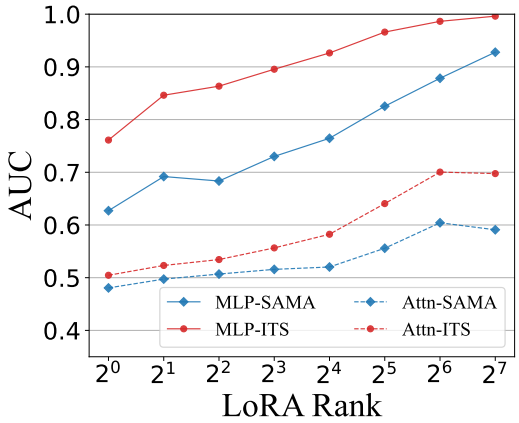


Figure 4: Training data detection performance on LoRA-tuned models across different ranks r and modules.

the rank $r \in \{1, \ldots, 128\}$ and compare the detectability of training data when targeting either MLP or Attention layers. As illustrated in Figure 4, two primary insights emerge:

Superiority of ITS. Across all ranks and modules, ITS consistently outperforms the SAMA baseline. For instance, when targeting the MLP at $r = 2$, ITS achieves an AUC of 0.85, significantly higher than SAMA's 0.69. This indicates that ITS maintains robust detection performance even with sparse model updates, outperforming baselines in regimes with limited training evidence.

MLP as the Knowledge Repository. The detection is substantially more effective when targeting MLP layers compared to attention blocks. At $r = 32$, the AUC for ITS targeting MLPs reaches a near-perfect 0.97, while its performance on Attention layers is only 0.64. This disparity aligns with interpretability research characterizing MLPs as key-value memories for training data (Meng et al., 2022; Geva et al., 2021), whereas attention layers primarily manage contextual routing. These results confirm that copyrighted information is predominantly internalized within MLP weights. Furthermore, they demonstrate that ITS provides a highly sensitive mechanism for identifying these localized training signals, even in constrained low-rank adaptation scenarios.

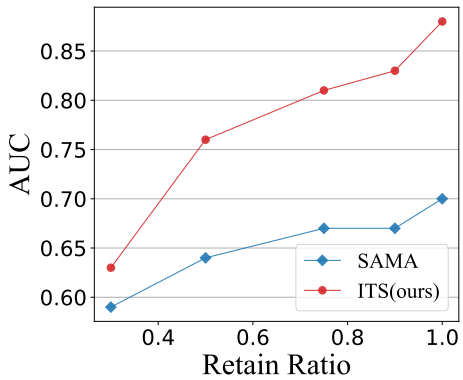
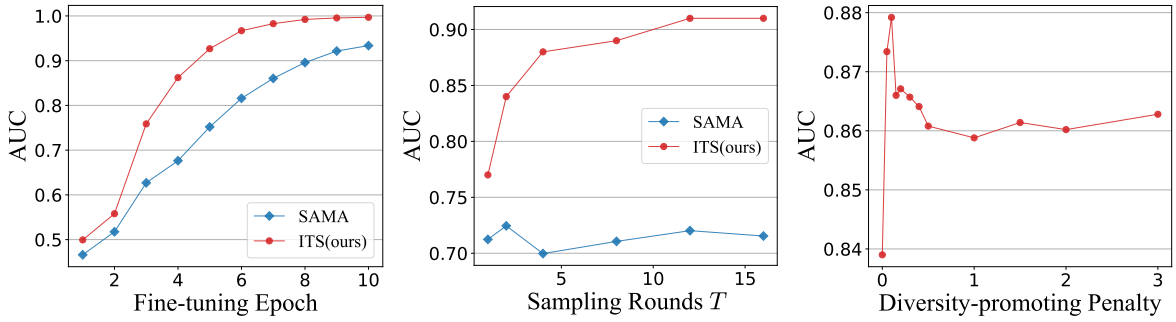


Figure 5: Detection performance across varying text retain ratios



(a) (b) (c)

Figure 6: Impact of key parameters on detection AUC. (a) Training epochs (b) Sampling rounds (c) Diversity-promoting penalty λ.

