

Are Fine-Tuned LLM Membership Signals Real? Separating Memorization from Likelihood Shortcuts

Zilong Wang Yingjie Xu Yutong Wu Ruiyang Li

Qinglin Ji Xueming Tang Songfeng Lu

School of Cyber Science and Engineering

Huazhong University of Science and Technology

Wuhan, China

{wzilong, xyingjie, wyt, ruiyangli}@hust.edu.cn

{qinglinji, xmtang, lusongfeng}@hust.edu.cn

Abstract

Membership inference attacks (MIAs) are increasingly used to audit whether protected examples were included in language model fine-tuning. Yet for fine-tuned LLMs, high attack AUC alone is not reliable evidence of data use: a score may separate members from non-members due to exposure, but also due to length, base model difficulty, typicality, or domain alignment. We propose DeMIA, a diagnostic membership inference framework that learns separate channels for membership scoring and difficulty modeling using likelihood features from the base and target models. DeMIA combines membership prediction with matched-pair ranking, base-difficulty anchoring, and adversarial length control, encouraging the final score to preserve membership-relevant evidence while reducing direct shortcut reliance. Across GPT-Neo-125M, Qwen2.5-1.5B, and Llama-3.1-8B on Wiki10k, AG-News10k, and XSum10k, DeMIA outperforms scalar likelihood, reference-calibrated, neighborhood, calibrated Min-K, and DF-style baselines while maintaining strong low-FPR recall. The signal remains strong on matched hard subsets, while blind controls using length, base loss, or other features from the base model remain near chance. These results show that fine-tuned LLM membership signals can persist under the tested nuisance controls and should be evaluated with reliability diagnostics rather than raw AUC alone.

1 Introduction

Membership inference attacks (MIAs) are increasingly used to assess whether protected or high-value data were used in LLM post-training. Beyond measuring privacy leakage, they are now discussed as tools for data-use auditing, provenance checking, and copyright-related verification (Zhao et al., 2024; Duarte et al., 2024; Zhang et al., 2025; Li et al., 2026; Wei et al., 2024). This shift changes the

evidentiary role of an attack. In a standard benchmark, a high AUC only shows that members and non-members are separable under the chosen split; it does not show that the separation is caused by exposure to the target record (Carlini et al., 2022; Zhang et al., 2025). For post-training auditing, this distinction matters because a score that separates members from non-members by exploiting length, difficulty, topic, or domain imbalance is not reliable evidence that the target record was used.

Fine-tuned LLMs make this evidentiary problem especially difficult. Fine-tuning induces a global likelihood shift toward the adaptation distribution, so memorization at the instance level is only one possible source of the observed score gap. A candidate example may appear member-like because it was exposed during fine-tuning, but also because it is shorter, easier, more typical, topically aligned, or closer to the fine-tuning domain. Consistent with this ambiguity, recent studies report a mixed empirical picture: some MIAs that rely on loss or confidence are close to random in controlled record-level settings, while stronger signals are reported when evidence is aggregated over documents or datasets (Duan et al., 2024; Maini et al., 2024; Chen et al., 2025; Puerto et al., 2025; Meeus et al., 2024). We therefore ask whether membership signals in fine-tuned LLMs are real in this evidentiary sense: do they persist after controlling for nuisance variation in language-model likelihood?

Existing MIAs for language models largely address this problem through score construction and calibration. Reference-free attacks derive membership scores from target-model likelihoods, such as loss, perplexity, pretraining data detection scores, or relative conditional likelihood, while reference-based and neighborhood-based attacks compare target scores against reference models, reference distributions, or local textual perturbations (Mireshghallah et al., 2022; Shi et al., 2024; Mattern et al., 2023; Xie et al., 2024). Recent vari-

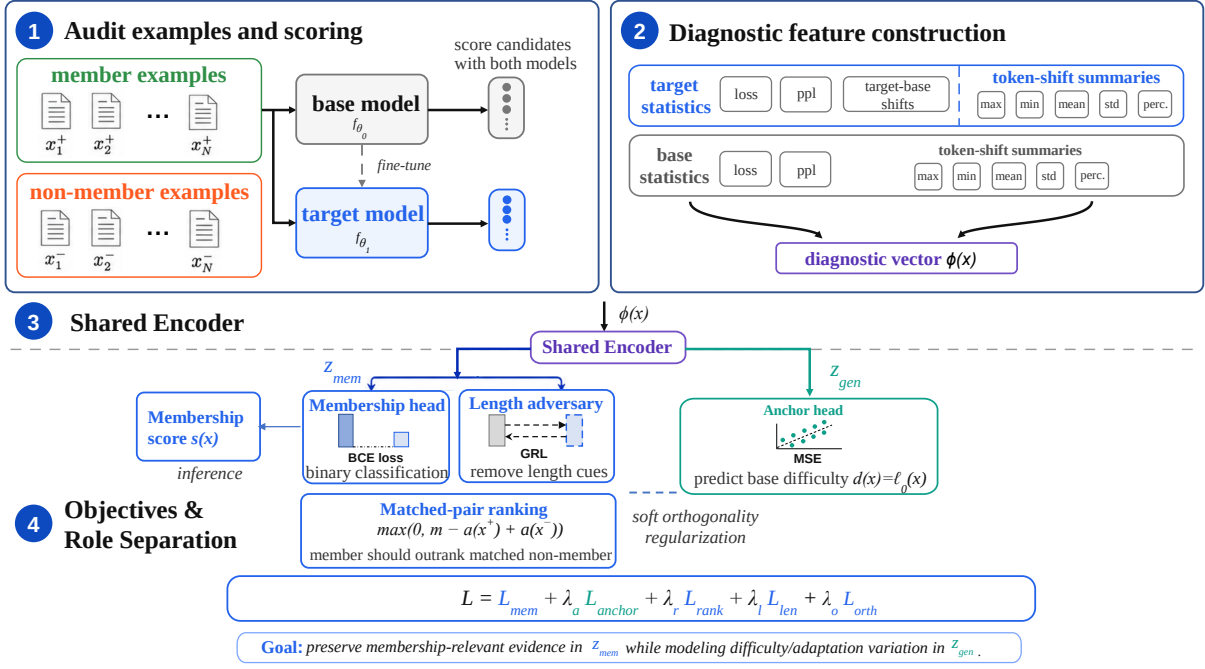


Figure 1: Overview of DeMIA. Candidate examples are scored by the public base model f_{θ_0} and the fine-tuned target model f_{θ_1} to construct diagnostic likelihood features $\phi(x)$. A shared encoder maps these features into a membership channel z_{mem} and a generalization channel z_{gen} . Their roles are shaped by membership prediction, matched ranking, length invariance, and base-difficulty anchoring. At inference time, only z_{mem} is used to produce the membership score.

ants further improve score construction through self-prompt calibration, distribution-free reference selection, contrastive or sampling-based likelihood estimation, in-context probing, and window-level aggregation (Fu et al., 2024; Huang et al., 2025; Wang et al., 2025; Kaneko et al., 2025; Lu et al., 2026; Chen et al., 2026). More recent work also reduces reliance on labeled non-members and evaluates membership inference under controlled distributional overlap and difficulty (Kim et al., 2026). However, calibration does not by itself resolve the attribution problem: if length, difficulty, typicality, topic, or domain fit remain predictive under the benchmark split, scalar scores can still encode these shortcuts. We therefore frame fine-tuned LLM membership inference as a problem of evidence attribution: the goal is not merely to obtain a higher AUC, but to learn membership evidence that remains predictive after controlling for direct nuisance factors.

To address this gap, we propose DeMIA, a role-separated evidence framework for reliable post-training auditing of fine-tuned LLMs. Instead of introducing another calibrated scalar score, DeMIA learns an evidence representation from diagnostic features extracted from the base and fine-tuned

models. It maps this representation into a membership channel z_{mem} , used for membership prediction, and a generalization channel z_{gen} , trained to model variation associated with difficulty and adaptation. The separation is induced by channel-specific objectives: a difficulty anchor for z_{gen} , matched-pair ranking under similar nuisance profiles, and adversarial length invariance for z_{mem} . Together, these objectives encourage the final score to rely on residual membership evidence rather than nuisance-induced separability.

We evaluate DeMIA under an evidentiary standard stricter than AUC alone. In addition to ROC-AUC, we consider low-FPR operating points, hard non-member subsets, matched comparisons, and blind shortcut controls. These diagnostics test whether the signal persists when easy separability is reduced and when simple explanations such as length or base model difficulty are isolated. Across multiple domains, model families, and fine-tuning checkpoints, DeMIA improves over scalar likelihood baselines and preserves membership separability under stricter evidence tests. Together, these tests provide stronger evidence than AUC alone that membership signals in fine-tuned LLMs persist after the tested nuisance factors are reduced.

2 Related Work

Membership inference as data-use auditing.

Membership inference attacks were originally introduced to detect whether a record was used in model training by exploiting behavioral differences between members and non-members (Shokri et al., 2017; Yeom et al., 2018; Carlini et al., 2022). With foundation models, MIAs are increasingly used not only to measure privacy leakage, but also for data-use auditing, provenance checking, and copyright-related verification (Maini et al., 2024; Zhao et al., 2024; Duarte et al., 2024; Zhang et al., 2025; Li et al., 2026; Wei et al., 2024; Panda et al., 2025). This broader use raises a stronger evidentiary requirement: a membership score should not merely separate benchmark splits, but should provide stronger evidence that the separation is related to exposure rather than benchmark-specific shortcuts. We study this issue in fine-tuned LLMs, where likelihood shifts may reflect either instance-level exposure or general adaptation.

Likelihood-based MIAs for language models.

