Delta-MIA: Measuring Membership Inference Attacks in Large Language Models via self-Contrast Framework

Anonymous ACL submission

Abstract

Membership inference attack (MIA) underpins privacy risk assessment, provenance, and compliance for large language models (LLMs). However, existing MIA evaluations are largely observational, often conflating membership with distribution shift while assuming access to proprietary corpora. We introduce Delta-MIA, an interventional self-contrast framework that reframes MIA evaluation as a fixed-record exposure intervention. By constructing matched model trajectories with and without exposure to the same target records, Delta-MIA controls member–non-member distribution shift by construction and enables paired sample-level analysis. We further introduce four paired diagnostic metrics: above-diagonal ratio (ADR), mean vertical distance (MVD), explained variance ratio (EVR), and noise-normalized MVD (n-MVD), which characterize directionality, separation, common-mode score variation, and normalized membership signal strength. Re-evaluating 9 MIA methods, several remain robust once the shift is removed, while others decline markedly. We show theoretically and empirically that larger models retain membership signals more durably through continual pre-training, a finding grounded in the geometry of gradient orthogonality in high-dimensional parameter spaces. Delta-MIA enables bias-free, interpretable, and transferable MIA evaluation.

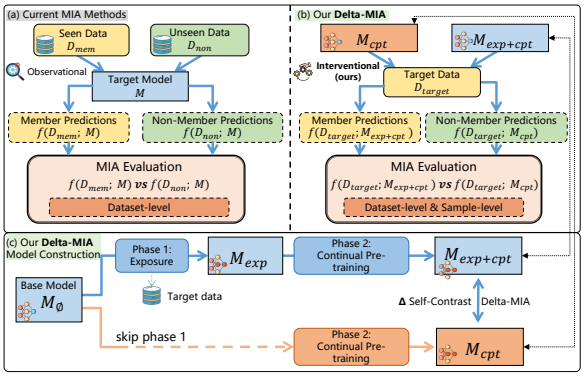


Figure 1: Schematic comparison of observational and interventional paradigms for MIA evaluation. **(a) Observational evaluation:** Existing pipelines typically compare separate member and non-member sets under a fixed target model, so aggregate scores can conflate membership evidence with distributional differences. **(b) Delta-MIA:** Delta-MIA holds the target records fixed and evaluates them under two matched model trajectories: $M_{\text{exp+cpt}}$, where D_{target} is exposed, and M_{cpt}, where it is not. This fixed-record contrast produces paired member/non-member scores for both dataset-level evaluation and sample-level diagnosis. **(c) Model construction:** Both trajectories start from the same base model $M_{\varnothing}$. The exposed trajectory undergoes target-data exposure followed by continual pre-training (CPT), while the unexposed trajectory skips exposure and applies the same CPT procedure. Thus, target-data exposure serves as the controlled intervention, enabling Δ self-contrast while controlling member/non-member distribution shift by construction.

1 Introduction

Membership inference attacks (MIAs) determine whether a given example was used to train a target model, and are widely used to assess memorization, data provenance, and compliance risks in large language models (LLMs) (Shokri et al., 2017; Shi et al., 2024). Reliable MIA evaluation is therefore important for privacy auditing, data governance, and emerging procedures such as machine unlearning (Mireshghallah et al., 2022a; Fry, 2025; Geng et al., 2025). At its core, such evaluation asks a counterfactual question: *how does a record receive different membership evidence if the model has been exposed to it during training?* Existing benchmarks rarely answer this question directly, because they typically compare different member and non-member records rather than different exposure states of the same records.

As shown in Fig. 1 (a), most current LLM MIA evaluations follow an **observational paradigm**: a fixed target model is evaluated on a member

set and a non-member set constructed from separate data sources. **(1) distributional confounding**: member and non-member examples may differ in time period (Shi et al., 2024; Meeus et al., 2024), domain (Chalkidis et al., 2023), or latent text difficulty, allowing attacks to exploit distributional shortcuts rather than membership evidence itself (Meeus et al., 2025). **(2) limited interpretability**: evaluations are often summarized by dataset-level metrics on a static pretrained model (Carlini et al., 2021; Song and Mittal, 2020), obscuring how individual examples contribute to attack success or failure. **(3) limited transferability**: some evaluation settings rely on access to, or assumptions about, the target model's training corpus (Shi et al., 2024), which is typically unavailable for frontier LLMs (Hardinges et al., 2024). Moreover, because membership signals can be forgotten during pre-training (Liao et al., 2025), a static observational comparison cannot determine whether a weak attack result reflects an ineffective method or a membership signal that has already decayed. Together, these limitations motivate an evaluation paradigm that *holds the evaluated records fixed while directly controlling their exposure status*.

Existing efforts have attempted to reduce distributional confounding, but they still largely remain observational. For example, MIMIR (Duan et al., 2024) constructs train-test splits from open-source LLM corpora and uses n-gram overlap for deduplication and shift estimation. While useful, such lexical controls cannot directly answer the fixed-record question: how would the attack score of the same example change if the model had, or had not, been exposed to it?

To address these challenges, we introduce **Delta-MIA** (Δ-MIA), an MIA evaluation framework under an **Interventional Paradigm** (Fig. 1 (b)). Rather than comparing different member and non-member records, Delta-MIA compares two exposure states of the same target records. Starting from a shared base model, we construct two matched trajectories: an exposed model that sees the target data before continual pre-training, and an unexposed model that skips this stage but undergoes the same continual pre-training. This fixed-record contrast controls member–non-member distribution shift by construction and enables paired sample-level diagnostics through EVR, MVD, ADR, and n-MVD.

Using Δ-MIA, we re-evaluate nine representative MIA methods across model scales and data domains. Under this controlled contrast, some methods preserve clear membership signals, whereas others show weaker or less stable behavior than suggested by observational evaluations. The paired sample-level view further explains otherwise hidden behaviors, including the skewed score distribution of ReCaLL (Xie et al., 2024) and the directional instability of Con-ReCaLL (Wang et al., 2025). Finally, Δ-MIA enables us to measure how membership signals survive continual pre-training, showing that larger models retain such signals more durably under a fixed CPT budget. Together, these results establish Δ-MIA as a new interventional paradigm for shift-controlled, interpretable, and transferable MIA evaluation in open-weight LLMs. *All models, data, and code will be made publicly available upon publication.*

2 Δ-MIA Evaluation Framework

2.1 Problem Definition

A membership inference attack (MIA) on a target model $\mathcal{M}$ aims to determine if a given sample x belongs to its training dataset $\mathcal{D}$, via a scoring function $f(x; \mathcal{M})$ to quantify the likelihood of membership. A prediction is then made by comparing the score against a threshold τ:

$$\text{prediction}(x, \mathcal{M}) = \begin{cases} 1, & \text{if } f(x; \mathcal{M}) \geq \tau, \\ 0, & \text{if } f(x; \mathcal{M}) < \tau. \end{cases} \quad (1)$$

2.2 Evaluation Procedure

Following Duan et al. (2024); Shi et al. (2024), we implement Δ-MIA using the Pythia (Biderman et al., 2023) model suite (410M to 6.9B parameters) and its public training corpus, the Pile dataset (Gao et al., 2021). This setup provides ground-truth membership knowledge required by certain MIA methods (e.g., Con-ReCaLL (Wang et al., 2025) requires member data as prefix), while Δ-MIA itself requires no access to the target model's original training corpus: for any open-weight model with a known release date, timestamp-verified held-out data and open-source pre-training datasets suffices to instantiate the framework.

Model Construction. We define a target dataset by sampling 1,500 instances from the Pile test split (500 each from Pile-CC, Wikipedia, and PubMed Abstracts). We construct four model checkpoints:

- $\mathcal{M}_{\varnothing}$: the original Pythia base model, unseen by target data.
- $\mathcal{M}_{\text{exp}}$: trained on the 1,500 target instances for one epoch (data exposure stage).

- $\mathcal{M}_{\text{cpt}}$: trained only on 1M instances from the Pile training split (continual pre-training, no exposure to target data).
- $\mathcal{M}_{\text{exp+cpt}}$: sequentially trained through both stages above.

The scale of continual pre-training (1M instances, exceeding 1B tokens) is chosen to approximate a realistic CPT setting for models up to 6.9B parameters, moving the target samples sufficiently past the short-term memorization spike that occurs immediately after data exposure to maintain stable model behavior.