5.4 Robustness Evaluation

To evaluate detection robustness on content snippets, we truncate test sequences to various retain ratios (30% to 100%). As shown in Fig. 5, ITS demonstrates superior data efficiency; even with 30% of the sequence, it maintains a robust lead over SAMA. Notably, the performance gap widens as more context is available, reaching an AUC of 0.88 at full length. This indicates that ITS more effectively aggregates evidence from longer dependencies by mitigating structural bias. See Appendix C.3 for detailed numerical results. Additionally, we explore the robustness of temperature scaling in Appendix C.6.

5.5 Scaling Analysis

We examine how detection performance scales with the degree of fine-tuning epochs and the computational budget. The detailed numerical results are presented in Appendix C.4.

Training Epochs. As shown in Fig. 6a, detectability increases with the number of training epochs as the model further internalizes the training distribution. Notably, ITS exhibits a significantly steeper improvement trajectory than SAMA in early epochs, indicating its superior sensitivity to initial stages of memorization.

Sampling Rounds. Increasing the query budget T enhances the precision of likelihood estimation. Fig. 6b shows that while SAMA's performance fluctuates within a narrow range due to approximation bias, ITS scales consistently with T, achieving an AUC of 0.91 at $T = 16$. This confirms that ITS more effectively leverages additional sampling to refine the detection signal.

5.6 Ablation Study

Impact of Dependency Levels. We investigate how the dependency level of sampled token subsets affects.Implementation details are provided in Appendix B.3. The low-dependency setting corresponds to our method. The medium-dependency setting adopts random sampling, whereas the high-dependency setting favors subsets with stronger inter-token dependency. As shown in Table 2, lower dependency among selected tokens leads to stronger detection performance.

Dependency level	AUC	T@10%	T@1%
High	0.57	0.16	0.02
Medium	0.81	0.49	0.13
Low (Ours)	**0.87**	**0.64**	**0.28**

Table 2: Abaltion study on dependency levels.

Diversity-promoting Penalty. We evaluate the impact of the penalty λ, designed to reduce sampling redundancy and enhance query efficiency. As shown in Fig. 6c, any $\lambda > 0$ improves AUC from 0.84 to a peak of 0.88. This confirms that penalty effectively surfaces more discriminative signals within a limited query budget.

6 Conclusion

This paper addresses the challenge of training data detection in dLLMs by identifying a structural error in probability estimation. To mitigate this, we propose Independent Token Sampling (ITS), which leverages internal attention structures to minimize token dependence. Our results demonstrate that ITS significantly improves both accuracy and query efficiency, providing a principled and scalable solution for verifying data compliance in dLLMs.

Limitations

Despite its effectiveness, ITS has several limitations. (1) Generalization to PEFT: Since ITS identifies independent tokens based on the reference model's attention structure, its efficacy may slightly decrease when auditing models adapted via Attention LoRA. We hypothesize that because LoRA updates are constrained to a narrow parameter subspace, the resulting memorization signals might be more localized or "diffused," making them less aligned with the global dependency patterns captured by the base model's Dependency Matrix. (2) Linguistic generalization: Our evaluation is primarily English-centric. The cross-lingual transferability of ITS—particularly for languages with distinct structural properties like Chinese (high information density) or French (complex morphosyntax)—requires further validation. (3) Lack of adaptive attack exploration: We have not yet explored the framework's resilience against adaptive fine-tuning attacks, where an adversary might employ "audit-aware" objectives to intentionally mask leakage within highly redundant token clusters. Investigating the robustness of ITS against such adversarial countermeasures is left as a critical direction for future work.

Ethical Considerations

This work aims to support responsible training data detection for diffusion large language models. Our method is designed to help determine whether a given candidate sample may have been used during training or fine-tuning, which can benefit privacy protection, copyright compliance, and model accountability.

We acknowledge that training data detection methods may also be misused to probe sensitive models or infer the presence of private records. Our method does not reconstruct training data or extract private content from the model; it only provides a likelihood-based detection score for a given input sample.

This work uses publicly available open-source datasets and open-source codebases for research purposes. We reviewed and complied with the licenses and terms of use associated with these artifacts. Our use of these artifacts is consistent with their intended research and academic use. No proprietary or privately collected data were used in this study. We will release our implementation code for research and reproducibility purposes, subject to the licenses and usage conditions of the original artifacts.