Most MIAs for language models derive membership evidence from token likelihoods, including loss, perplexity, likelihood ratios, low-probability token scores, relative conditional likelihood, document aggregation, and neighborhood comparison (Miresghallah et al., 2022; Shi et al., 2024; Duan et al., 2024; Xie et al., 2024; Meeus et al., 2024; Mattern et al., 2023). These methods establish likelihood as the primary observable signal for LLM membership inference. However, likelihood also varies with difficulty, length, typicality, topic, and domain fit, so score separability alone does not determine whether a signal reflects exposure to a particular example.

Calibration and reference correction. A major line of work attempts to reduce likelihood confounding through calibration. Difficulty calibration shows that raw membership scores can be misleading when members and non-members differ in intrinsic difficulty (Watson et al., 2022). Language model attacks further refine or calibrate membership scores using reference models, local perturbations, self-prompted surrogate data, distribution-free reference signals, noisy neighbors, contrastive decoding, sampling-based pseudo-likelihood, in-context probing, or window-level aggregation (Miresghallah et al., 2022; Mattern et al., 2023; Fu et al., 2024; Huang et al., 2025;

Galli et al., 2024; Wang et al., 2025; Kaneko et al., 2025; Lu et al., 2026; Chen et al., 2026). These methods improve attack performance or relax access assumptions, but still primarily construct or calibrate scalar membership scores. In contrast, DeMIA uses separate channels for membership scoring and modeling difficulty-related variation.

Evaluation validity and shortcut confounding.

Recent studies question whether reported LLM MIA success reflects membership-related evidence or artifacts of evaluation design (Duan et al., 2024; Chen et al., 2025). A complementary SoK identifies distribution shifts introduced by post-hoc member/non-member construction as a major source of evaluation confounding (Meeus et al., 2025).

Large-scale evaluations show that many attacks based on loss or confidence are close to random in controlled record-level settings, while stronger results can appear with document aggregation, dataset-level inference, scaling regimes, or specific benchmark protocols (Maini et al., 2024; Puerto et al., 2025; Meeus et al., 2024). Blind baselines further show that member and non-member splits can be separable from surface statistics alone, without querying the target model (Das et al., 2025). These findings motivate our focus on diagnostics that test whether membership signals persist after direct nuisance factors are controlled.

3 Method: DeMIA

3.1 Problem Setup and Threat Model

We study record-level membership inference for fine-tuned large language models. Let f_{θ_0} denote a public base checkpoint, and let $\mathcal{D}_{\text{ft}}$ be a private fine-tuning dataset. The target model f_{θ_1} is obtained by adapting f_{θ_0} on $\mathcal{D}_{\text{ft}}$. Given a candidate text sequence x , membership inference asks whether x was used during fine-tuning:

$$y(x) = \mathbf{1}\{x \in \mathcal{D}_{\text{ft}}\}, \quad y(x) \in \{0, 1\}, \quad (1)$$

where $\mathbf{1}\{\cdot\}$ denotes the indicator function.

We consider a post-training auditing setting with likelihood access. The auditor can evaluate the token log-likelihood of a candidate sequence under both the public base model f_{θ_0} and the fine-tuned target model f_{θ_1} , but does not require gradients, optimizer states, training trajectories, or internal activations. This access model is weaker than full white-box access but stronger than generation-only black-box access.

The auditor is additionally assumed to have a candidate audit pool containing labeled member and non-member examples for calibration. In our empirical evaluation, these examples are partitioned into disjoint audit-training, validation, and test splits, and the supervised audit head is fitted only on the audit-training split. This setting models post-training audits in which candidate training examples and matched reference examples are available for calibration. Candidate-free or unknown-domain auditing is not considered here.

Given a candidate x , the auditor outputs a membership score

$$s(x) \in [0, 1], \quad (2)$$

where larger values indicate stronger membership evidence.

The main challenge is that fine-tuning changes model behavior at both the distribution and instance levels. A lower loss under f_{θ_1} may indicate exposure to x , but it may also reflect that x is short, easy, typical, or close to the fine-tuning domain. Thus, we do not treat raw likelihood separation as sufficient membership evidence. Instead, our objective is to learn a membership statistic that remains predictive after nuisance variation is controlled.

3.2 Diagnostic Features

For each candidate sequence $x = (x_1, \dots, x_T)$, we compute per-token negative log-likelihoods under the base and target models:

$$n_{m,t}(x) = -\log p_{\theta_m}(x_t | x_{<t}), \quad m \in \{0, 1\}, \quad (3)$$

where $m = 0$ denotes the base model and $m = 1$ denotes the fine-tuned target model. The corresponding average losses are

$$\ell_m(x) = \frac{1}{T} \sum_{t=1}^T n_{m,t}(x), \quad m \in \{0, 1\}. \quad (4)$$

We also compute perplexities $\text{ppl}_m(x) = \exp(\ell_m(x))$.

The basic diagnostic vector contains base and target likelihood statistics and their signed sequence-level shifts:

$$\begin{aligned} \phi_6(x) = [\ell_0(x), \text{ppl}_0(x), \ell_1(x), \text{ppl}_1(x), \\ \ell_1(x) - \ell_0(x), \text{ppl}_1(x) - \text{ppl}_0(x)]. \end{aligned} \quad (5)$$

This vector summarizes how fine-tuning changes the likelihood of x relative to the base model.

To capture variation across tokens, we define the signed per-token loss shift

$$\delta_t(x) = n_{1,t}(x) - n_{0,t}(x). \quad (6)$$

Negative values of $\delta_t(x)$ indicate tokens whose negative log-likelihood decreases after fine-tuning. The full feature vector $\phi_{15}(x)$ augments $\phi_6(x)$ with nine summary statistics of $\{\delta_t(x)\}_{t=1}^T$: the minimum, maximum, mean, standard deviation, and the 10th, 25th, 50th, 75th, and 90th percentiles. These statistics describe whether the likelihood change is broadly distributed across the sequence or concentrated on a small subset of tokens.

Raw sequence length is not included in $\phi(x)$. We treat length as a nuisance variable and use it only for matching, adversarial invariance, and diagnostic evaluation.

3.3 Role-Separated Evidence Representation

The diagnostic vector $\phi(x)$ can contain both membership-relevant evidence and nuisance information. A likelihood change after fine-tuning may indicate exposure to x , but it may also arise because x is easy, typical, or close to the adaptation distribution. DeMIA therefore treats membership inference as a problem of structuring evidence rather than calibrating a scalar score.

A shared encoder first maps $\phi(x)$ into a hidden representation,

$$h(x) = E_\psi(\phi(x)), \quad (7)$$

which is then projected into two latent channels:

$$\begin{aligned} z_{\text{mem}}(x) &= P_{\text{mem}}(h(x)), \\ z_{\text{gen}}(x) &= P_{\text{gen}}(h(x)). \end{aligned} \quad (8)$$

The membership channel z_{mem} is used to preserve membership-relevant evidence for prediction, while the generalization channel z_{gen} is trained to model variation associated with difficulty and adaptation.

The final membership score is computed only from z_{mem} :

$$s(x) = \sigma(C_{\text{mem}}(z_{\text{mem}}(x))). \quad (9)$$

The two channels are separated by functional role rather than by an explicit statistical-independence constraint. We assign objectives accordingly: membership prediction and matched ranking act on z_{mem} , difficulty anchoring acts on z_{gen} , and adversarial invariance discourages z_{mem} from encoding length shortcuts.

3.4 Training Objectives

DeMIA is trained with four primary signals that assign different roles to the two latent channels, together with an auxiliary soft orthogonality regularizer. Let

$$a(x) = C_{\text{mem}}(z_{\text{mem}}(x)), \quad s(x) = \sigma(a(x)), \quad (10)$$

where $a(x)$ is the membership logit and $s(x)$ is the membership score.

Membership prediction. The primary objective is binary membership prediction from z_{mem} :

$$\mathcal{L}_{\text{mem}} = -\mathbb{E}_{(x,y)} \left[y \log s(x) + (1-y) \log(1-s(x)) \right]. \quad (11)$$

This makes z_{mem} predictive of the audit label, but BCE alone can exploit any separable nuisance shortcut in $\phi(x)$.

Difficulty anchoring. We use base model difficulty as an anchor for the generalization channel. Let $d(x)$ denote an observable difficulty signal from the base model; in our default implementation, $d(x) = \ell_0(x)$, the base model loss. A difficulty head predicts

$$\hat{d}(x) = C_{\text{gen}}(z_{\text{gen}}(x)), \quad (12)$$

and is trained with

$$\mathcal{L}_{\text{anchor}} = \mathbb{E}_x \left[\left(\hat{d}(x) - d(x) \right)^2 \right]. \quad (13)$$

Because $d(x)$ is computed from the base checkpoint and does not depend on the target fine-tuning update, it captures difficulty present before fine-tuning rather than changes induced by the target update. This gives z_{gen} a concrete role: modeling difficulty-related variation that would otherwise be attractive to the membership classifier.

Adversarial length invariance. Length is a direct surface shortcut because it affects likelihood estimates and may differ across benchmark splits. Let $r(x)$ denote a discretized length bin. We attach a length predictor to z_{mem} through a gradient reversal layer:

$$\hat{r}(x) = Q_{\eta}(\text{GRL}(z_{\text{mem}}(x))). \quad (14)$$

The length loss is

$$\mathcal{L}_{\text{len}} = \mathbb{E}_x [\text{CE}(\hat{r}(x), r(x))]. \quad (15)$$

The length head learns to predict length, while the reversed gradient updates the encoder and membership projection to make length harder to recover from z_{mem} . This regularizer specifically reduces recoverable length information in z_{mem} ; other nuisance factors are addressed through matching and diagnostic evaluation.