Self-Contrast Evaluation. The core evaluation contrasts $\mathcal{M}_{\text{cpt}}$ against $\mathcal{M}_{\text{exp+cpt}}$ on the identical 1,500 target instances. Since the evaluated records are identical across the two models, this contrast controls member–non-member distribution shift by construction and allows score differences to be interpreted as exposure-induced membership signals. We note that although $\mathcal{M}_{\text{cpt}}$ and $\mathcal{M}_{\text{exp+cpt}}$ share the same CPT data and budget, their CPT trajectories begin from different initializations ($\mathcal{M}_{\varnothing}$ and $\mathcal{M}_{\text{exp}}$ respectively); we verify empirically that this path-dependent difference induces negligible behavioral change on non-target data (Appendix A). In Appendix F, we further relate the interventional contrast to the observational evaluation protocol.

2.3 Metric Design

Raw MIA scores from $\mathcal{M}_{\text{cpt}}$ (non-member) and $\mathcal{M}_{\text{exp+cpt}}$ (member) are preprocessed by removing outliers beyond the 1st and 99th percentiles and applying min-max normalization, yielding comparable score pairs $(s_i^{\text{non}}, s_i^{\text{mem}})$ for the i-th of K filtered samples.

Above-Diagonal Ratio (ADR) measures fundamental detection capability as the proportion of samples where the member score exceeds the non-member score:

$$\text{ADR} = \frac{1}{K}\sum_{i=1}^{K} \mathbb{I}(s_i^{\text{mem}} > s_i^{\text{non}}). \tag{2}$$

A low ADR indicates frequent pairwise inversions and unstable membership directionality.

Mean Vertical Distance (MVD) quantifies the average discrimination strength:

$$\text{MVD} = \frac{1}{K}\sum_{i=1}^{K} (s_i^{\text{mem}} - s_i^{\text{non}}). \tag{3}$$

Explained Variance Ratio (EVR) quantifies confounding noise via PCA (Maćkiewicz and Ratajczak, 1993) on the score pairs, where noise includes difference of intrinsic text complexity among samples. Forming data matrix $\mathbf{X} \in \mathbb{R}^{K\times 2}$ with rows $[s_i^{\text{non}}, s_i^{\text{mem}}]$, EVR is the variance fraction explained by the first principal component:

$$\text{EVR} = \frac{\lambda_1}{\lambda_1 + \lambda_2}, \tag{4}$$

where $\lambda_1 \geq \lambda_2$ are the eigenvalues of the covariance matrix. A high EVR indicates score variance dominated by sample-difficulty noise; a low EVR indicates the membership signal is relatively significant as the scores are more clustered rather than linear.

Noise-Normalized MVD (n-MVD) combines signal strength and noise level into a single signal-to-noise ratio:

$$\text{n-MVD} = \frac{\text{MVD}}{\text{EVR}}. \tag{5}$$

Dataset-Level Metrics. We additionally report AUC and TPR@low FPR for comparability with prior work (Carlini et al., 2021; Mireshghallah et al., 2022a; Shi et al., 2024).

Detailed dataset-level and sample-level analyses are presented in Sec. 4.1 and Sec. 4.2, respectively.

3 Experimental Settings

We standardize MIA scores so that higher values indicate a greater likelihood of membership, without altering each method's intrinsic mechanism. Nine representative MIA methods are evaluated (hyperparameter details in Appendix B.2):

1. **Loss** (Yeom et al., 2018) uses the log-likelihood (LL) of the sample under the target model $\mathcal{M}$: $f(\mathbf{x}; \mathcal{M}) = \text{LL}(x; \mathcal{M})$
2. **Ref** (Carlini et al., 2020) calibrates the Loss score by subtracting the log-likelihood from a reference model ($\mathcal{M}_{\text{ref}}$): $f(x; \mathcal{M}) = \text{LL}(x; \mathcal{M}) - \text{LL}(x; \mathcal{M}_{\text{ref}})$
3. **Zlib** (Carlini et al., 2020) uses the Loss score normalized by the compressed size of the sample via the zlib algorithm, which serves as a proxy for sample complexity: $f(x; \mathcal{M}) = \text{LL}(x; \mathcal{M})/\text{zlib}(x)$
4. **Neighborhood** (Mattern et al., 2023) uses the difference between the log-likelihood of the target sample and the average log-likelihood of its syntactically perturbed "neighbor" samples: $f(x; \mathcal{M}) = \text{LL}(x; \mathcal{M}) - \frac{1}{n}\sum_{i=1}^{n} \text{LL}(x_i'; \mathcal{M}_{\text{ref}})$

Method	Pythia-410M		Pythia-1B		Pythia-2.8B		Pythia-6.9B	
	AUC	T@5%F	AUC	T@5%F	AUC	T@5%F	AUC	T@5%F
Loss	0.529	0.059	0.544	0.059	0.568	0.063	0.596	0.067
Zlib	0.522	0.068	0.533	0.087	0.550	0.104	0.571	0.120
Neighborhood	0.533	0.061	0.551	0.063	0.583	0.072	0.615	0.100
Con-ReCaLL	0.527	0.061	0.541	0.071	0.574	0.069	0.569	0.082
DC-PDD	0.532	0.059	0.549	0.063	0.580	0.076	0.615	0.100
Min-K%	0.538	0.058	0.555	0.062	0.586	0.072	0.619	0.077
ReCaLL	0.555	0.064	0.585	0.059	0.612	0.066	0.642	0.073
Ref	**0.591**	**0.084**	**0.614**	**0.092**	0.641	**0.124**	0.677	**0.155**
Min-K%++	0.571	0.065	0.611	0.076	**0.663**	0.091	**0.716**	0.112

Table 1: Dataset-level results on the Δ-MIA Benchmark. We report AUC and TPR at 5% FPR (T@5%F) for MIA methods evaluated on the Pythia model family (410M, 1B, 2.8B, and 6.9B). Higher is better. The best score in each column is **bold**; the second best is underlined.

Method	Pile-CC		PubMed		Wikipedia	
	AUC	T@5%F	AUC	T@5%F	AUC	T@5%F
Loss	0.576	0.071	0.582	0.102	0.582	0.091
Zlib	0.545	0.075	0.541	0.084	0.586	0.103
Neighborhood	0.585	0.073	0.586	0.120	0.577	0.065
Con-ReCaLL	0.573	0.088	0.621	0.098	0.574	0.081
DC-PDD	0.587	0.088	0.586	0.086	0.600	0.067
Min-K%	0.591	0.094	0.600	0.132	0.602	0.086
ReCaLL	0.624	0.094	0.650	0.128	0.583	0.074
Ref	0.642	0.117	0.647	0.159	0.646	0.096
Min-K%++	**0.661**	**0.119**	**0.676**	**0.216**	**0.669**	**0.110**

Table 2: Dataset-level results on the Δ-MIA Benchmark for the Pythia-2.8B model across three domains (Pile-CC, PubMed Abstracts, Wikipedia (en)). We report AUC and TPR at 5% FPR (T@5%F); higher is better. The best score in each column is **bold**; the second best is underlined.