References

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2023. The reversal curse: Llms trained on" a is b" fail to learn" b is a". *arXiv preprint arXiv:2309.12288*.

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.

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*.

Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. 2024. Do membership inference attacks work on large language models? In *First Conference on Language Modeling*.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495.

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.

Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schölkopf, Mrinmaya Sachan, and Taylor Berg-Kirkpatrick. 2023. Membership inference attacks against language models via neighbourhood comparison. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11330–11343.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, JUN ZHOU, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. 2025. Large language

diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2024. Detecting pretraining data from large language models. In *The Twelfth International Conference on Learning Representations*.

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*.

Cheng Wang, Yiwei Wang, Bryan Hooi, Yujun Cai, Nanyun Peng, and Kai-Wei Chang. 2025. Con-recall: Detecting pre-training data in llms via contrastive decoding. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1013–1026.

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*.

Bingyang Wen, KP Subbalakshmi, and Fan Yang. 2022. Revisiting attention weights as explanations from an information theoretic perspective. *arXiv preprint arXiv:2211.07714*.

Chengyue Wu, Hao Zhang, Shuchen Xue, Zhijian Liu, Shizhe Diao, Ligeng Zhu, Ping Luo, Song Han, and Enze Xie. 2025. Fast-dllm: Training-free acceleration of diffusion llm by enabling kv cache and parallel decoding. *arXiv preprint arXiv:2505.22618*.

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.

Jingyang Zhang, Jingwei Sun, Eric Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Frank Yang, and Hai Li. 2025. Min-k%++: Improved baseline for detecting pre-training data from large language models. In *The Thirteen International Conference on Learning Representations*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Fengqi Zhu, Zebin You, Yipeng Xing, Zenan Huang, Lin Liu, Yihong Zhuang, Guoshan Lu, Kangyu Wang, Xudong Wang, Lanning Wei, and 1 others. 2025. Llada-moe: A sparse moe diffusion language model. *arXiv preprint arXiv:2509.24389*.

A Proof of Theorems

The proof of Theorem 1:

Proof.

$$\begin{aligned}
&\operatorname{error}\Big[\hat{\mathcal{P}}(x)\Big], && (15)\\
&= \mathcal{KL}\Big(\mathcal{P}(x)\Big\|\hat{\mathcal{P}}(x)\Big), && (16)\\
&= \mathcal{KL}\Big(\mathcal{P}(x)\Big\|\prod_{i=1}^{|x|}\mathcal{P}(x_i)\Big), && (17)\\
&= \sum_x \mathcal{P}(x)\log\frac{\mathcal{P}(x)}{\prod_{i=1}^{|x|}\mathcal{P}(x_i)}, && (18)\\
&= \sum_x \mathcal{P}(x)\log\prod_{i=2}^{|x|}\frac{\mathcal{P}(x_{\leq i})}{\mathcal{P}(x_{<i})\mathcal{P}(x_i)}, && (19)\\
&= \sum_x \mathcal{P}(x)\sum_{i=2}^{|x|}\log\frac{\mathcal{P}(x_{\leq i})}{\mathcal{P}(x_{<i})\mathcal{P}(x_i)}, && (20)\\
&= \sum_{i=2}^{|x|}\sum_x \mathcal{P}(x)\log\frac{\mathcal{P}(x_{\leq i})}{\mathcal{P}(x_{<i})\mathcal{P}(x_i)}, && (21)\\
&= \sum_{i=2}^{|x|}\sum_{x_{\leq i}} \mathcal{P}(x_{\leq i})\log\frac{\mathcal{P}(x_{\leq i})}{\mathcal{P}(x_{<i})\mathcal{P}(x_i)}, && (22)\\
&= \sum_{i=2}^{|x|}\mathcal{I}\big(X_{<i};X_i\big). && (23)
\end{aligned}$$

□

The proof of Theorem 2 is similar to above, replacing all $\mathcal{P}(\cdot)$ with $\mathcal{P}(\cdot|c)$.