Matched-pair ranking. To encourage membership separation under comparable nuisance profiles, we construct matched pairs (x_i^+, x_j^-) , where x_i^+ is a member and x_j^- is a non-member matched by sequence length and base model difficulty. We apply a margin ranking loss to membership logits:

$$\mathcal{L}_{\text{rank}} = \mathbb{E}_{(x^+, x^-)} \left[\max \{0, m - a(x^+) + a(x^-)\} \right]. \quad (16)$$

where $m > 0$ is a margin. This encourages a member to receive a higher membership logit than a non-member with a similar nuisance profile.

Pair construction. For each member example x_i^+ , we retrieve a matched non-member x_j^- using only pre-target nuisance statistics. In our implementation, matching uses sequence length and base model loss $\ell_0(x)$. We standardize both quantities on the audit-training split and select the nearest non-member using Euclidean distance with equal weighting. Target-model scores, DeMIA predictions, and membership logits are not used for pair construction. This training-time matching is distinct from the hard-subset evaluation in Section 4, but both use only pre-target nuisance statistics.

Overall objective. The final objective is

$$\mathcal{L} = \mathcal{L}_{\text{mem}} + \lambda_a \mathcal{L}_{\text{anchor}} + \lambda_r \mathcal{L}_{\text{rank}} + \lambda_l \mathcal{L}_{\text{len}} + \lambda_o \mathcal{L}_{\text{orth}}. \quad (17)$$

where $\mathcal{L}_{\text{orth}}$ is an auxiliary soft orthogonality regularizer between z_{mem} and z_{gen} . It encourages complementary structure between the two channels without imposing an independence constraint. The coefficients λ_a , λ_r , λ_l , and λ_o control difficulty anchoring, matched ranking, length invariance, and soft channel regularization, respectively.

3.5 Inference

At inference time, DeMIA computes $\phi(x)$, maps it to $z_{\text{mem}}(x)$, and outputs

$$s(x) = \sigma(C_{\text{mem}}(z_{\text{mem}}(x))). \quad (18)$$

The generalization channel, difficulty head, and length adversary are used only during training to shape the learned representation. At inference time, the deployed output remains a scalar membership score produced from z_{mem} .

4 Experiments

We evaluate DeMIA along three complementary dimensions. First, we examine its membership inference performance relative to existing score-based attacks across models, datasets, and fine-tuning stages. Second, we test whether the role-separated training objectives provide benefits beyond simply fitting a supervised classifier on the same diagnostic features. Third, we examine whether the resulting membership signal persists under increasingly stringent controls for nuisance variation, textual overlap, and run-specific effects. Detailed per-setting results, statistical analyses, access assumptions, and additional robustness diagnostics are reported in the appendix.

4.1 Evaluation Protocol

Benchmarks and target models. We use three 10k-example text benchmarks: Wiki10k (Merity et al., 2017), AGNews10k (Zhang et al., 2015), and XSum10k (Narayan et al., 2018), covering encyclopedic, news-style, and summarization-style text. We evaluate three fine-tuned LLM families: GPT-Neo-125M (Black et al., 2021), Qwen2.5-1.5B (Yang et al., 2024), and Llama-3.1-8B (Grattafiori et al., 2024).

Compared methods. We compare DeMIA with representative score-based MIAs, including target loss, reference/relative likelihood calibration (REF/RCL), Min-K% token loss, neighborhood perturbation, calibrated Min-K variants, DF-style candidate calibration, and WBC-MARGIN where aligned token-window traces are available. These methods cover target-only scoring, base/reference calibration, token-tail statistics, local perturbation sensitivity, candidate-pool calibration, and localized window scoring. We also compare against *Strong MLP*, a supervised classifier using the same diagnostic features and audit labels as DeMIA. Role-specific ablations are examined separately in Section 4.3. Appendix B.2 summarizes access assumptions and implementation coverage.

Metrics and diagnostics. We report ROC-AUC and TPR at 5% and 1% FPR, with emphasis on

Method	#	AUC	TPR@5	TPR@1	Hard
Loss	7	0.819	0.395	0.125	0.831
Ref/RCL	7	0.929	0.656	0.266	0.916
Min-K	7	0.582	0.067	0.017	0.606
Cal-MinK	7	0.683	0.150	0.032	0.669
DF-style	7	0.838	0.227	0.026	0.809
WBC-margin	6	0.946	0.621	0.177	0.945
DeMIA	7	0.960	0.853	0.624	0.942

Table 1: Aggregate performance against score-based MIA baselines on the representative reliability settings. # denotes the number of available settings; TPR@5/1 denotes TPR at 5%/1% FPR, and Hard denotes hard-subset AUC.

the low-FPR regime relevant to auditing. For key comparisons, we additionally report paired bootstrap confidence intervals. We evaluate validation-threshold transfer by selecting thresholds on validation non-members and applying them unchanged to held-out test splits. Reliability diagnostics include hard non-member and matched subsets, blind controls using pre-target or text-only features, text-similarity filtering, and an independent-rerun evaluation. Detailed construction procedures and additional robustness results are reported in the appendix.

4.2 RQ1: Overall Membership Inference Performance

Figure 2 shows the performance trends across model families, datasets, and fine-tuning checkpoints. Across all nine model-dataset combinations, DeMIA consistently achieves higher AUC than the plotted likelihood baselines, including target loss, Min-K%, reference-style likelihood calibration, and neighborhood perturbation. The advantage generally becomes more pronounced at later fine-tuning checkpoints, indicating that the membership signal emerging during fine-tuning is not fully captured by any single scalar likelihood score.

Table 1 summarizes performance on the representative reliability settings. DeMIA achieves the highest mean AUC and low-FPR recall among the score-based methods, with an AUC of 0.960, TPR@5 of 0.853, and TPR@1 of 0.624. In particular, TPR@1 increases from 0.266 for REF/RCL to 0.624 for DeMIA, showing a substantially larger advantage in the strict low-FPR regime than is apparent from AUC alone. WBC-MARGIN reaches a comparable hard-subset AUC but has lower low-FPR recall and is available for fewer representative

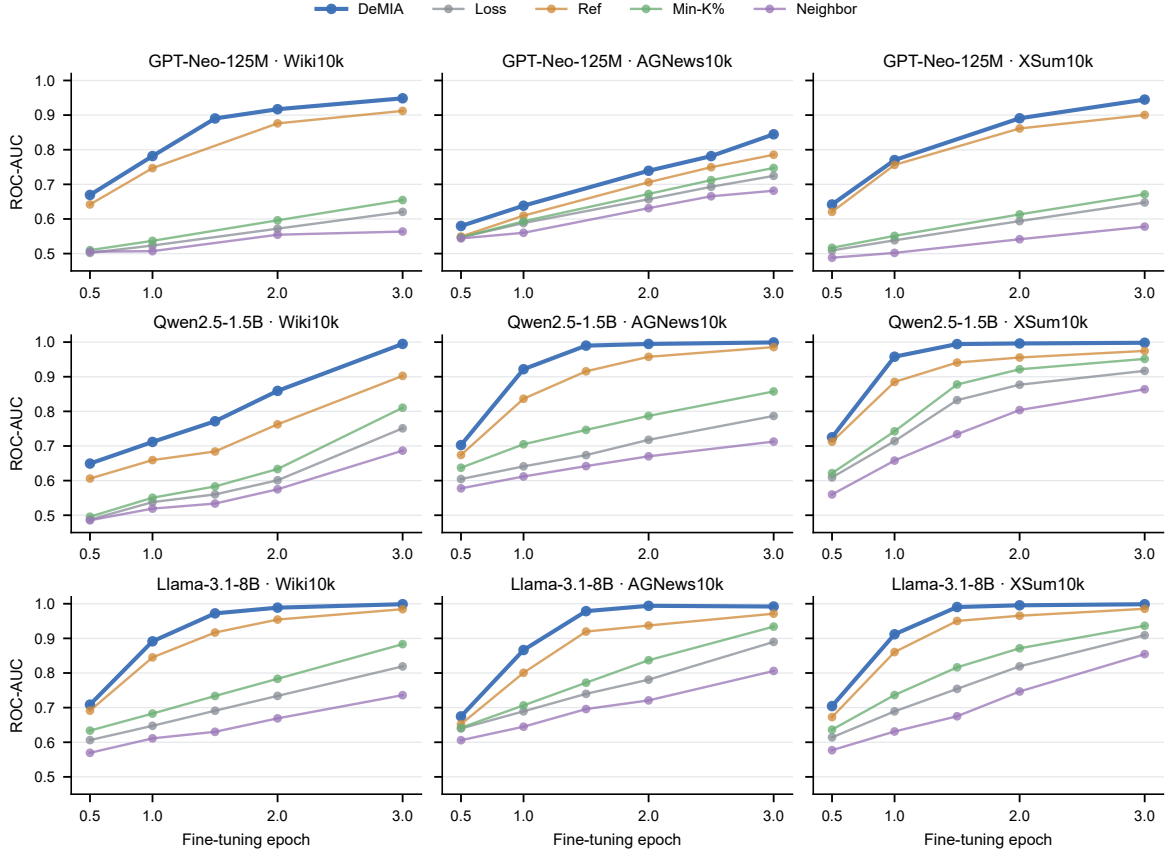


Figure 2: Overall membership inference AUC across fine-tuning epochs. Rows correspond to target model families and columns to datasets. DeMIA consistently achieves higher AUC than the likelihood-based baselines across the evaluated model–dataset settings, with the performance gap generally widening as fine-tuning progresses.

settings because it requires aligned token-window traces. Same-feature supervised comparisons are examined separately in Section 4.3.

4.3 RQ2: Does Role-Separated Training Add Beyond Supervision?

We compare DeMIA with supervised controls that use the same diagnostic features and audit labels. *Strong MLP* applies an unconstrained membership classifier to the full diagnostic vector, while the no-role-separation ablation removes the role-separated architecture. *WeakReg* and *OverReg* vary the strength of nuisance regularization. Table 2 reports averages over nine settings.