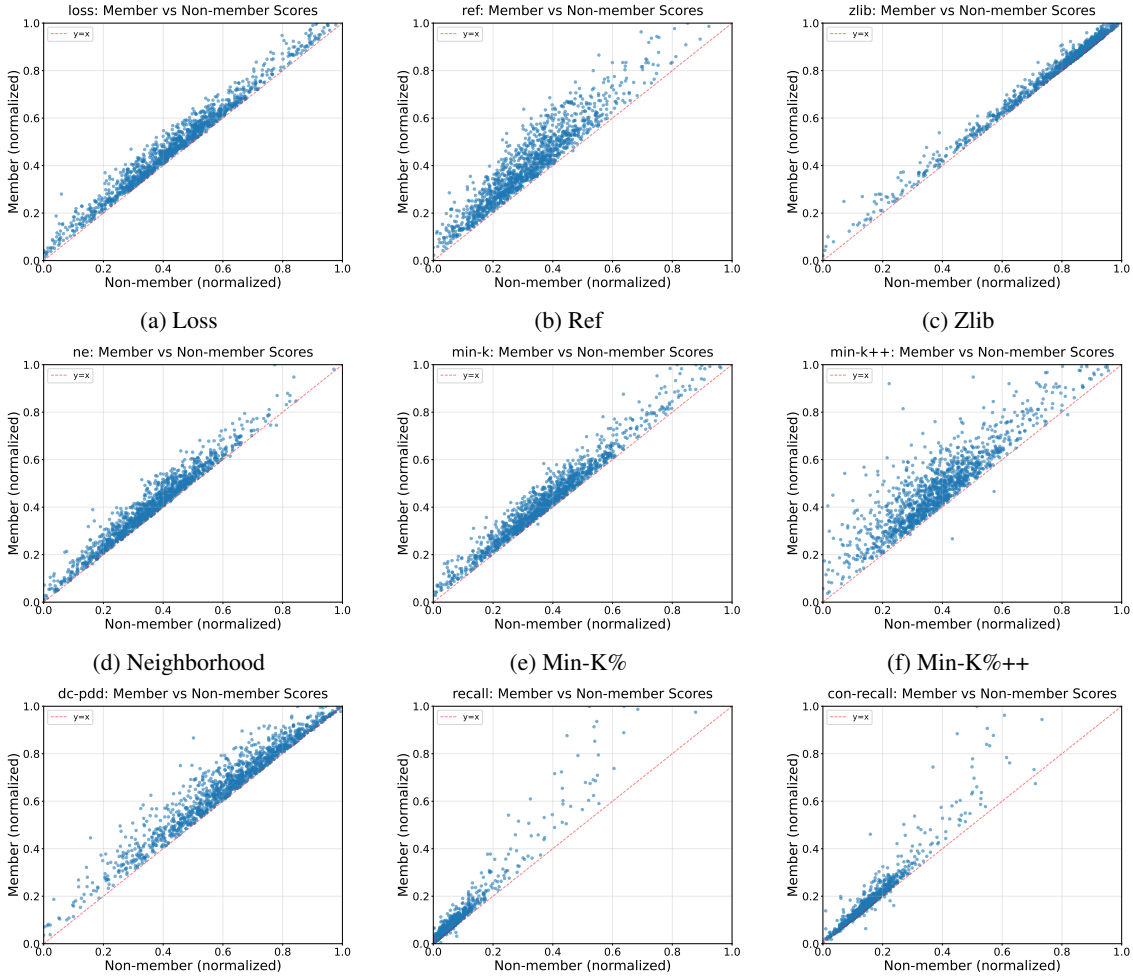
5. **Min-K%** (Shi et al., 2024) uses the average log-likelihood of the tokens with the lowest K% predicted probabilities within the sample: $f(x;\mathcal{M}) = (1/|\text{min-}k(x)|)\sum_{x_i\in\text{min-}k(x)} \log p(x_i|x_{<i})$
6. **Min-K%++** (Zhang et al., 2025) enhances Min-K% by applying z-score normalization to the log-probability of each token before averaging: $f(x;\mathcal{M}) = (1/|\text{min-k}(x)|)$ $\sum_{x_t\in\text{min-k}(x)}(\log p(x_t|x_{<t}) - \mu_{x_{<t}})/\sigma_{x_{<t}}$
7. **DC-PDD** (Zhang et al., 2024) measures the divergence between the model predictive distribution and an estimated token frequency distribution from an external corpus D': $f(x;\mathcal{M}) = (1/|x|)\sum_{x_i\in x} p(x_i;\mathcal{M}) \log p(x_i;D')$
8. **ReCaLL** (Xie et al., 2024) uses the ratio of the log-likelihood of the sample conditioned on a non-member prefix ($P_{\text{non-member}}$) to its unconditional log-likelihood: $f(x;\mathcal{M}) = \text{LL}(x|P_{\text{non-member}};\mathcal{M})/\text{LL}(x;\mathcal{M})$
9. **Con-ReCaLL** (Wang et al., 2025). This method extends ReCaLL by incorporating a contrastive term based on the log-likelihood conditioned on a member prefix (P_{member}): $f(x;\mathcal{M}) = (\text{LL}(x|P_{\text{non-member}};\mathcal{M}) - \gamma \cdot \text{LL}(x|P_{\text{member}};\mathcal{M}))/\text{LL}(x;\mathcal{M})$

4 Main Results

4.1 Dataset-Level Analysis: Model Scales and Data Domains

We evaluate nine MIA methods on the Δ-MIA benchmark across model scales (Sec. 3) and data domains (Sec. 3).

Analysis Across Model Scales. First, AUC and TPR@5%FPR increase monotonically with model



(a) Loss (b) Ref (c) Zlib

(d) Neighborhood (e) Min-K% (f) Min-K%++

(g) DC-PDD (h) ReCaLL (i) Con-ReCaLL

Figure 2: Sample-level scatter plots for Pythia-2.8B across 9 attack methods. Each subfigure shows the normalized member score (y-axis) versus the corresponding non-member score (x-axis), with the identity line $y = x$ for reference. Points above the line indicate correct membership reflection.

scale across nearly all methods, suggesting that larger models retain more pronounced membership signals. Second, a clear performance hierarchy exists: Ref leads in TPR@5%FPR across all scales, while Min-K%++ achieves the highest AUC on larger models (2.8B and 6.9B), establishing itself as the strongest reference-free method at scale. ReCaLL achieves competitive AUC but consistently lower TPR@5%FPR. Third, Con-ReCaLL underperforms relative to its reported effectiveness in prior benchmarks (Wang et al., 2025), suggesting that its prior gains may be sensitive to distributional confounding or prefix-related effects. We provide a mechanistic explanation in Sec. 4.2.

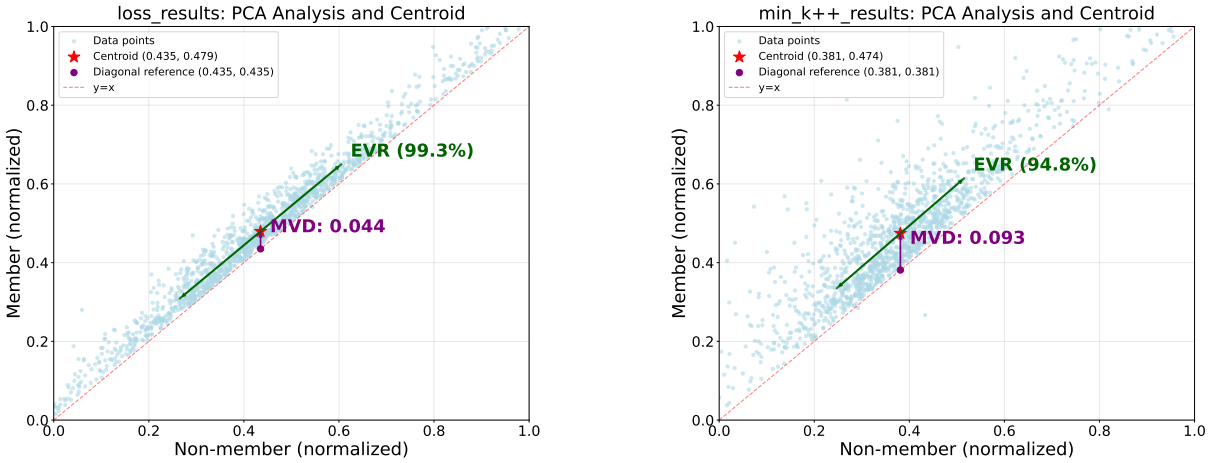
Analysis Across Data Domains. MIA efficacy varies considerably across domains (Sec. 3). ReCaLL shows the strongest domain sensitivity, with notably higher AUC on PubMed than Wikipedia, reflecting its dependence on in-domain prefix calibration. Min-K%++ is the most domain-robust top performer, maintaining leading AUC across all three domains. Further per-scale results are in Appendix C.

4.2 Sample-Level Analysis: Qualitative and Quantitative Diagnostics

Qualitative Analysis via Scatter Plots. Fig. 2 visualizes normalized score pairs for Pythia-2.8B. Three patterns emerge. (1) Most methods concentrate points above the diagonal, indicating that most methods capture a positive exposure effect. (2) Loss exhibits a tight linear pattern just above the diagonal, its signal dominated by sample-difficulty noise. (3) ReCaLL's scatter is markedly different: points cluster heavily in the low-score region with only sparse coverage elsewhere, producing a skewed distribution that explains its high AUC but

Method	410M		1B		2.8B		6.9B	
	ADR(%)	n-MVD	ADR(%)	n-MVD	ADR(%)	n-MVD	ADR(%)	n-MVD
Loss	100.00	0.019	100.00	0.029	99.93	0.045	99.85	0.064
Zlib	100.00	0.009	100.00	0.013	99.93	0.021	99.85	0.029
Neighborhood	98.38	0.019	99.63	0.029	99.63	0.044	99.63	0.069
Con-ReCaLL	95.79*	0.012	98.52	0.017	98.15	0.022	91.42*	0.027
DC-PDD	98.74	0.022	99.85	0.034	99.63	0.055	99.63	0.081
Min-K%	99.93	0.023	100.00	0.034	99.93	0.051	99.85	0.073
ReCaLL	99.63	0.018	99.70	0.022	99.55	0.025	99.18	0.030
Ref	100.00	**0.053**	100.00	**0.069**	99.93	0.087	99.85	0.115
Min-K%++	99.18	0.040	99.48	0.063	99.70	**0.098**	99.78	**0.142**

Table 3: Sample-level results on the Δ-MIA Benchmark. Performance of MIA methods across the Pythia family (410M, 1B, 2.8B, 6.9B) are evaluated with ADR and n-MVD. Values of ADR substantially below 98% are denoted by an asterisk*, indicating anomalous detection behavior. The highest value of n-MVD in each column is **bold**; the second highest is underlined.