B Experimental Details

B.1 Model Training

Our experimental suite comprises six datasets: AG News (Zhang et al., 2015) and five subsets from the MIMIR benchmark (Duan et al., 2024) (ArXiv, Pile-CC, Wikipedia (en), PubMed Central, and GitHub). To ensure a fair comparison, we adopt the training and test split configurations proposed by Chen et al. (2026), fine-tuning the model on each full training partition. Unless otherwise specified, the ArXiv dataset serves as the representative corpus for all robustness evaluations, scaling experiments, and ablation studies.

We consider two types of models: LLaDA-8B-Base (Nie et al., 2025) and Dream-v0-Base-7B (Ye et al., 2025). The former is trained from scratch, while the latter is adapted from an AR-LLM. For the reference models used in Ratio (Watson et al.,

Method	ArXiv			Pile CC			AG News		
	AUC	T@10%	T@1%	AUC	T@10%	T@1%	AUC	T@10%	T@1%
Loss	$0.52_{\pm 0.01}$	$0.13_{\pm 0.01}$	$0.01_{\pm 0.00}$	$0.51_{\pm 0.01}$	$0.11_{\pm 0.01}$	$0.02_{\pm 0.01}$	$0.53_{\pm 0.01}$	$0.11_{\pm 0.01}$	$0.01_{\pm 0.00}$
Zlib	$0.49_{\pm 0.01}$	$0.12_{\pm 0.01}$	$0.01_{\pm 0.00}$	$0.48_{\pm 0.01}$	$0.09_{\pm 0.02}$	$0.01_{\pm 0.00}$	$0.54_{\pm 0.01}$	$0.14_{\pm 0.01}$	$0.02_{\pm 0.01}$
Min-k%	$0.52_{\pm 0.01}$	$0.13_{\pm 0.01}$	$0.01_{\pm 0.00}$	$0.51_{\pm 0.01}$	$0.11_{\pm 0.01}$	$0.02_{\pm 0.01}$	$0.53_{\pm 0.01}$	$0.11_{\pm 0.01}$	$0.01_{\pm 0.00}$
Min-k%++	$0.50_{\pm 0.01}$	$0.08_{\pm 0.01}$	$0.01_{\pm 0.00}$	$0.49_{\pm 0.01}$	$0.07_{\pm 0.02}$	$0.01_{\pm 0.00}$	$0.51_{\pm 0.01}$	$0.11_{\pm 0.01}$	$0.01_{\pm 0.00}$
DF-MIA	$0.52_{\pm 0.01}$	$0.10_{\pm 0.01}$	$0.00_{\pm 0.00}$	$0.51_{\pm 0.02}$	$0.08_{\pm 0.01}$	$0.01_{\pm 0.00}$	$0.60_{\pm 0.01}$	$0.19_{\pm 0.02}$	$0.02_{\pm 0.01}$
Ratio	$0.52_{\pm 0.01}$	$0.10_{\pm 0.01}$	$0.01_{\pm 0.01}$	$0.52_{\pm 0.01}$	$0.11_{\pm 0.02}$	$0.01_{\pm 0.00}$	$0.56_{\pm 0.01}$	$0.14_{\pm 0.01}$	$0.02_{\pm 0.00}$
SAMA	$0.65_{\pm 0.01}$	$0.23_{\pm 0.02}$	$0.05_{\pm 0.01}$	$0.74_{\pm 0.01}$	$0.38_{\pm 0.02}$	$0.10_{\pm 0.02}$	$0.65_{\pm 0.01}$	$0.22_{\pm 0.02}$	$0.00_{\pm 0.00}$
ITS (Ours)	$\mathbf{0.77}_{\pm 0.01}$	$\mathbf{0.46}_{\pm 0.03}$	$\mathbf{0.16}_{\pm 0.03}$	$\mathbf{0.87}_{\pm 0.01}$	$\mathbf{0.68}_{\pm 0.03}$	$\mathbf{0.30}_{\pm 0.06}$	$\mathbf{0.74}_{\pm 0.01}$	$\mathbf{0.33}_{\pm 0.02}$	$\mathbf{0.02}_{\pm 0.04}$