Across the nine settings, DeMIA and Strong MLP are close in average AUC (0.960 versus 0.956), with somewhat larger differences in the low-FPR metrics. Aggregate averages, however, hide substantial variation across settings. We therefore perform paired bootstrap analysis on two representative settings to quantify this variation.

Table 3 shows that the benefit of the role-

Variant	AUC	Hard	TPR@5	TPR@1
Strong MLP	0.956	0.936	0.831	0.611
No role sep.	0.954	0.939	0.852	0.626
DeMIA	0.960	0.942	0.857	0.628
WeakReg	0.960	0.942	0.856	0.623
OverReg	0.942	0.925	0.738	0.473

Table 2: Same-feature supervised comparisons and ablations averaged over nine settings. Hard denotes hard-subset AUC; TPR@5/1 denotes TPR at 5%/1% FPR.

separated objectives varies across the two representative datasets. On Wiki10k, the paired improvements over Strong MLP are positive for both AUC and TPR@1, with confidence intervals that exclude zero. On AGNews10k, the point estimates are also positive, but both confidence intervals include zero. The additional benefit over the same-feature supervised classifier is therefore clear on Wiki10k but weaker on AGNews10k. Full paired comparisons with REF/RCL and the no-role-separation ablation are reported in Appendix B.6.

The ablations show a complementary trade-off.

Dataset	AUC			TPR@1		
	DeMIA	Strong MLP	Δ [95% CI]	DeMIA	Strong MLP	Δ [95% CI]
Wiki10k	0.980	0.972	+0.008 [0.005, 0.011]	0.675	0.527	+0.148 [0.051, 0.285]
AGNews10k	0.819	0.802	+0.017 [-0.005, 0.038]	0.032	0.012	+0.020 [-0.011, 0.125]

Table 3: Paired comparison between DeMIA and the same-feature supervised Strong MLP on two representative Qwen2.5-1.5B datasets after two fine-tuning epochs. Both methods use the same diagnostic features and audit-label supervision. Δ denotes DeMIA minus Strong MLP; brackets report 95% confidence intervals obtained by paired bootstrap resampling.

WeakReg remains close to the main configuration, whereas OverReg reduces AUC to 0.942 and TPR@1 to 0.473. Stronger regularization therefore does not monotonically improve reliability; excessive suppression of nuisance information can also remove useful membership signal.

The low-FPR gains persist when operating points are fixed without using test labels. With thresholds selected from validation non-members, DeMIA reaches mean test TPRs of 0.861 and 0.727 at the nominal 5% and 1% FPR levels, with realized test FPRs of 0.056 and 0.016, respectively (Appendix Table 13). Across seeds 0–2, performance is stable, with AUC 0.962 ± 0.002 and TPR@5 0.863 ± 0.007 .

4.4 RQ3: Does the Signal Persist Under Stricter Reliability Tests?

Matched nuisance controls. High overall AUC can be inflated when members and non-members differ along simple nuisance dimensions. We therefore evaluate both hard non-member subsets and a stricter matched subset constructed using only pre-target statistics. The matched subset pairs members and non-members using token length, log character length, base loss, and base perplexity; target-model scores, DeMIA predictions, and membership logits are not used for construction. Matching reduces the mean absolute standardized mean difference from 0.030 to 0.003 and the maximum absolute SMD from 0.050 to 0.007 (Appendix Table 15).

On the matched subset, DeMIA retains an AUC of 0.968, compared with 0.959 for REF/RCL, 0.873 for Loss, 0.706 for calibrated Min-K, and 0.580 for Min-K% (Appendix Table 16). The original hard-subset evaluation shows the same qualitative pattern across model–dataset settings (Appendix Figure 3 and Table 17). Membership separability therefore remains strong after explicitly balancing the measured length and base-difficulty variables.

Method	AUC	TPR@5	TPR@1	Hard
Length only	0.500	0.044	0.008	0.502
Base loss	0.501	0.055	0.014	0.511
Length+base	0.502	0.053	0.013	0.501
Base MLP	0.495	0.052	0.011	0.509
Length+base MLP	0.490	0.045	0.009	0.491
DeMIA	0.960	0.853	0.624	0.942

Table 4: Blind controls using only pre-target nuisance features; DeMIA is included for reference. TPR@5/1 denotes TPR at 5%/1% FPR, and Hard denotes hard-subset AUC.

Blind shortcut controls. We next test whether membership can be predicted without any statistic from the fine-tuned target model. Controls based on length, base loss, and classifiers over base-model features remain near chance (Table 4). Text-only TF-IDF classifiers are also near chance across the representative settings (Appendix Table 18). In contrast, DeMIA retains 0.960 AUC and 0.624 TPR@1, showing that the observed separation is not reproduced by these pre-target or lightweight text-only controls.

Text-similarity and near-duplicate controls. We further remove the 20% most text-overlapping opposite-class examples using SimHash, with filtering based only on the input text. Table 5 shows that the text-only blind control remains near chance on both representative settings, while REF/RCL, Strong MLP, and DeMIA retain high membership separability. A separate TF-IDF/character n-gram filtering test yields the same qualitative pattern.

Calibration-construction robustness. Because our audit setting uses labeled calibration examples, we additionally examine sensitivity to calibration construction using a same-feature supervised probe over the 15-dimensional diagnostic vector. We keep the target model and held-out test split fixed and vary only the calibration non-members, considering random negatives, length/base-loss matched

Dataset	Method	AUC	TPR@1
Wiki10k	Text-only blind	0.509	0.010
	Ref/RCL	0.966	0.276
	Strong MLP	0.974	0.271
	DeMIA	0.974	0.289
AGNews10k	Text-only blind	0.526	0.012
	Ref/RCL	0.996	0.926
	Strong MLP	0.997	0.965
	DeMIA	0.997	0.970

Table 5: SimHash-20 near-duplicate stress test for Qwen2.5-1.5B after two fine-tuning epochs. The 20% most text-overlapping opposite-class examples are removed using input text only.

Construction	AUC	TPR@5	TPR@1
<i>Wiki10k</i>			
Random	.996±.001	.998±.002	.879±.043
Matched	.997±.001	.997±.003	.915±.025
Hard	.997±.001	.997±.002	.910±.022
Noisy10	.991±.001	.994±.006	.735±.034
<i>GitHub-CPP</i>			
Random	.968±.006	.823±.039	.326±.124
Matched	.970±.004	.826±.020	.430±.102
Hard	.971±.004	.828±.017	.441±.092
Noisy10	.963±.006	.795±.049	.278±.056

Table 6: Calibration-construction robustness over five seeds. Values are mean $\pm$ standard deviation. The target and test split are fixed while only the calibration non-member construction is varied.

negatives, hard negatives, and calibration sets with 10% label noise. Table 6 reports results over five seeds. On Wiki10k, performance remains stable across the random, matched, and hard-negative constructions. On GitHub-CPP, AUC and TPR@5 are similarly stable, while TPR@1 is more sensitive to calibration construction and is higher with matched or hard negatives than with random negatives. The 10% label-noise condition produces the largest degradation, particularly in the strict low-FPR regime.

Independent-rerun diagnostic. Finally, we test whether the learned audit signal persists across independent target runs. A probe trained on one Qwen2.5-1.5B/Wiki10k fine-tuning run achieves 0.801 AUC, 0.432 TPR@5, and 0.246 TPR@1 on an independently rerun target with the same model and dataset configuration, showing that the signal remains detectable beyond a single fine-tuning realization.

4.5 Summary of Findings

Together, the experiments show three main patterns. First, DeMIA consistently outperforms the evaluated score-based MIA baselines across model families, datasets, and fine-tuning stages, with particularly large gains in the low-FPR regime. Second, the benefit of the role-separated training objectives over a same-feature supervised classifier varies across settings: paired bootstrap analysis shows a clear gain in one representative setting, whereas the confidence intervals include zero in another, and excessive regularization substantially reduces low-FPR recall. Third, the membership signal persists under validation-threshold transfer, matched nuisance controls, blind pre-target controls, text-similarity and near-duplicate filtering, and an independently rerun target. A complementary calibration-construction analysis further shows stable AUC under random, matched, and hard-negative calibration, with larger degradation under label noise. Taken together, these results support evaluating membership signals in fine-tuned LLMs with explicit reliability diagnostics rather than aggregate attack accuracy alone.

5 Conclusion

We studied membership inference for fine-tuned LLMs from an evidentiary perspective, where high AUC alone can be misleading because member-non-member separability may also arise from length, base-model difficulty, topic imbalance, or other likelihood shortcuts. We proposed DeMIA, a role-separated evidence framework that learns membership and generalization channels from likelihood diagnostics of the base and target models using difficulty anchoring, matched-pair ranking, and adversarial length control. Across multiple models, datasets, and fine-tuning checkpoints, DeMIA consistently outperforms the evaluated score-based MIA baselines, particularly in the low-FPR regime, while its additional benefit over a same-feature supervised classifier is setting-dependent. The resulting membership signal remains detectable under stricter tests that control nuisance variation and textual overlap, and it persists across an independently rerun target. These findings show that reliable post-training membership auditing requires evaluating not only how well a score separates members from non-members, but also whether that separation remains under targeted reliability checks.

Limitations

Our results establish membership-relevant separability under the tested reliability controls, but they do not provide record-level causal attribution to training exposure or memorization. DeMIA uses matched ranking, base-difficulty anchoring, and adversarial length control to reduce reliance on selected nuisance cues, while the matched-subset, blind-control, text-similarity, and rerun evaluations test whether the resulting signal persists under complementary reliability diagnostics. These controls address several plausible shortcut explanations, but they do not exhaust the possible sources of member–non-member separability. Likewise, the role-separated representation should be interpreted functionally rather than as a statistically independent decomposition: the probing analysis in Appendix C.8 shows that membership and base-difficulty information remain recoverable from both channels.