(a) Loss baseline: High Noise, Weak Signal

(b) Min-K%++: Low Noise, Strong Signal

Figure 3: PCA visualization of sample-level evaluations for MIA methods. Compared to **(a) the Loss baseline**, **(b) Min-K%++** demonstrates superior performance by effectively reducing sample-induced noise (lower EVR) while amplifying the membership signal (higher MVD). This validates the use of n-MVD as a signal-to-noise metric for MIA evaluation.

low TPR@5%FPR; a geometric analysis is in Appendix E.

Quantitative Analysis with Diagnostic Metrics. Sec. 4.1 provides sample-level results. First, n-MVD increases with model scale for most methods, suggesting that larger models leave more distinct membership traces; Ref leads on smaller models while Min-K%++ overtakes on larger ones (2.8B and 6.9B), reflecting the increasing effectiveness of z-score normalization at scale. Second, Con-ReCaLL stands out anomalously as the only method with substantially degraded ADR, dropping to 91.42% at 6.9B. This reflects a structural issue: its score function is not well-defined in the sense that the two conditional log-likelihoods in the numerator, $\mathrm{LL}(x|P_{\text{non-mem}})$ and $\mathrm{LL}(x|P_{\text{mem}})$, admit no strict ordering — their relative magnitude is governed by uncontrolled factors including the intrinsic complexity of each prefix and its distributional proximity to the target text. When their weighted difference changes sign, the score direction inverts, undermining sample-level discriminability. A formal analysis is in Appendix D.

5 Analysis of Membership Signal Retention Across Model Scales

The Δ-MIA framework naturally enables a controlled analysis of how membership signals evolve through continual pre-training. By comparing

MIA performance under the $(\mathcal{M}_{\varnothing}, \mathcal{M}_{\text{exp}})$ contrast against the $(\mathcal{M}_{\text{cpt}}, \mathcal{M}_{\text{exp+cpt}})$ contrast, we can directly measure how much of the membership signal survives the CPT process, and how this decay interacts with model scale.

5.1 Geometric Explanation

Let θ_{exp} and $\theta_{\varnothing}$ denote the parameters of $\mathcal{M}_{\text{exp}}$ and $\mathcal{M}_{\varnothing}$ respectively. We define the **membership signal vector** as $\Delta\theta = \theta_{\text{exp}} - \theta_{\varnothing} \in \mathbb{R}^d$, which encodes the directional change in parameter space induced by training on the target data. During CPT, each gradient step g_t erodes this signal in proportion to its alignment with $\Delta\theta$, quantified by the cosine similarity:

$$\rho_t = \frac{\langle g_t, \Delta\theta \rangle}{\|g_t\| \cdot \|\Delta\theta\|} \tag{6}$$

When $|\rho_t| \approx 0$, the update operates in a subspace nearly orthogonal to $\Delta\theta$, leaving the membership signal largely intact. This is directly analogous to the principle underlying Orthogonal Gradient Descent (Farajtabar et al., 2020): gradient updates orthogonal to directions encoding prior knowledge produce minimal interference with that knowledge.

Since CPT trains on general-purpose text unrelated to the target data, we treat the gradients g_t as approximately unstructured vectors in $\mathbb{R}^d$ relative to $\Delta\theta$, carrying no systematic directional alignment with $\Delta\theta$ induced by the CPT data distribution. In high-dimensional spaces, two independent random vectors tend toward orthogonality, with their cosine similarity concentrating around zero. As model scale and thus d grow, each CPT gradient step becomes geometrically less capable of projecting onto $\Delta\theta$, and cumulative signal erosion over T steps diminishes accordingly. This geometric view predicts that, under a fixed CPT budget, larger models should retain membership signals more durably — a prediction we test empirically below.

5.2 Empirical Validation: Gradient–Signal Alignment Across Scales

To directly test the orthogonality hypothesis, we sample 10,000 instances from the CPT corpus and compute per-sample gradient vectors with respect to $\mathcal{M}_{\varnothing}$ parameters. Fig. 4 reports the distribution of $|\rho_t|$ across the four Pythia model scales. Across all scales, the distributions are heavily concentrated near zero, confirming that CPT gradients are largely orthogonal to $\Delta\theta$. Critically, as model size increases from 410M to 6.9B, the distribution

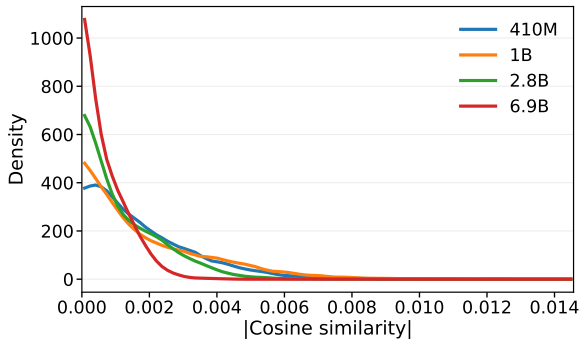


Figure 4: Distribution of $|\rho_t|$ between per-sample CPT gradients and $\Delta\theta$ across model scales. As parameter dimensionality increases, the distribution concentrates closer to zero, indicating that CPT gradient updates become increasingly orthogonal to the membership signal direction.

shifts progressively closer to zero with decreasing spread, providing direct empirical support for the scaling argument: in larger models, CPT interferes less with the parameter directions encoding target data membership.

5.3 MIA Performance Under Signal Decay

To quantify how membership signals evolve through CPT, we evaluate all nine MIA methods under both the $(\mathcal{M}_{\varnothing}, \mathcal{M}_{\text{exp}})$ and $(\mathcal{M}_{\text{cpt}}, \mathcal{M}_{\text{exp+cpt}})$ contrasts and measure the resulting performance change. We report AUC in Fig. 5 and the n-MVD decay rate in Sec. 5.3. We use n-MVD rather than AUC for this purpose as it more linearly reflects the underlying signal strength, making it better suited for quantifying decay.

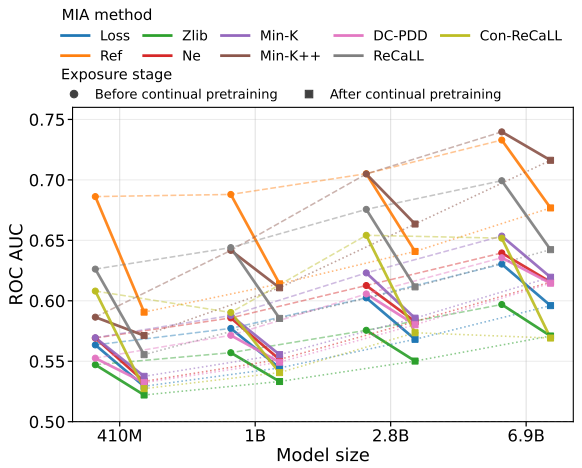


Figure 5: ROC AUC of nine MIA methods evaluated at $\mathcal{M}_{\text{exp}}$ and $\mathcal{M}_{\text{exp+cpt}}$ across four Pythia model scales.

Three findings emerge from this analysis. First,

Method	410M	1B	2.8B	6.9B
Loss	54.2%	42.0%	34.3%	26.5%
Zlib	55.3%	45.1%	36.0%	31.1%
Neighborhood	53.8%	42.5%	30.0%	16.8%
Con-ReCaLL	71.5%	49.9%	47.2%	49.8%
DC-PDD	39.2%	31.0%	24.7%	16.5%
Min-K%	46.5%	36.8%	30.4%	22.0%
ReCaLL	53.2%	39.0%	33.0%	24.1%
Ref	55.0%	42.7%	34.7%	26.8%
Min-K%++	16.6%	25.5%	23.7%	15.2%

Table 4: Decay rate of n-MVD after continual pretraining. Lower values indicate more robust retention of the membership signal through CPT.

while all methods exhibit performance drops after CPT, decay rates vary substantially: MIN-K%++ proves the most robust, whereas CON-RECALL suffers the largest erosion, consistent with its structurally fragile score function discussed in Sec. 4.2. Second, decay rates decrease consistently with model scale across nearly all methods, directly corroborating the prediction in Sec. 5.1: in higher-dimensional parameter spaces, CPT gradients are more orthogonal to $\Delta\theta$, leaving the membership signal less disturbed. This implies that privacy risks do not simply diminish as pre-training proceeds — larger models retain membership signals more durably. Third, despite the overall performance drop, the relative ordering of methods is largely preserved, confirming that Δ-MIA's assessments of relative method quality are robust to training recency.