Method	PubMed Central			Wikipedia (en)			Github		
	AUC	T@10%	T@1%	AUC	T@10%	T@1%	AUC	T@10%	T@1%
Loss	$0.52_{\pm 0.01}$	$0.13_{\pm 0.01}$	$0.01_{\pm 0.00}$	$0.52_{\pm 0.01}$	$0.14_{\pm 0.01}$	$0.01_{\pm 0.00}$	$0.56_{\pm 0.01}$	$0.18_{\pm 0.02}$	$0.02_{\pm 0.01}$
Zlib	$0.48_{\pm 0.01}$	$0.10_{\pm 0.01}$	$0.01_{\pm 0.01}$	$0.49_{\pm 0.01}$	$0.10_{\pm 0.01}$	$0.01_{\pm 0.00}$	$0.55_{\pm 0.01}$	$0.19_{\pm 0.02}$	$0.03_{\pm 0.01}$
Min-k%	$0.52_{\pm 0.01}$	$0.13_{\pm 0.01}$	$0.01_{\pm 0.00}$	$0.52_{\pm 0.01}$	$0.13_{\pm 0.01}$	$0.01_{\pm 0.00}$	$0.56_{\pm 0.01}$	$0.18_{\pm 0.02}$	$0.02_{\pm 0.01}$
Min-k%++	$0.49_{\pm 0.01}$	$0.10_{\pm 0.01}$	$0.01_{\pm 0.00}$	$0.52_{\pm 0.01}$	$0.12_{\pm 0.02}$	$0.00_{\pm 0.00}$	$0.53_{\pm 0.01}$	$0.09_{\pm 0.02}$	$0.01_{\pm 0.00}$
DF-MIA	$0.50_{\pm 0.01}$	$0.11_{\pm 0.02}$	$0.01_{\pm 0.00}$	$0.51_{\pm 0.01}$	$0.11_{\pm 0.01}$	$0.01_{\pm 0.00}$	$0.59_{\pm 0.01}$	$0.16_{\pm 0.02}$	$0.03_{\pm 0.01}$
Ratio	$0.53_{\pm 0.01}$	$0.12_{\pm 0.01}$	$0.01_{\pm 0.01}$	$0.51_{\pm 0.01}$	$0.12_{\pm 0.02}$	$0.01_{\pm 0.00}$	$0.50_{\pm 0.01}$	$0.10_{\pm 0.01}$	$0.02_{\pm 0.01}$
SAMA	$0.68_{\pm 0.01}$	$0.30_{\pm 0.01}$	$0.08_{\pm 0.01}$	$0.73_{\pm 0.01}$	$0.35_{\pm 0.02}$	$0.08_{\pm 0.01}$	$0.68_{\pm 0.01}$	$0.29_{\pm 0.02}$	$0.05_{\pm 0.01}$
ITS (Ours)	$\mathbf{0.78}_{\pm 0.01}$	$\mathbf{0.46}_{\pm 0.03}$	$\mathbf{0.12}_{\pm 0.02}$	$\mathbf{0.84}_{\pm 0.01}$	$\mathbf{0.59}_{\pm 0.02}$	$\mathbf{0.25}_{\pm 0.02}$	$\mathbf{0.75}_{\pm 0.01}$	$\mathbf{0.43}_{\pm 0.03}$	$\mathbf{0.14}_{\pm 0.02}$

Table 3: Detection Performance on Dream.

2021), SAMA (Chen et al., 2026), and ITS (ours), we adopt the corresponding pre-trained versions. Fine-tuning is conducted using the AdamW optimizer, with a learning rate of 5×10^{-5} for full fine-tuning and 5×10^{-4} for LoRA, a weight decay of 0.1, a batch size of 48, and up to 4 epochs. Training is performed in bf16 precision on one NVIDIA A6000 GPU with DeepSpeed Stage 1 optimization.

B.2 Detection Inference

We follow the implementation of SAMA (Chen et al., 2026) for all experimental settings, except for the number of progressive sampling rounds T. Specifically, SAMA uses $T = 16$, whereas we adopt a substantially smaller budget of $T = 4$ in our experiments. The masking ratio increases progressively from $\rho_{\min} = 5\%$ to $\rho_{\max} = 50\%$, and $N = 128$ subsets, each containing $M = 10$ tokens, are sampled. All baselines are evaluated under the same query budget, *i.e.*, using T queries to the target model for each sample. For our method, the default diversity-promoting penalty λ is set to 0.1.