Semantic and source-level confounding remain important limitations. We evaluate both TF-IDF/character n-gram filtering and SimHash-based near-duplicate filtering, and the membership signal remains strong after removing the most text-overlapping opposite-class examples. These tests reduce simple lexical-overlap and near-duplicate explanations, but they do not remove higher-level semantic similarity, paraphrastic or templated overlap, common source provenance, temporal artifacts, or similarity under learned embedding spaces. More aggressive semantic matching or filtering based on independently trained representations would provide a stronger test of these factors.

Our threat model assumes likelihood access to both a public base checkpoint and the fine-tuned target model, together with labeled audit examples for training or calibrating the attack. This setting is appropriate for controlled post-training audits and open-weight models, but it does not directly cover generation-only black-box APIs or candidate-free auditing. We evaluate sensitivity to calibration construction using random, matched, hard-negative, and noisy calibration sets; clean calibration constructions are comparatively stable, while label noise causes larger degradation in the strict low-FPR regime. This analysis uses the same likelihood diagnostic features as the audit probe and should not be interpreted as a full evaluation of DeMIA on every calibration regime. The independent-rerun experiment further shows that separability trans-

fers across two fine-tuning realizations of the same Qwen2.5-1.5B/Wiki10k configuration, but cross-model and cross-domain transfer remain open questions.

Our empirical comparison covers representative likelihood, reference/calibration, token-tail, neighborhood, DF-style, supervised same-feature, and WBC-compatible attacks under the access assumptions supported by our experimental artifacts. It does not cover every recent learned or automated MIA, particularly methods requiring different generation procedures, reference construction, shadow-model training, or transfer assumptions. The experiments also cover only three main text benchmarks and three model families. Different behavior may arise under instruction tuning, RLHF, parameter-efficient fine-tuning, substantially longer contexts, multilingual data, or specialized domains such as code, biomedical, and legal text. The GitHub-CPP experiment in this work is limited to calibration-construction sensitivity and is not a full code-domain evaluation of DeMIA. More broadly, membership inference should be interpreted as statistical evidence for data-use auditing; its strength depends on the access model, calibration data, evaluation protocol, and the extent to which alternative sources of separability have been controlled. Because false positives or distribution-shift failures could be over-interpreted as definitive evidence of training-data use, DeMIA should not be treated as standalone forensic or legal proof of record-level data use.

References

- Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. [Gpt-neo: Large scale autoregressive language modeling with mesh-tensorflow](#).
- 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.
- Bowen Chen, Namgi Han, and Yusuke Miyao. 2025. [A statistical and multi-perspective revisiting of the membership inference attack in large language models](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22854–22874.
- Yuetian Chen, Yuntao Du, Kaiyuan Zhang, Ashish Kundu, Charles Fleming, Bruno Ribeiro, and Ninghui Li. 2026. [Window-based membership inference attacks against fine-tuned large language mod-](#)

- els. In *35th USENIX Security Symposium (USENIX Security 26)*, Baltimore, MD. USENIX Association.
- Debeshee Das, Jie Zhang, and Florian Tramèr. 2025. [Blind baselines beat membership inference attacks for foundation models](#). In *2025 IEEE Security and Privacy Workshops (SPW)*, pages 118–125. IEEE.
- Michael Duan, Anshuman Suri, Niloofar Miresghallah, 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 *Conference on Language Modeling (COLM)*.
- André Vicente Duarte, Xuandong Zhao, Arlindo L. Oliveira, and Lei Li. 2024. DE-COP: Detecting copyrighted content in language models training data. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 11940–11956.
- 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](#). *Advances in Neural Information Processing Systems*, 37:134981–135010.
- Filippo Galli, Luca Melis, and Tommaso Cucinotta. 2024. Noisy neighbors: Efficient membership inference attacks against llms. In *Proceedings of the Fifth Workshop on Privacy in Natural Language Processing*, pages 1–6.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- 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.
- Masahiro Kaneko, Youmi Ma, Yuki Wata, and Naoaki Okazaki. 2025. [Sampling-based pseudo-likelihood for membership inference attacks](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8894–8907.
- Gyuwan Kim, Yang Li, Evangelia Spiliopoulou, Jie Ma, and William Yang Wang. 2026. [Detecting training data of large language models via expectation maximization](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1115–1129.
- Yanming Li, Cédric Eichler, Nicolas Anciaux, Alexandra Bensamoun, Lorena Gonzalez-Manzano, and Seifeddine Ghazzi. 2026. [Data provenance auditing of fine-tuned large language models with a text-preserving technique](#). In *International Conference on Machine Learning*.
- Zhexi Lu, Hongliang Chi, Nathalie Baracaldo, Swanand Ravindra Kadhe, Yuseok Jeon, and Lei Yu. 2026. [In-context probing for membership inference in fine-tuned language models](#). In *Network and Distributed System Security Symposium (NDSS)*.
- Pratyush Maini, Hengrui Jia, Nicolas Papernot, and Adam Dziedzić. 2024. [Llm dataset inference: Did you train on my dataset?](#) *Advances in Neural Information Processing Systems*, 37:124069–124092.
- Disha Makhija, Manoj Ghuhane, Arivazhagan, Vinayashankar Bannihatti Kumar, and Rashmi Gangadharaiyah. 2026. [Neural breadcrumbs: Membership inference attacks on llms through hidden state and attention pattern analysis](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5608–5620.
- 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.
- 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 24)*, pages 2369–2385.
- 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 *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 385–401. IEEE.
- Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *International Conference on Learning Representations*.
- Fatemehsadat Mireshghallah, Kartik Goyal, Archit Uniyal, Taylor Berg-Kirkpatrick, and Reza Shokri. 2022. [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.
- Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. [Don’t give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization](#). In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 1797–1807.
- Ashwinee Panda, Xinyu Tang, Christopher Choquette-Choo, Milad Nasr, and Prateek Mittal. 2025. Privacy auditing of large language models. In *International Conference on Learning Representations*.

- Haritz Puerto, Martin Gubri, Sangdoo Yun, and Seong Joon Oh. 2025. [Scaling up membership inference: When and how attacks succeed on large language models](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4165–4182.
- 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 *International Conference on Learning Representations*.
- Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. [Membership inference attacks against machine learning models](#). In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE.
- 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 Alexandre Sablayrolles. 2022. On the importance of difficulty calibration in membership inference attacks. In *International Conference on Learning Representations*.
- Johnny Wei, Ryan Wang, and Robin Jia. 2024. [Proving membership in llm pretraining data via data watermarks](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13306–13320.
- 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.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2024. [Qwen2.5 technical report](#). *arXiv preprint arXiv:2412.15115*.
- 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.
- Jie Zhang, Debeshee Das, Gautam Kamath, and Florian Tramèr. 2025. [Position: Membership inference attacks cannot prove that a model was trained on your data](#). In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 333–345. IEEE.
- Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28:649–657.
- Shuai Zhao, Linchao Zhu, Ruijie Quan, and Yi Yang. 2024. Protecting copyrighted material with unique identifiers in large language model training. *arXiv preprint arXiv:2403.15740*.

A AI Usage Statement

We used LLMs only to assist with the language of this manuscript, including proofreading, polishing, and improving clarity and readability. LLMs were not used to generate research ideas, formulate hypotheses, design methods, run experiments, analyze results, or draw scientific conclusions. All technical content, interpretations, and final decisions were produced and verified by the authors, who take full responsibility for the paper.

B Additional Experimental Details

We provide an anonymized implementation and reproducibility scripts at https://anonymous.4open.science/r/demia_attack-3628. The repository contains reproducibility scripts but does not redistribute original raw datasets or model weights. Public benchmark datasets and publicly released model checkpoints are used only for research evaluation under their respective licenses or terms of use. This section documents the experimental comparisons that support the main performance claims. We report baseline access assumptions, representative late-checkpoint values, additional same-access calibration baselines, and a WBC-compatible subset comparison. Reliability-oriented diagnostics are reported separately in appendix C.

B.1 Computational Environment and Hyperparameters

Computational environment. Experiments were conducted using an NVIDIA A100-SXM4-80GB GPU on a server running Ubuntu 24.04.2 LTS with NVIDIA driver 580.95.05, Python 3.10.19, PyTorch 2.7.1+cu118, two AMD EPYC 7B13 64-core CPUs, and 125 GiB RAM. The target model families contain approximately 125M parameters for GPT-Neo-125M, 1.5B parameters for Qwen2.5-1.5B, and 8B parameters for Llama-3.1-8B. The DeMIA audit probe is a lightweight MLP with approximately 36K trainable parameters in the main configuration.

Compute budget. The additional baseline and robustness evaluations required approximately 5 single-A100 GPU-hours. Including the representative target-model fine-tuning runs used for the reported results, the complete reported experiment suite required approximately 20–30 single-A100 GPU-hours.

Target-model fine-tuning. All target models are fine-tuned with a maximum sequence length of 256. The reported GPT-Neo-125M runs use batch size 32, gradient accumulation 1, learning rate 5×10^{-5} , weight decay 0, AdamW optimization, a linear learning-rate schedule, warmup ratio 0.03, seed 42, and setting-dependent training epochs. Qwen2.5-1.5B and Llama-3.1-8B use batch size 1, gradient accumulation 8, learning rate 1×10^{-5} , weight decay 0, bf16 precision, gradient checkpointing, and maximum sequence length 256. The representative Llama-3.1-8B runs use Adafactor with a linear learning-rate schedule, seed 42, and two training epochs.