6 Related Work

Fine-tuning in privacy leakage. Several works use controlled fine-tuning to study privacy leakage in language models. Mireshghallah et al. (2022b) show that different adaptation strategies exhibit substantially different memorization risks, while Lukas et al. (2023) connect record-level membership inference to PII reconstruction in fine-tuned models. Fu et al. (2024) propose self-prompt calibration to reduce sample-difficulty confounding in MIA against fine-tuned LLMs. These works treat the fine-tuned model as the auditing target. In contrast, Δ-MIA uses controlled exposure as an experimental intervention to evaluate pretraining-data MIA methods under a fixed-record contrast.

Evaluation and distributional confounding. MIA performance is highly sensitive to how member and non-member sets are constructed (Meeus et al., 2025; Das et al., 2025). Ye et al. (2022) formalize membership inference as hypothesis testing and distinguish evaluation games by which sources of uncertainty are fixed or randomized. Many LLM benchmarks follow a fixed-model observational design, where members and non-members are drawn from different sources or time periods (Shi et al., 2024; Meeus et al., 2024; Duan et al., 2024). Such designs can conflate membership evidence with distributional differences and inflate apparent attack performance. Δ-MIA instead adopts a fixed-record self-contrast design: target records are held constant, while two matched model trajectories, $\mathcal{M}_{\text{cpt}}$ and $\mathcal{M}_{\text{exp+cpt}}$, differ by the controlled target-exposure stage. This design controls member–non-member distributional confounding by construction and enables paired sample-level diagnosis.

Forgetting and membership-signal retention. Membership inference in pretraining is shaped by forgetting dynamics. Recent work shows that information acquired during pretraining can be forgotten as training proceeds (Liao et al., 2025), while Hayes et al. (2025) demonstrates that membership signals remain detectable in large models under strong attack conditions. These findings suggest MIA evaluation should account for whether signals survive subsequent training. Δ-MIA incorporates this perspective by measuring membership-signal decay across model scales after CPT, and offers a geometric account based on gradient–signal orthogonality in high-dimensional parameter spaces.

7 Conclusion

We introduced Δ-MIA, an interventional self-contrast framework that reframes MIA evaluation as a controlled exposure intervention over fixed target records. By comparing matched exposed and unexposed model trajectories, Δ-MIA controls member–non-member distribution shift by construction and enables paired sample-level diagnosis beyond aggregate attack metrics. Using this framework, we re-evaluated nine representative MIA methods and showed that controlled contrasts reveal both robust membership signals and method-specific instabilities that are hidden in observational evaluations. We further found that, under a fixed CPT budget, larger models retain membership signals more durably, consistent with gradient–signal orthogonality in high-dimensional parameter spaces. Overall, Δ-MIA establishes a new interventional paradigm for interpretable MIA evaluation in open-weight LLMs.

Limitations

Δ-MIA requires access to model weights for fine-tuning, making it directly applicable only to open-weight models. Evaluation of closed-source proprietary models remains an open challenge shared by the broader MIA community; for models with known release dates, timestamp-based data collection provides a partial path forward without requiring access for the non-member set construction, though the controlled exposure procedure itself still requires open weights.

The framework also incurs higher computational cost than purely observational benchmarks, as it requires training multiple model checkpoints per evaluation. We would mitigate this in practice by releasing Pythia checkpoints ($\mathcal{M}_{\text{cpt}}$, $\mathcal{M}_{\text{exp+cpt}}$) publicly, allowing future work to reuse them without retraining.

References

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 2397–2430. PMLR.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. 2021. Membership inference attacks from first principles. *CoRR*, abs/2112.03570.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2020. Extracting training data from large language models. *CoRR*, abs/2012.07805.

Ilias Chalkidis, Nicolas Garneau, Catalina Goanta, Daniel Martin Katz, and Anders Søgaard. 2023. Lexfiles and legallama: Facilitating english multinational legal language model development. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 15513–15535. Association for Computational Linguistics.

Debeshee Das, Jie Zhang, and Florian Trantèr. 2025. Blind baselines beat membership inference attacks for foundation models. In *2025 IEEE Security and Privacy, SP 2025 - Workshops, San Francisco, CA, USA, May 15, 2025*, pages 118–125. IEEE.

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? *CoRR*, abs/2402.07841.

Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. 2020. Orthogonal gradient descent for continual learning. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 3762–3773. PMLR.

Veve Fry. 2025. An overview of the risk-based model of ai governance. *Preprint*, arXiv:2507.15299.

Wenjie Fu, Huandong Wang, Chen Gao, Guanghua Liu, Yong Li, and Tao Jiang. 2024. Membership inference attacks against fine-tuned large language models via self-prompt calibration. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2021. The pile: An 800gb dataset of diverse text for language modeling. *CoRR*, abs/2101.00027.

Jiahui Geng, Qing Li, Herbert Woisetschlaeger, Zongxiong Chen, Fengyu Cai, Yuxia Wang, Preslav Nakov, Hans-Arno Jacobsen, and Fakhri Karray. 2025. A comprehensive survey of machine unlearning techniques for large language models. *Preprint*, arXiv:2503.01854.

Jack Hardinges, Elena Simperl, and Nigel Shadbolt. 2024. We Must Fix the Lack of Transparency Around the Data Used to Train Foundation Models. *Harvard Data Science Review*, (Special Issue 5). Https://hdsr.mitpress.mit.edu/pub/xau9dza3.

Jamie Hayes, Ilia Shumailov, Christopher A Choquette-Choo, Matthew Jagielski, Georgios Kaissis, Milad Nasr, Meenatchi Sundaram Muthu Selva Annamalai, Niloofar Mireshghallah, Igor Shilov, Matthieu Meeus, and 1 others. 2025. Exploring the limits of strong membership inference attacks on large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Chonghua Liao, Ruobing Xie, Xingwu Sun, Haowen Sun, and Zhanhui Kang. 2025. Exploring forgetting in large language model pre-training. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2112–2127, Vienna, Austria. Association for Computational Linguistics.

Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Beguelin. 2023. Analyzing Leakage of Personally Identifiable Information in Language Models . In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 346–363, Los Alamitos, CA, USA. IEEE Computer Society.

Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schoelkopf, 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, Toronto, Canada. Association for Computational Linguistics.

Andrzej Maćkiewicz and Waldemar Ratajczak. 1993. Principal components analysis (pca). *Computers & Geosciences*, 19(3):303–342.

Matthieu Meeus, Shubham Jain, Marek Rei, and Yves-Alexandre de Montjoye. 2024. Did the neurons read your book? document-level membership inference for large language models. In *33rd USENIX Security Symposium, USENIX Security 2024, Philadelphia, PA, USA, August 14-16, 2024*. USENIX Association.

Matthieu Meeus, Igor Shilov, Shubham Jain, Manuel Faysse, Marek Rei, and Yves-Alexandre de Montjoye. 2025. Sok: Membership inference attacks on llms are rushing nowhere (and how to fix it). In *IEEE Conference on Secure and Trustworthy Machine Learning, SaTML 2025, Copenhagen, Denmark, April 9-11, 2025*, pages 385–401. IEEE.

Fatemehsadat Mireshghallah, Kartik Goyal, Archit Uniyal, Taylor Berg-Kirkpatrick, and Reza Shokri. 2022a. Quantifying privacy risks of masked language models using membership inference attacks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8332–8347, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Fatemehsadat Mireshghallah, Archit Uniyal, Tianhao Wang, David Evans, and Taylor Berg-Kirkpatrick. 2022b. An empirical analysis of memorization in fine-tuned autoregressive language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1816–1826, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pages 3–18. IEEE Computer Society.

Liwei Song and Prateek Mittal. 2020. Systematic evaluation of privacy risks of machine learning models. In *USENIX Security Symposium*.

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

Roy Xie, Junlin Wang, Ruomin Huang, Minxing Zhang, Rong Ge, Jian Pei, Neil Zhenqiang Gong, and Bhuwan Dhingra. 2024. ReCaLL: Membership inference via relative conditional log-likelihoods. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8671–8689, Miami, Florida, USA. Association for Computational Linguistics.

Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, Vincent Bindschaedler, and Reza Shokri. 2022. Enhanced membership inference attacks against machine learning models. *Preprint*, arXiv:2111.09679.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. Privacy risk in machine learning: Analyzing the connection to overfitting. *Preprint*, arXiv:1709.01604.

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

Weichao Zhang, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. 2024. Pretraining data detection for large language models: A divergence-based calibration method. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5263–5274, Miami, Florida, USA. Association for Computational Linguistics.

Appendix

A Controlled Behavior Verification for Δ-MIA

A key property of Δ-MIA's experimental design is that $\mathcal{M}_{\text{cpt}}$ and $\mathcal{M}_{\text{exp+cpt}}$ differ *only* in whether they were exposed to the target data during training. A subtle concern is that although $\mathcal{M}_{\text{cpt}}$ and $\mathcal{M}_{\text{exp+cpt}}$ are trained with the same CPT data and budget, their CPT trajectories start from different initializations: $\mathcal{M}_{\varnothing}$ and $\mathcal{M}_{\text{exp}}$. Thus, the target exposure could in principle induce path-dependent changes that propagate to the final model beyond the target records themselves. To empirically confirm this controllability, we verify that the two models behave near-identically on *any* data outside the target set. If this holds, score differences observed on the target instances can be unambiguously attributed to data exposure, rather than to incidental behavioral drift induced by the CPT process itself.

We conduct two complementary experiments using Pythia-2.8B, evaluating $\mathcal{M}_{\text{cpt}}$ and $\mathcal{M}_{\text{exp+cpt}}$ on held-out data that neither model has been exposed to during training.

A.1 Validation with a Distribution-Mismatched Dataset (C4)

We evaluate both models on 2,000 text samples from C4, which lie outside the distribution of the target data. For each sample, we record inference loss from both models and compute the token-level KL divergence between their output distributions, using $\mathcal{M}_{\text{cpt}}$ as reference.

As shown in Fig. 6, inference losses exhibit near-perfect linear correlation between the two models, and KL divergences are concentrated near zero. This confirms that data exposure during Phase 1 does not meaningfully alter the model's behavior on out-of-distribution data.

A.2 Verification on a Distribution-Matched Dataset (Pile Validation Split)

A more stringent concern is whether the models might diverge specifically on data sharing the same distribution as the target instances, which could artificially inflate MVD measurements. We address this using 1,500 samples from the Pile validation split (500 each from Pile-CC, PubMed Abstracts, and Wikipedia), which are distributionally matched to the target data but unseen by both models.

As shown in Fig. 7, KL divergences remain near zero across all three domains. Even on distribution-matched held-out data, the two models behave near-identically, confirming that any score difference observed on the target instances reflects genuine membership effects rather than CPT-induced distributional drift.

B Experimental Details

B.1 Training Configuration

Target Data Selection. Our target-domain selection generally follows MIMIR (Duan et al., 2024) with two adjustments. First, we exclude the GitHub subset from Pile due to substantial non-English and malformed content that destabilizes MIA evaluation. Second, we select domains whose sample lengths remain within passage-level, ensuring consistent applicability across methods sensitive to context length (e.g., ReCaLL).

Training Procedure. All training uses full-parameter fine-tuning via LLaMA-Factory with DeepSpeed ZeRO-2, on 8 NVIDIA H100 GPUs. The procedure constructs four model checkpoints from each base Pythia model $\mathcal{M}_{\varnothing}$:

- $\mathcal{M}_{\text{exp}}$: fine-tuned on the 1,500 target instances for one epoch (Phase 1). All samples are processed in a small number of optimizer steps with effective batch size 128, learning rate 1×10^{-5}, constant scheduler, weight decay 0.01, gradient clipping 1.0, FP16.
- $\mathcal{M}_{\text{exp+cpt}}$: initialized from $\mathcal{M}_{\text{exp}}$, then trained on 1M Pile training records for one epoch (Phase 2). Effective batch size 1,024, learning rate 1×10^{-5}, cosine scheduler with 5% warmup, weight decay 0.01, gradient clipping 1.0, FP16, cutoff length 2,048.
- $\mathcal{M}_{\text{cpt}}$: initialized from $\mathcal{M}_{\varnothing}$ directly (Phase 1 skipped), then trained with the identical Phase 2 configuration as $\mathcal{M}_{\text{exp+cpt}}$.
- $\mathcal{M}_{\varnothing}$: the original Pythia checkpoint, unchanged.

Per-model batch size configurations are summarized in Appendix B.1.

Auxiliary Data for Traditional Setups. To support MIA methods requiring additional member/non-member sets (e.g., Con-ReCaLL), we provide a small auxiliary dataset: 60 instances from the Pile training split (seen by $\mathcal{M}_{\varnothing}$ and thus by all models) as members, and 60 instances from the Pile validation split (unseen by all models) as non-members.

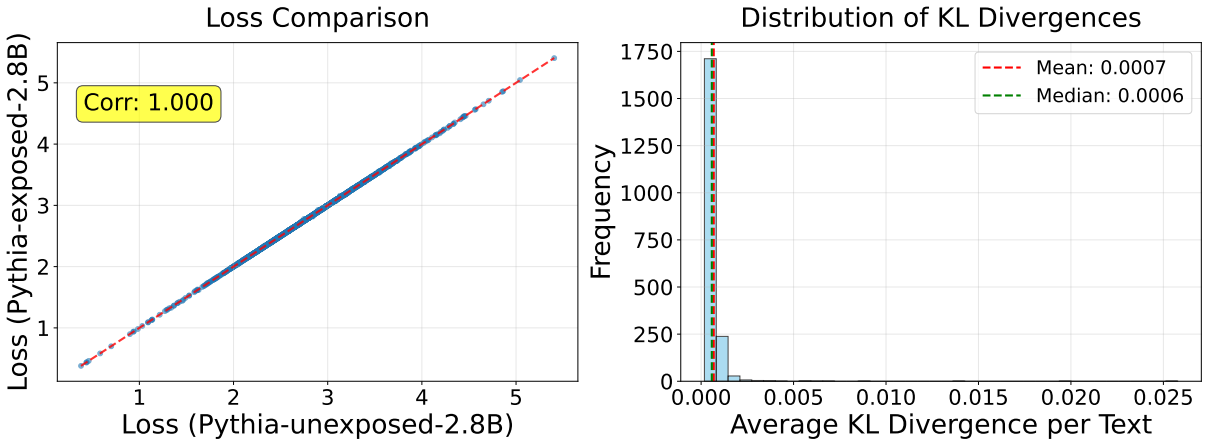


Figure 6: Left: Scatter plot of inference losses from $\mathcal{M}_{\text{cpt}}$ and $\mathcal{M}_{\text{exp+cpt}}$ (Pythia-2.8B) on 2,000 held-out C4 samples. Right: Frequency histogram of average token-level KL divergence per sample between the two models' output distributions.

Model	**Phase 1**		**Phase 2**		
	Per-device BS	Grad. Accum	Per-device BS	Grad. Accum	Eff. BS
Pythia-410M	16	1	64	2	1024
Pythia-1B	16	1	64	2	1024
Pythia-2.8B	8	2	32	4	1024
Pythia-6.9B	4	4	16	8	1024

Table 5: Per-model training configurations. Effective batch size (Eff. BS) = per-device batch size $\times$ gradient accumulation steps $\times$ 8 GPUs.

B.2 Details for Hyperparameters in MIA Evaluation

To ensure fair evaluation, we adopt hyperparameter settings from each method's original publication without additional tuning.

1. **Ref** (Carlini et al., 2020): Pythia-70M as reference model $\mathcal{M}_{\text{ref}}$.
2. **Neighborhood** (Mattern et al., 2023): 10 neighbor samples per target sample.
3. **Min-K% & Min-K%++** (Shi et al., 2024; Zhang et al., 2025): Bottom 20% of tokens by predicted probability.
4. **DC-PDD** (Zhang et al., 2024): Upper bound $a = 0.01$ for cross-entropy between token probability and frequency distributions.
5. **ReCaLL** (Xie et al., 2024): num_shots $= 1$.
6. **Con-ReCaLL** (Wang et al., 2025): num_shots $= 1$, $\gamma = 0.5$.

C Details of Domain-Specific MIA Performance

Appendix B.1, Appendix B.2, and Appendix B.2 provide domain-specific results across Pile-CC, PubMed Abstracts, and Wikipedia for Pythia models at 410M, 1B, and 6.9B scales, supplementing the 2.8B results in the main text.