B.3 Implementation of Ablation Study on Dependency Levels

To examine how the dependency level of sampled token subsets affects detection performance, we conduct an ablation study while keeping the diversity-promoting penalty unchanged. Specifically, Line 7 in Algorithm **??** is modified as

$$u_v \leftarrow \alpha h_v + \lambda C_v. \tag{24}$$

In the default ITS setting, we use $\alpha = 1$, which encourages the selection of tokens with low cumulative dependency. For the medium-dependency setting, token affinity is ignored and tokens are sampled randomly, corresponding to $\alpha = 0$. For the high-dependency setting, we set $\alpha = -1$, which encourages the selection of token sets with high mutual affinity.

C Additional Experimental Results

C.1 Comparison on Dream Model

To further validate the scalability of our method, we evaluate ITS on the Dream 7B model (Ye et al., 2025) across six diverse datasets: ArXiv, Pile CC, AG News, PubMed Central, Wikipedia (en), and Github. As detailed in Table 3, ITS consistently achieves the best performance across all datasets and metrics. Conventional AR-based methods (e.g., Min-k%, Ratio, DF-MIA) struggle significantly on dLLMs, with AUCs often hovering near chance level (≈ 0.52). Compared to SAMA, ITS provides a substantial boost in both detection accuracy and high-precision regimes. For instance, on the Pile CC dataset, ITS improves the AUROC from 0.74 to 0.87 and triples the TPR@1% ($0.10 \rightarrow 0.30$).These results underscore that the structural bias mitigated by ITS is a universal challenge across different dLLM scales and data domains. The consistent lead in TPR@1% highlights ITS's effectiveness in low-false-positive scenarios, which is critical for real-world training data detection.

Rank	AUC		T@10%		T@1%	
	SAMA	ITS	SAMA	ITS	SAMA	ITS
1	0.48	**0.50**	**0.10**	0.09	0.01	0.01
2	0.50	**0.52**	0.09	**0.12**	0.01	0.01
4	0.51	**0.53**	0.11	**0.13**	0.01	0.01
8	0.52	**0.56**	0.13	0.13	0.01	**0.02**
16	0.52	**0.58**	0.10	**0.17**	0.02	0.02
32	0.56	**0.64**	0.15	**0.24**	0.01	**0.03**
64	0.60	**0.70**	0.21	**0.30**	0.02	**0.06**
128	0.59	**0.70**	0.18	**0.31**	0.04	**0.06**

Table 4: Detection performance with different LoRA ranks. The target modules are attention layers.

Rank	AUC		T@10%		T@1%	
	SAMA	ITS	SAMA	ITS	SAMA	ITS
1	0.63	**0.76**	0.20	**0.40**	0.03	**0.08**
2	0.69	**0.85**	0.28	**0.54**	0.06	**0.23**
4	0.68	**0.86**	0.27	**0.59**	0.06	**0.25**
8	0.73	**0.90**	0.34	**0.68**	0.10	**0.24**
16	0.76	**0.93**	0.40	**0.76**	0.08	**0.43**
32	0.83	**0.97**	0.52	**0.90**	0.10	**0.62**
64	0.88	**0.99**	0.65	**0.96**	0.23	**0.78**
128	0.93	**1.00**	0.79	**0.99**	0.40	**0.93**

Table 5: Detection performance with different LoRA ranks. The target modules are MLP layers.

C.2 Detailed Results for LoRA Experiments

In this section, we provide the comprehensive numerical breakdown of the LoRA experiments. These experiments evaluate the sensitivity of training data detection when targeting different model components (Attention vs. MLP) across various Low-Rank Adaptation (LoRA) ranks $r \in \{1, 2, 4, 8, 16, 32, 64, 128\}$.

Experimental Setup. For each rank r, the model is fine-tuned on the target dataset using LoRA applied exclusively to either the MLP or Attention modules. During detection, ITS and the SAMA baseline are applied to the same modules that were adapted during training. The sampling rounds T and other hyperparameters are kept identical to ensure a fair comparison.