DeMIA audit probe. The audit examples are partitioned into training, validation, and test splits using a 70/15/15 ratio. The probe is trained for 20 epochs with batch size 64, AdamW, learning rate 1×10^{-3} , and weight decay 0. The main architecture takes the 15-dimensional diagnostic vector as input and uses hidden dimension 128 and latent dimension 64. The membership and ranking objectives use unit weights with ranking margin 0.2; the length-adversarial and orthogonality weights are set to 2.0 and 0.1, respectively. The Strong MLP baseline uses the same diagnostic features and audit-label supervision as DeMIA, but replaces the role-separated architecture with a single unconstrained membership classifier.

B.2 Baseline Coverage and Access Assumptions

Our main comparisons use baselines that can be implemented under the same score-access setting as DeMIA: target-model likelihood statistics, base/target likelihood differences, token-tail summaries, and perturbation-based likelihood scores. This includes target loss, REF/RCL, Min-K%, calibrated Min-K, DF-style candidate calibration, neighborhood perturbation, and WBC-margin when row-aligned token-window traces are available.

Some previously proposed attacks require additional access that is outside our threat model.

For example, internal-representation attacks use hidden states or attention patterns (Makhija et al., 2026), while self-prompt calibration methods rely on generation APIs and surrogate/reference construction. Full DF-MIA-style attacks may additionally require distribution-free reference construction beyond the stored diagnostic features. We therefore report same-access approximations where possible, such as DF-style candidate calibration, and exclude methods whose required information is unavailable or not comparable under our score-access protocol. This keeps the main comparisons aligned with the information available to DeMIA.

B.3 Late-Checkpoint Numerical Results

Table 7 provides numerical values for the late-stage checkpoints shown in the main epoch curves. The table is intended as a reference for the plotted trends rather than as a separate model-selection procedure.

B.4 Additional Same-Access Calibration Baselines

Tables 8 and 9 show that stronger calibration improves over raw token-tail scores, but still leaves a substantial gap in the strict low-FPR regime.

B.5 WBC-Compatible Subset Comparison

WBC-margin requires aligned token-window likelihoods. Eight settings satisfy this requirement. Table 10 therefore reports all methods on this shared subset, providing a direct comparison under the same evaluation settings.

B.6 Paired Bootstrap and Confidence Intervals

To quantify uncertainty in the comparisons with strong supervised and likelihood-based references, we perform paired bootstrap resampling on the representative settings highlighted in the main text. The bootstrap is paired at the example level so that differences between methods are computed on the same resampled audit examples. We report both confidence intervals for individual methods and confidence intervals for paired performance differences. Positive paired differences indicate an advantage for DeMIA.

Table 11 shows that the absolute performance of DeMIA and Strong MLP is close on both settings, but the paired analysis provides a more direct test because it controls for variation induced by the sampled audit examples.

Model	Dataset	Epoch	DeMIA AUC	Loss AUC	Ref/RCL AUC	Min-K% AUC
GPT-Neo-125M	Wiki10k	5.0	0.9758	0.7087	0.9555	0.7559
GPT-Neo-125M	AGNews10k	5.0	0.9434	0.8404	0.8967	0.8672
GPT-Neo-125M	XSum10k	5.0	0.9708	0.7481	0.9304	0.7762
Qwen2.5-1.5B	Wiki10k	3.0	0.9948	0.7510	0.9025	0.8105
Qwen2.5-1.5B	AGNews10k	3.0	0.9991	0.7867	0.9856	0.8575
Qwen2.5-1.5B	XSum10k	3.0	0.9981	0.9169	0.9746	0.9515
Llama-3.1-8B	Wiki10k	3.0	0.9987	0.8190	0.9843	0.8833
Llama-3.1-8B	AGNews10k	2.0	0.9940	0.7808	0.9372	0.8369
Llama-3.1-8B	XSum10k	3.0	0.9986	0.9092	0.9854	0.9363

Table 7: Representative late-checkpoint AUC values corresponding to the main epoch curves. The table reports the late-stage checkpoints used for the numerical summary; full epoch trends are shown in Figure 2.

Method	AUC	TPR@5	TPR@1	Hard
Cal-MinK-HGB	0.6240	0.1330	0.0430	0.6297
Cal-MinK-HGB-3	0.6170	0.1178	0.0280	0.6199
Cal-MinK-LR	0.5809	0.0762	0.0153	0.5859
Min-K%	0.5502	0.0629	0.0163	0.5733
Base-target Min-K%	0.5436	0.0708	0.0164	0.5178
Delta Min-K%	0.5436	0.0708	0.0164	0.5178
Base-target Min-5%	0.5295	0.0598	0.0135	0.4962

Table 8: Calibrated Min-K variants. Hard denotes hard-subset AUC. The HGB variant is the strongest same-access token-tail calibration baseline in the available feature set.

Method	AUC	TPR@5	TPR@1	Hard
DF-style LR	0.7581	0.1864	0.0272	0.7442
DF-style robust	0.7043	0.0841	0.0147	0.6786
DF-style HGB	0.6927	0.1677	0.0559	0.6559

Table 9: DF-style same-access candidate-calibration baselines averaged over 37 settings. These variants use target/base diagnostic features and candidate-pool pseudo-labels rather than training an additional reference LLM.

The paired results reveal substantial variation across the two representative datasets. On Wiki10k, the improvements over Strong MLP are separated from zero for both AUC and TPR@1, with gains of +0.008 and +0.148, respectively. The improvements over Ref/RCL are also separated from zero. On AGNews10k, by contrast, the confidence intervals for the differences relative to Strong MLP and Ref/RCL include zero, and the no-role-separation variant is essentially tied with DeMIA in AUC. These results support the interpretation in Section 4.3: the benefit of the role-separated objectives is setting-dependent, with a clear gain on Wiki10k but weaker evidence on AGNews10k.

The stronger regularization ablation provides a complementary observation. As shown in Table 2, OverReg reduces TPR@1 from 0.628 to 0.473.

This indicates that increasing the strength of nuisance suppression does not monotonically improve membership inference and can remove useful membership information.

C Additional Reliability Diagnostics

This section provides additional diagnostics for evaluating the reliability of the membership signal reported in the main text. We examine validation-threshold transfer, hard and matched subsets, blind and text-similarity controls, calibration-construction sensitivity, full shortcut diagnostics, and representation-level probing. Together, these analyses test whether the observed membership separation persists when specific nuisance factors and alternative explanations are reduced or isolated.

C.1 Validation-Threshold Low-FPR Transfer

The main text reports TPR at low false-positive-rate operating points as a key auditing metric. To test whether these results depend on selecting operating points from the held-out test ROC, we additionally evaluate validation-threshold transfer. For each method and setting, we choose a threshold using only validation non-members to target a nominal FPR of 5% or 1%, and then apply that threshold unchanged to the held-out test split. We report both the resulting test TPR and the realized test FPR. WBC-margin is evaluated separately on its WBC-compatible subset in Appendix B.5, since it requires available row-aligned token-window traces.

Table 13 shows that DeMIA retains strong low-FPR recall under validation-selected thresholds. At the nominal 5% FPR operating point, DeMIA achieves a TPR of 0.861 with a realized test FPR of 0.056, compared with 0.683 for Ref/RCL at the same realized FPR. At the nominal 1% FPR operating point, DeMIA achieves a TPR of 0.727 with a realized test FPR of 0.016, compared with 0.414

Method	#Settings	AUC	TPR@5%FPR	TPR@1%FPR	Hard AUC
Loss	8	0.843	0.436	0.139	0.860
Ref/RCL	8	0.960	0.722	0.297	0.953
Min-K%	8	0.591	0.068	0.018	0.610
Cal-MinK-HGB	8	0.704	0.162	0.034	0.691
DF-style LR	8	0.860	0.242	0.027	0.835
WBC-margin	8	0.946	0.621	0.177	0.945
DeMIA	8	0.988	0.943	0.699	0.982

Table 10: WBC-compatible shared-subset comparison on the eight settings satisfying the aligned token-window requirement of WBC-margin. All methods are evaluated on the same settings; DeMIA achieves the highest AUC, low-FPR recall, and Hard AUC.

Dataset	Method	AUC [95% CI]	TPR@1 [95% CI]
Wiki10k	DeMIA	0.980 [0.975, 0.985]	0.675 [0.453, 0.790]
	Strong MLP	0.972 [0.967, 0.977]	0.527 [0.402, 0.655]
	Ref/RCL	0.967 [0.959, 0.974]	0.358 [0.186, 0.513]
AGNews10k	DeMIA	0.819 [0.771, 0.864]	0.032 [0.000, 0.209]
	Strong MLP	0.802 [0.755, 0.849]	0.012 [0.000, 0.095]
	No role sep.	0.820 [0.773, 0.866]	0.013

Table 11: Bootstrap confidence intervals for individual methods on two representative Qwen2.5-1.5B datasets after two fine-tuning epochs. TPR@1 denotes TPR at 1% FPR. The no-role-separation TPR@1 result is reported as a point estimate because its bootstrap interval was unavailable.

Setting	Comparison	Δ AUC [95% CI]	Δ TPR@1 [95% CI]
Wiki10k	DeMIA – Ref/RCL	+0.013 [0.009, 0.017]	+0.317 [0.168, 0.411]
	DeMIA – Strong MLP	+0.008 [0.005, 0.011]	+0.148 [0.051, 0.285]
AGNews10k	DeMIA – Ref/RCL	+0.004 [-0.001, 0.009]	+0.013 [-0.095, 0.044]
	DeMIA – Strong MLP	+0.017 [-0.005, 0.038]	+0.020 [-0.011, 0.125]
	DeMIA – No role sep.	-0.001 [-0.009, 0.006]	+0.019

Table 12: Paired-bootstrap differences between DeMIA and the principal comparison methods on two representative Qwen2.5-1.5B settings after two fine-tuning epochs. Positive values favor DeMIA. The TPR@1 interval for the no-role-separation comparison was not available and is therefore reported only as a point difference.