D Formal Analysis of Con-ReCaLL Score Inversion

Con-ReCaLL defines its membership score as:

$$f(x;\mathcal{M}) = \frac{\text{LL}(x|P_{\text{non-mem}};\mathcal{M}) - \gamma \cdot \text{LL}(x|P_{\text{mem}};\mathcal{M})}{\text{LL}(x;\mathcal{M})} \quad (7)$$

where $P_{\text{non-mem}}$ and P_{mem} are non-member and member prefixes drawn from the auxiliary sets described in Appendix B.1, and $\text{LL}(x;\mathcal{M}) < 0$ holds by definition.

Expected behavior. The method's design assumption is that a non-member prefix, being distributionally distant from the target text, perturbs the model's conditional predictions more strongly than a member prefix. Under this assumption:

$$\text{LL}(x|P_{\text{non-mem}};\mathcal{M}) < \gamma \cdot \text{LL}(x|P_{\text{mem}};\mathcal{M}) \quad (8)$$

making the numerator negative. Combined with a negative denominator, this yields a positive score.

Method	Pile-CC		PubMed		Wikipedia	
	AUC	T@5%F	AUC	T@5%F	AUC	T@5%F
Loss	0.533	0.061	0.535	0.067	0.536	0.069
Zlib	0.518	0.067	0.516	0.067	0.540	0.067
Neighborhood	0.537	0.067	0.534	0.073	0.533	0.057
Con-ReCaLL	0.481	0.044	0.496	0.051	0.510	0.048
DC-PDD	0.536	0.067	0.536	0.065	0.540	0.053
Min-K%	0.539	0.071	0.542	0.079	0.547	0.079
ReCaLL	0.574	0.061	0.574	0.069	0.549	0.069
Ref	**0.588**	**0.080**	**0.591**	**0.116**	**0.597**	0.069
Min-K%++	0.574	0.065	0.577	0.077	0.570	**0.084**

Table 6: Dataset-level results on the Δ-MIA Benchmark for the Pythia-410M model across three domains (Pile-CC, PubMed Abstracts, Wikipedia (en)). We report AUC and TPR at 5% FPR (T@5%F); higher is better. The best score in each column is **bold**; the second best is underlined.

Method	Pile-CC		PubMed		Wikipedia	
	AUC	T@5%F	AUC	T@5%F	AUC	T@5%F
Loss	0.550	0.063	0.555	0.077	0.553	0.077
Zlib	0.529	0.061	0.526	0.073	0.561	0.077
Neighborhood	0.555	0.059	0.553	0.086	0.550	0.062
Con-ReCaLL	0.458	0.036	0.496	0.059	0.514	0.060
DC-PDD	0.555	0.075	0.554	0.065	0.559	0.072
Min-K%	0.560	0.080	0.565	0.086	0.565	0.079
ReCaLL	0.603	0.067	0.617	0.084	0.565	0.084
Ref	**0.612**	**0.090**	0.617	**0.114**	**0.621**	0.077
Min-K%++	**0.612**	0.084	**0.623**	0.102	0.608	**0.091**

Table 7: Dataset-level results on the Δ-MIA Benchmark for the Pythia-1B model across three domains (Pile-CC, PubMed Abstracts, Wikipedia (en)). We report AUC and TPR at 5% FPR (T@5%F); higher is better. The best score in each column is **bold**; the second best is underlined.

Method	Pile-CC		PubMed		Wikipedia	
	AUC	T@5%F	AUC	T@5%F	AUC	T@5%F
Loss	0.608	0.105	0.616	0.141	0.611	0.096
Zlib	0.564	0.086	0.559	0.120	0.619	0.127
Neighborhood	0.616	0.103	0.622	0.134	0.607	0.057
Con-ReCaLL	0.416	0.038	0.474	0.053	0.506	0.067
DC-PDD	0.624	0.101	0.624	0.141	0.637	0.079
Min-K%	0.632	0.128	0.641	0.177	0.634	0.105
ReCaLL	0.669	0.126	0.687	0.145	0.605	0.081
Ref	0.680	0.136	0.687	0.204	0.679	0.108
Min-K%++	**0.718**	**0.178**	**0.730**	**0.287**	**0.720**	**0.144**

Table 8: Dataset-level results on the Δ-MIA Benchmark for the Pythia-6.9B model across three domains (Pile-CC, PubMed Abstracts, Wikipedia (en)). We report AUC and TPR at 5% FPR (T@5%F); higher is better. The best score in each column is **bold**; the second best is underlined.

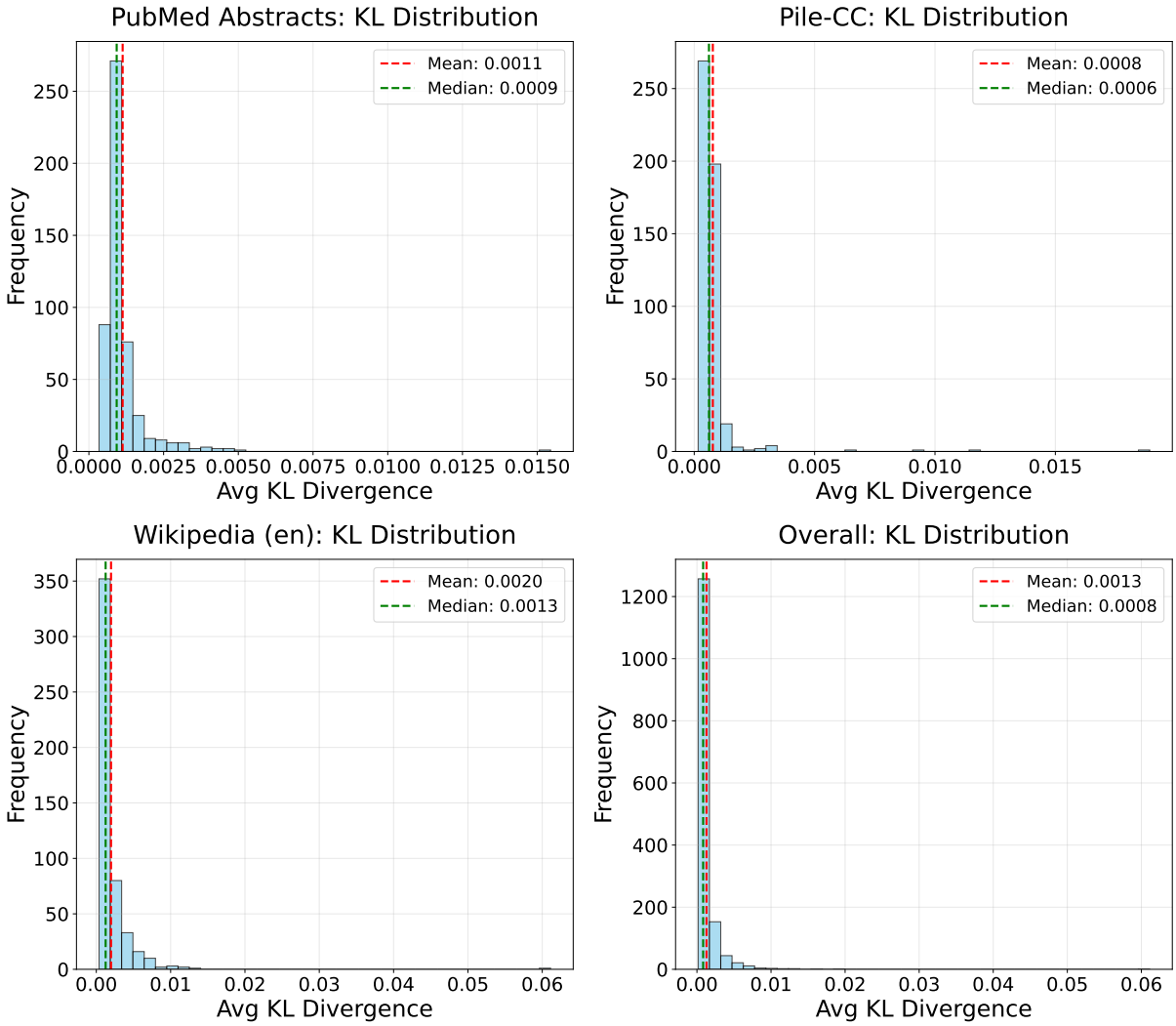


Figure 7: Distributions of average KL divergence between $\mathcal{M}_{\text{cpt}}$ and $\mathcal{M}_{\text{exp+cpt}}$ (Pythia-2.8B) on the Pile validation split (distribution-matched, held-out). From top-left to bottom-right: Pile-CC, PubMed Abstracts, Wikipedia, and the combined dataset. All distributions concentrate near zero, confirming that data exposure introduces no behavioral drift on non-target data sharing the same distribution.