Full Numerical Analysis. Tables 4 and 5 list the AUROC, TPR@10%, and TPR@1% for both methods.

1. **Attention Layers (Table 4):** Detection performance in Attention layers remains relatively low even as rank increases. ITS consistently maintains a lead over SAMA, particularly in TPR metrics at higher ranks, but the overall signal remains sparse, supporting the view that Attention primarily manages contextual routing rather than long-term knowledge storage.

2. **MLP Layers (Table 5):** In contrast, detection signals in MLP layers are remarkably strong. ITS demonstrates exceptional efficiency, achieving an AUC of 0.8394 at $r = 2$, a performance level that SAMA does not reach until $r = 32$. At $r = 128$, ITS achieves a near-perfect AUC of 0.9984 (rounded to 1.00 in the table), confirming that MLPs are the primary repository for internalized training signatures.

Retain Ratio	AUC		T@10%		T@1%	
	SAMA	ITS	SAMA	ITS	SAMA	ITS
30%	0.59	**0.63**	0.17	**0.20**	0.01	**0.02**
50%	0.64	**0.76**	0.21	**0.40**	0.05	**0.07**
75%	0.67	**0.81**	0.28	**0.47**	0.05	**0.15**
90%	0.67	**0.83**	0.27	**0.53**	0.06	**0.22**
100%	0.70	**0.88**	0.34	**0.64**	0.06	**0.26**

Table 6: Detection performance with different text retain ratios.

C.3 Robustness Evaluation

We evaluate detection performance on partial sequences to simulate real-world copyright auditing of short snippets. Sequences are truncated from the end to retain $30\%, 50\%, 75\%$, and 90% of the original tokens. Table 6 provides the full AUROC breakdown. The results show that ITS scales more effectively than SAMA as the observation window expands, confirming that our method captures training signals more accurately from increased context.

Epoch	AUC		T@10%		T@1%	
	SAMA	ITS	SAMA	ITS	SAMA	ITS
1	0.47	**0.50**	0.07	**0.09**	0.01	**0.02**
2	0.52	**0.56**	0.11	**0.16**	0.01	**0.02**
3	0.63	**0.76**	0.21	**0.38**	0.02	**0.12**
4	0.68	**0.86**	0.28	**0.61**	0.05	**0.26**
5	0.75	**0.93**	0.38	**0.78**	0.11	**0.38**
6	0.82	**0.97**	0.51	**0.92**	0.11	**0.60**
7	0.86	**0.98**	0.62	**0.95**	0.18	**0.75**
8	0.90	**0.99**	0.70	**0.98**	0.27	**0.88**
9	0.92	**1.00**	0.78	**0.99**	0.30	**0.94**
10	0.93	**1.00**	0.79	**0.99**	0.43	**0.96**

Table 7: Detection performance with different fine-tuning epochs.

T	AUC		T@10%		T@1%	
	SAMA	ITS	SAMA	ITS	SAMA	ITS
1	0.71	**0.77**	0.34	**0.41**	0.07	**0.09**
2	0.72	**0.84**	0.34	**0.58**	0.07	**0.18**
4	0.70	**0.88**	0.34	**0.64**	0.06	**0.26**
8	0.71	**0.89**	0.33	**0.66**	0.08	**0.33**
12	0.72	**0.91**	0.36	**0.75**	0.06	**0.33**
16	0.72	**0.91**	0.31	**0.74**	0.07	**0.43**

Table 8: Detection performance with different sampling rounds T.

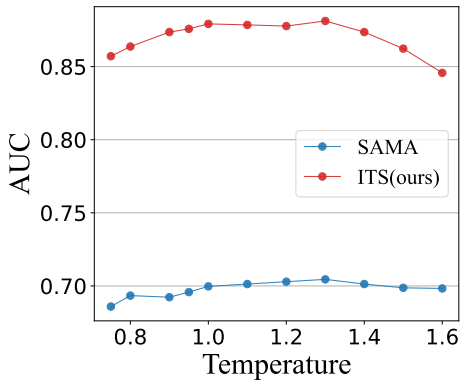


Figure 7: Caption

C.4 Scaling Analysis

Training Epochs: Table 7 details the AUROC across 10 epochs. The rapid saturation of ITS performance (reaching >0.98 by epoch 7) demonstrates its effectiveness in auditing models that have moderately internalized protected data.