Method	Test AUC	Val-thr TPR@5% FPR	Val-thr TPR@1% FPR
Loss	0.819	0.443 (0.056)	0.130 (0.011)
Ref/RCL	0.929	0.683 (0.056)	0.414 (0.017)
Min-K%	0.582	0.080 (0.060)	0.016 (0.009)
Cal-MinK-HGB	0.683	0.158 (0.052)	0.028 (0.010)
DeMIA	0.960	0.861 (0.056)	0.727 (0.016)

Table 13: Validation-threshold low-FPR transfer. Thresholds are selected from validation non-member score quantiles and applied unchanged to the held-out test split. Parentheses report the realized test FPR.

for Ref/RCL with a realized test FPR of 0.017. The same ordering therefore persists when thresholds are selected from validation data rather than from the held-out test ROC.

C.2 Hard-Subset Construction and Matched Evaluation

The hard-subset diagnostics test whether membership separability persists after easy non-members

are removed. We report both the original hard-subset construction diagnostics and the stricter matched hard-subset construction used in the main text. All hard subsets are constructed using pre-target nuisance statistics; target-model likelihoods, target-base shifts, DeMIA scores, and membership logits are not used.

Table 14 documents the original hard-subset stress test. It is useful for transparency, but the

Model	Dataset	Epoch	#Hard NM	Length SMD		Base-loss SMD		Base-PPL SMD	
				Before	After	Before	After	Before	After
GPT-Neo-125M	AGNews10k	1.5	79	-0.04	0.13	-0.00	-0.10	-0.00	-0.10
	Wiki10k	2.0	81	-0.04	0.10	0.01	-0.02	0.01	-0.03
Llama-3.1-8B	AGNews10k	2.0	79	-0.04	0.20	-0.05	-0.23	-0.04	-0.23
	Wiki10k	2.0	82	-0.05	0.05	0.02	-0.04	0.01	-0.05
	XSum10k	2.0	77	-0.03	0.02	-0.00	-0.12	-0.00	-0.12
Qwen2.5-1.5B	AGNews10k	2.0	79	-0.03	0.08	-0.01	-0.14	-0.00	-0.14
	Wiki10k	2.0	81	-0.05	0.10	0.03	-0.12	0.03	-0.12

Table 14: Original hard-subset construction diagnostics for representative settings with a usable hard non-member pool. Each setting contains 745 members and 755 non-members. The construction uses length, base loss, and base perplexity as matching variables. SMD denotes standardized mean difference computed using the training-split scale. Two XSum settings produced no hard non-members under the original construction and are omitted; the stricter matched-subset construction covers all representative settings in Table 15.

Statistic	Original test set	Matched hard subset	Ratio
Mean absolute SMD	0.030	0.003	0.095
Max absolute SMD	0.050	0.007	0.132
Matched pairs per setting	—	76.0	—

Table 15: Matched hard-subset construction diagnostics. Members and non-members are paired using only pre-target controls: token length, log character length, base loss, and base perplexity. The reduction in SMD shows that the matched subset aggressively balances the most direct nuisance variables before attack scores are evaluated.

main reliability interpretation relies on the stricter matched construction below.

Table 15 reports balance statistics for the matched hard subset. The standardized mean differences decrease substantially after matching, showing that the matched subset directly reduces the most immediate length and base-difficulty gaps. This construction is a stress test for shortcut explanations, not a guarantee that all semantic or distributional confounders are removed.

Table 16 reports performance on the matched hard subset. DeMIA remains strong after matching, reaching 0.968 matched hard AUC and outperforming Loss, Min-K%, calibrated Min-K, and Ref/RCL. This indicates that the membership signal does not vanish when direct length and base-difficulty differences are explicitly reduced.

C.3 Hard-Subset All-Baseline Comparison

In addition to the matched hard subset, we report the original hard-subset comparison across major baselines. This analysis examines how scalar likelihood, calibrated token-tail, DF-style, localized-

window, and learned diagnostic methods behave when easy non-members are replaced with harder alternatives.

Table 17 shows that hard-subset evaluation remains challenging for scalar likelihood, calibrated token-tail, and DF-style baselines. DeMIA retains strong hard-subset separability, with a Hard AUC of 0.942, TPR@5%FPR of 0.768, and TPR@1%FPR of 0.569. WBC-margin reaches a comparable Hard AUC of 0.945 on the six settings satisfying its aligned token-window requirement. For comparison under identical setting coverage, Appendix B.5 reports the WBC-compatible shared-subset evaluation.

C.4 Blind Shortcut and Topic Controls

Blind controls test whether membership labels can be predicted without using target-model likelihoods or target-base likelihood shifts. We evaluate controls based on surface length, base-model loss, base-side MLP features, and raw-text TF-IDF features. These controls use only pre-target or text-only information and therefore cannot exploit fine-tuning-induced likelihood changes.

Table 18 reports text-only TF-IDF controls. These controls provide a lightweight check for topic or lexical separability in the member/non-member split. Their near-chance performance weakens a direct text-topic shortcut explanation, although it does not rule out all possible semantic confounders.

Method	AUC	TPR@5%	TPR@1%	Matched hard AUC
Loss	0.819	0.395	0.125	0.873
Ref/RCL	0.929	0.656	0.266	0.959
Min-K%	0.582	0.067	0.017	0.580
Cal-MinK-HGB	0.683	0.150	0.032	0.706
DeMIA	0.960	0.853	0.624	0.968

Table 16: Matched hard-subset comparison among formal likelihood-based baselines. Unlike the original hard split, this diagnostic evaluates matched member–non-member pairs constructed only from pre-target nuisance statistics. The result checks whether each method remains effective after direct length and base-difficulty differences are explicitly balanced.

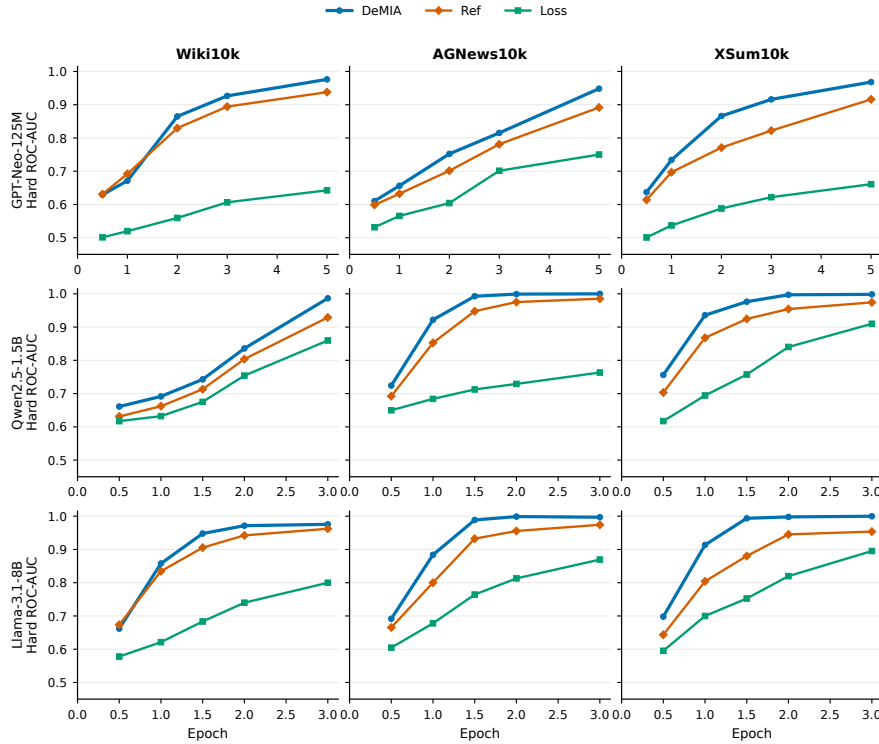


Figure 3: AUC on the original hard membership inference subsets. Easy non-members are replaced with harder alternatives. DeMIA retains strong separability across the evaluated settings.

Method	#Settings	Hard AUC	Hard TPR@5%FPR	Hard TPR@1%FPR
Loss	7	0.831	0.420	0.230
Ref/RCL	7	0.916	0.639	0.285
Min-K%	7	0.606	0.073	0.014
Cal-MinK-HGB	7	0.669	0.164	0.061
DF-style LR	7	0.809	0.195	0.024
WBC-margin	6	0.945	—	—
DeMIA	7	0.942	0.768	0.569

Table 17: Original hard-subset comparison across major baselines. Metrics are averaged over settings with a nonempty original hard non-member pool. WBC-margin is evaluated on the six settings that also satisfy its aligned token-window requirement; threshold-specific hard-subset TPRs are not reported for WBC-margin.