A more negative numerator (stronger perturbation by the non-member prefix) and a less negative denominator (higher unconditional likelihood, consistent with membership) both increase the score — a coherent signal direction.

Score inversion under violated ordering. The ordering above is not guaranteed. The relative magnitude of the two conditional log-likelihoods depends on factors external to membership status, including the intrinsic linguistic complexity of each prefix and its distributional proximity to the target text x. When a member prefix happens to be distributionally more distant from x than the non-member prefix, the ordering reverses and the numerator becomes positive. In this case:

$$f(x; \mathcal{M}) = \frac{\text{(positive)}}{\text{(negative)}} < 0 \tag{9}$$

The score becomes negative, and its behavior with respect to the denominator inverts: a more negative denominator — indicating lower unconditional likelihood, inconsistent with membership — now *increases* the score. The signal direction is thus opposite to the intended semantics, rendering the score unreliable for that sample.

Aggregate consequences. When score inversion occurs randomly across samples, the method's ADR degrades because a fraction of samples are assigned scores that contradict their true membership status. Importantly, this is an intrinsic property of

the score function and is not specific to the Δ-MIA evaluation setting. In observational benchmarks, the issue may be partially masked: when member and non-member data are drawn from training and held-out corpora respectively, their distributional differences correlate with membership status and can provide an auxiliary signal that partially compensates for per-sample inversions, inflating aggregate metrics without resolving the underlying structural issue.

E Interpreting Scatter Plots for ROC Metrics

The scatter plot geometry — normalized non-member scores (x-axis) versus member scores (y-axis) — directly encodes the TPR–FPR trade-off. For a threshold τ, define four quadrants:

- TL(τ): $\{s_{\text{mem}} \geq \tau \wedge s_{\text{non}} < \tau\}$
- TR(τ): $\{s_{\text{mem}} \geq \tau \wedge s_{\text{non}} \geq \tau\}$
- BL(τ): $\{s_{\text{mem}} < \tau \wedge s_{\text{non}} < \tau\}$
- BR(τ): $\{s_{\text{mem}} < \tau \wedge s_{\text{non}} \geq \tau\}$

For TPR@5%FPR, τ is the 95th percentile of non-member scores, fixing FPR at 5%. TPR and the TPR–FPR gap are then:

$$\text{TPR} = \frac{|\text{TL}(\tau)| + |\text{TR}(\tau)|}{N} \tag{10}$$

$$\text{TPR} - \text{FPR} = \frac{|\text{TL}(\tau)| - |\text{BR}(\tau)|}{N} \tag{11}$$

Superior performance at low FPR requires maximizing TL(τ) density while minimizing BR(τ) density.

ReCaLL's distribution is heavily left-skewed, placing most non-member scores near zero. This yields a very low 95th percentile threshold ($\tau = 0.300$), which also falls below a substantial fraction of member scores, resulting in a sparse TL(τ) and low TPR. Ref's more dispersed distribution produces $\tau = 0.673$, above which a larger fraction of member scores lie, yielding a dense TL(τ) and correspondingly high TPR (Fig. 8).

F Supplementary Observational Evaluation and Paradigm Equivalence Verification

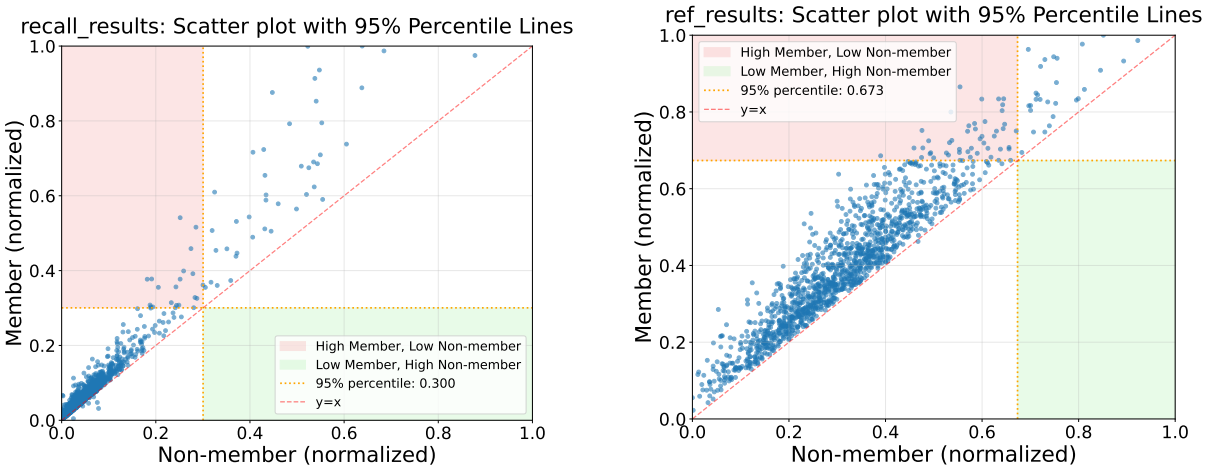
To verify that Δ-MIA isolates the same membership signal targeted by observational MIA, we conduct a standard observational evaluation using $\mathcal{M}_{\text{exp+cpt}}$ as the target model. The member set consists of the 1,500 target instances used in Δ-MIA; the non-member set comprises 1,500 matched samples from the Pile validation split (Pile-CC, PubMed, Wikipedia). All methods use the same hyperparameters as in Δ-MIA (Appendix B.2).

Paradigm Equivalence. Performance trends are highly consistent across paradigms (Appendix F). Ref and Min-K%++ remain the top performers with similar scaling behavior; Loss, Zlib, Neighborhood, and Min-K% also show aligned relative rankings. This confirms that Δ-MIA captures the same core membership signal as observational evaluation, while offering stronger internal validity through controlled distribution.

Method-Specific Discrepancies. Two methods diverge notably across paradigms.

Con-ReCaLL performs substantially better under the observational paradigm. As discussed in Sec. 4.2, its score function lacks a guaranteed sign for the numerator, making it structurally dependent on distributional differences between member and non-member prefixes. In the observational setting, residual distributional gaps — even in carefully matched sets — provide an incidental corrective signal. Δ-MIA removes this entirely, exposing the method's structural limitation.

DC-PDD similarly benefits from residual distributional differences in the observational setting. Its token frequency estimation is sensitive to subtle member/non-member distributional gaps that persist even in matched observational benchmarks. Δ-MIA eliminates these gaps, revealing the method's limited ability to capture genuine membership signals.



(a) ReCaLL (95th percentile = 0.300)

(b) Ref (95th percentile = 0.673)

Figure 8: Visualizing the impact of score distribution on TPR@5%FPR. (a) The scatter plot for ReCaLL exhibits a distribution skewed toward low non-member scores. This results in a low 95th percentile threshold ($\tau = 0.300$) and a sparsely populated top-left (TL) region, leading to a low TPR. (b) In contrast, the Ref method produces a more dispersed distribution, yielding a higher threshold ($\tau = 0.673$) and a dense TL region, which corresponds to a high TPR. The diagonal dashed line indicates a random-guess baseline. The red shaded area represents the TL(τ) quadrant, while the green area represents the BR(τ) quadrant.

Method	Pythia-410M		Pythia-1B		Pythia-2.8B		Pythia-6.9B	
	AUC	T@5%F	AUC	T@5%F	AUC	T@5%F	AUC	T@5%F
Loss	0.557	0.077	0.573	0.080	0.594	0.081	0.620	0.087
Zlib	0.544	**0.084**	0.557	0.098	0.575	0.105	0.594	0.129
Neighborhood	0.552	0.076	0.573	0.084	0.599	0.104	0.631	0.131
Con-ReCaLL	0.563	0.081	0.596	0.079	0.626	0.084	0.644	0.093
DC-PDD	0.564	0.072	0.587	0.092	0.616	0.114	0.646	0.115
Min-K%	0.567	0.079	0.588	0.077	0.617	0.089	0.647	0.105
ReCaLL	0.564	0.076	0.602	0.076	0.629	0.082	0.653	0.090
Ref	0.572	0.076	0.602	0.095	0.629	**0.131**	0.665	0.162
Min-K%++	**0.595**	0.082	**0.639**	**0.101**	**0.688**	0.118	**0.739**	**0.184**

Table 9: Dataset-level results of the supplementary observational evaluation on post-exposure models. We report AUC and TPR at 5% FPR (T@5%F) for MIA methods evaluated on the Pythia model family (410M, 1B, 2.8B, and 6.9B). Higher is better. The best score in each column is **bold**; the second best is underlined.