Sampling Rounds: Table 8 compares the impact of sampling rounds T. The results highlight that ITS is not only more accurate but also more sample-efficient, as ITS at $T = 2$ already outperforms SAMA at $T = 16$.

λ	AUC	T@10%	T@1%
0.00	0.84	0.59	0.17
0.05	0.87	0.63	0.24
0.10	0.88	0.64	0.26
0.15	0.87	0.61	0.25
0.20	0.87	0.63	0.25
0.30	0.87	0.64	0.25
0.40	0.86	0.62	0.28
0.50	0.86	0.61	0.24
1.00	0.86	0.60	0.27
1.50	0.86	0.60	0.28
2.00	0.86	0.60	0.29
3.00	0.86	0.62	0.27

Table 9: Detection performance with different diversity-promoting penalty λ.

C.5 Detailed Results for Penalty λ

Detailed Analysis of the Diversity-promoting Penalty. The penalty λ is specifically integrated to discourage the redundant selection of overlapping token sets, thereby maximizing the information gain per query. Table 9 demonstrates the efficacy of this strategy. The AUC and T@10% scores peak between $\lambda = 0.05$ and $\lambda = 0.15$. While the AUC slightly retracts to 0.86 for larger λ values, it remains remarkably stable even as the penalty increases to 3.00. This plateau indicates that while the balance between independence and coverage is sensitive at lower scales, the ITS framework is not overly dependent on meticulous hyperparameter tuning.

C.6 Temperature Scaling

Temperature scaling is often used to obfuscate signals by smoothing output distributions with temperature t:

$$p_i = \frac{\exp(z_i/t)}{\sum_j \exp(z_j/t)}. \tag{25}$$

However, our results in Table 10 and Figure 7 show that ITS is remarkably resilient to this perturbation, maintaining a dominant performance lead over SAMA across all tested temperatures $t \in [0.75, 1.60]$. As shown in Table 10, ITS consistently achieves an AUC between 0.85 and 0.88, whereas SAMA plateaus at approximately 0.70. Even under aggressive scaling ($T = 1.5$), ITS retains a TPR@1%FPR of 0.28, which has a four-fold improvement over SAMA's 0.07. The stability of ITS suggests that by minimizing mutual dependency, our method captures intrinsic memorization signals that are largely invariant to monotonic logit transformations. While temperature scaling increases output entropy, the relative information captured by our independent token sets remains intact. Consequently, temperature scaling fails as a viable defense against ITS, as our method continues to provide a robust and discriminative signal where heuristic baselines falter.

t	AUC		T@10%		T@1%	
	SAMA	ITS	SAMA	ITS	SAMA	ITS
0.75	0.69	**0.86**	0.29	**0.62**	0.06	**0.26**
0.80	0.69	**0.86**	0.31	**0.60**	0.06	**0.23**
0.90	0.69	**0.87**	0.30	**0.64**	0.08	**0.26**
0.95	0.70	**0.88**	0.30	**0.64**	0.07	**0.26**
1.00	0.70	**0.88**	0.34	**0.64**	0.06	**0.26**
1.10	0.70	**0.88**	0.31	**0.64**	0.07	**0.26**
1.20	0.70	**0.88**	0.30	**0.67**	0.08	**0.27**
1.30	0.70	**0.88**	0.30	**0.66**	0.07	**0.28**
1.40	0.70	**0.87**	0.30	**0.63**	0.07	**0.22**
1.50	0.70	**0.86**	0.30	**0.62**	0.07	**0.28**
1.60	0.70	**0.85**	0.31	**0.58**	0.05	**0.23**

Table 10: Detection performance with different temperature.

D Use of AI Assistants

AI-assisted tools were used solely for language polishing and improving the clarity, grammar, and readability of the manuscript. The AI assistant was not used to generate scientific ideas, design the study, perform data analysis, interpret results, or draw conclusions. All content, arguments, experimental results, and conclusions were developed and verified by the authors, who take full responsibility for the final manuscript.