C.5 Text-Similarity and Near-Duplicate Stress Tests

The text-only controls in Appendix C.4 test whether coarse lexical or topic information alone is sufficient to predict membership. We further

examine whether membership separability is concentrated among highly similar or near-duplicate member–non-member examples. Both filtering procedures are constructed using input text only and do not use target-model likelihoods, DeMIA scores,

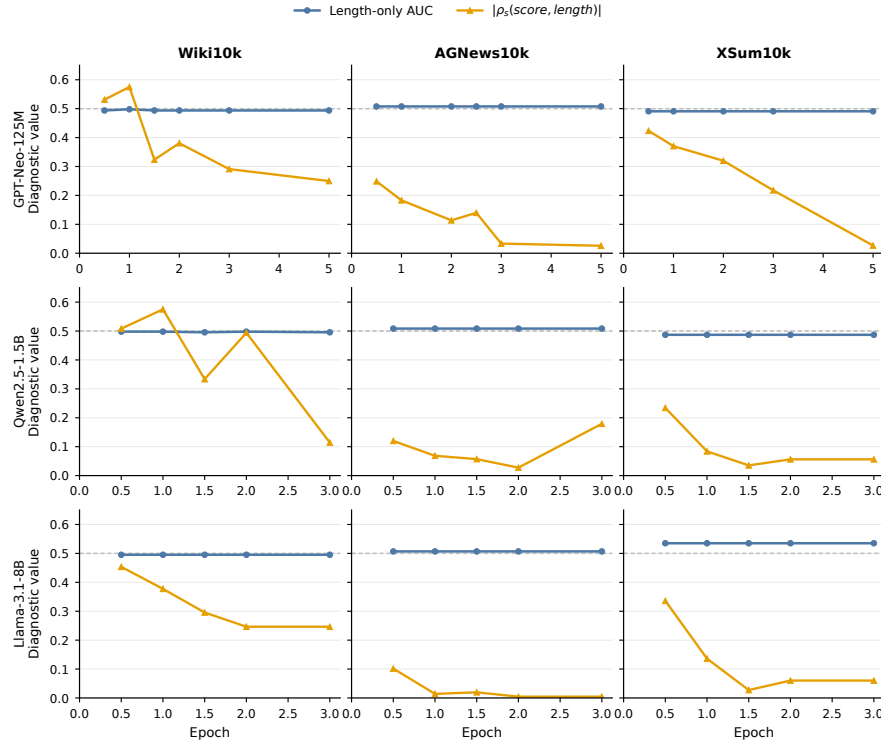


Figure 4: Length-only and score-length diagnostics. Length-only AUC remains close to chance across model families and datasets, while score-length correlations vary by setting. These diagnostics support the conclusion that the main gains are not explained by a direct length rule.

Method	#Settings	AUC	TPR@5%FPR	TPR@1%FPR	Hard AUC
TF-IDF char LR	9	0.494	0.052	0.008	0.512
TF-IDF word LR	9	0.505	0.049	0.010	0.528

Table 18: Text-only blind controls using raw text. TF-IDF classifiers do not use target or base likelihoods. Their near-chance performance weakens direct lexical or topic-separability explanations on the representative settings.

membership logits, or model-internal features.

We consider two complementary procedures. The first computes nearest-opposite similarity from TF-IDF features with character n-grams and removes the 20% most similar opposite-class examples. The second uses SimHash to remove the 20% most text-overlapping opposite-class cases. In both cases, the filtered evaluation subset is fixed before membership scores are examined.

TF-IDF and character n-gram filtering.

Table 19 reports the complementary TF-IDF/character n-gram stress test. The text-only blind classifier remains near chance on both datasets, while REF/RCL, Strong MLP, and DeMIA retain high membership separability.

Dataset	Method	AUC	TPR@1
Wiki10k	Text-only blind	0.492	0.008
	Ref/RCL	0.984	0.434
	Strong MLP	0.989	0.580
	DeMIA	0.990	0.618
AGNews10k	Text-only blind	0.483	0.012
	Ref/RCL	0.999	0.986
	Strong MLP	1.000	0.998
	DeMIA	1.000	0.998

Table 19: TF-IDF/character n-gram similarity filtering. The 20% most similar opposite-class examples are removed using input text only. Values are rounded to three decimals.

SimHash filtering. We independently repeat the stress test using SimHash-style near-duplicate filtering. The complete SimHash results are reported in Table 5 in the main text. The text-only blind control remains near chance on both datasets, while REF/RCL, Strong MLP, and DeMIA retain substantial membership separability after the most text-overlapping opposite-class cases are removed.

Across the two filtering procedures, membership separability persists after direct lexical overlap and near-duplicate cases are reduced. The effect is weaker under SimHash on Wiki10k, particularly

at TPR@1, but the signal remains substantially above the text-only blind control. These tests reduce simple lexical-overlap and near-duplicate explanations, while higher-level semantic similarity, shared source provenance, temporal effects, and similarity under learned embedding spaces remain outside the scope of these text-only filters.

C.6 Calibration-Construction Details

The calibration-construction experiment in Table 6 examines whether the learned audit signal is sensitive to the choice of calibration non-members. Across all conditions, the target model and held-out test split are fixed, and only the calibration pool is changed. The probe uses the same 15-dimensional likelihood diagnostic vector as in the main audit setting; raw token or character length is not included as an input feature.

We consider four calibration conditions. Random uses non-members sampled from the available calibration pool. Matched selects non-members with similar sequence length and base-model loss. Hard uses the harder non-member construction from the available candidate pool, while Noisy10 corrupts 10% of the calibration labels to test sensitivity to imperfect supervision. Each condition is evaluated over five seeds.

The analysis is performed on Wiki10k and on a code-like GitHub-CPP diagnostic corpus. GitHub-CPP is included only to test whether calibration sensitivity changes in a substantially different code domain; it is not used as evidence of full cross-domain transfer of DeMIA. The quantitative results are reported in Table 6: the main distinction is that clean calibration constructions are relatively stable in AUC, whereas strict low-FPR recall is more sensitive on GitHub-CPP and label noise produces the clearest degradation.

C.7 Full Shortcut Diagnostic Matrix

The main text reports aggregate shortcut diagnostics. Here we provide the full model–dataset matrix, including length-only AUC, major baseline AUCs, WBC-margin where available, and score–length correlations.

Table 20 shows that length-only AUC remains close to chance across most settings. It also shows that WBC-margin is a strong localized likelihood baseline in some settings, motivating the separate WBC-compatible subset comparison in Appendix B.5.

C.8 Channel Probing Diagnostics

Channel probing provides an auxiliary view of what information remains linearly recoverable from the learned representations. We train lightweight probes to predict membership labels, length bins, base loss, and base perplexity from each available representation. The probes are fitted only for analysis and are not used to produce the membership scores reported in the main experiments.

Table 21 shows substantial shared recoverability across the two channels. Membership information remains strongly recoverable from both z_{mem} and z_{gen} , and base-side difficulty information is also present in both representations. This is consistent with the strong correlations among the likelihood diagnostics used as input and indicates that the learned channels should not be interpreted as statistically independent representations.

The purpose of the role-separated objectives is instead to assign different functions to the two channels during training: z_{mem} supplies the final membership score, whereas z_{gen} supports the base-difficulty anchoring objective. The probing results therefore complement, rather than determine, the reliability analysis. Evidence about shortcut dependence comes primarily from the matched and hard-subset evaluations, blind pre-target controls, and the additional stress tests in Appendix C.5.

C.9 Group-Level Length Diagnostics

We additionally examine token-length distributions for members, easy non-members, and hard non-members. This analysis is descriptive rather than an attack evaluation: it provides a group-level view of whether the hard-subset construction is associated with a pronounced difference in sequence length.

Figure 5 does not show a pronounced group-level length separation that tracks the membership split across the evaluated settings. Length nevertheless remains a relevant nuisance variable because it can correlate with both base difficulty and likelihood statistics. We therefore treat this figure as an auxiliary diagnostic; the stronger evidence against direct length shortcuts comes from the matched evaluation in Section 4.4 and the length-only blind controls in Table 4.

Model	Dataset	Len	Loss	Ref	Min-K	Cal-MK	DF	WBC	Neigh.	DeMIA	$ \rho_\ell $
GPT-Neo-125M	AGNews10k	0.5077	0.6750	0.7161	0.6898	0.5064	0.5565	0.6768	0.6339	0.7544	0.1240
GPT-Neo-125M	Wiki10k	0.4945	0.5853	0.8265	0.6107	0.5642	0.8215	0.9051	0.5518	0.8638	0.3919
GPT-Neo-125M	XSum10k	0.4909	0.6074	0.8138	0.6257	0.5330	0.6426	0.7862	0.5476	0.8437	0.2715
Llama-3.1-8B	AGNews10k	0.5069	0.7479	0.8562	0.7782	0.6305	0.7489	0.8315	0.6947	0.9011	0.0289
Llama-3.1-8B	Wiki10k	0.4953	0.6995	0.8784	0.7433	0.7295	0.8251	0.9379	0.6432	0.9119	0.3239
Llama-3.1-8B	XSum10k	0.5351	0.7572	0.8868	0.7993	0.7577	0.7882	0.8488	0.6969	0.9202	0.1240
Qwen2.5-1.5B	AGNews10k	0.5085	0.6847	0.8739	0.7466	0.6091	0.7432	0.8105	0.6429	0.9216	0.0904
Qwen2.5-1.5B	Wiki10k	0.4970	0.5875	0.7229	0.6147	0.6324	0.7911	0.8412	0.5600	0.7972	0.4050
Qwen2.5-1.5B	XSum10k	0.4870	0.7899	0.8937	0.8229	0.7080	0.8458	0.8050	0.7238	0.9343	0.0931

Table 20: Full shortcut diagnostic matrix by model–dataset pair. Len denotes length-only AUC; Cal-MK denotes calibrated Min-K; DF denotes DF-style candidate calibration; Neigh. denotes neighborhood perturbation; $|\rho_\ell|$ denotes absolute score–length correlation. WBC is reported only where row-aligned token-window traces are recoverable.

Method	Representation	# Settings	Mem. AUC	Len. Acc.	Base-loss R^2	Base-PPL R^2
DeMIA-Zmem	z_{gen}	10	0.964	0.573	0.983	0.301
DeMIA-Zmem	z_{mem}	10	0.964	0.566	0.976	0.457
DeMIA-Zmem-8d	z_{gen}	10	0.961	0.520	0.760	0.549
DeMIA-Zmem-8d	z_{mem}	10	0.962	0.520	0.715	0.362
DeMIA	z_{gen}	11	0.965	0.573	0.981	-0.040
DeMIA	z_{mem}	11	0.965	0.564	0.976	0.444
No role sep.	z_{gen}	10	0.964	0.571	0.945	0.491
No role sep.	z_{mem}	10	0.964	0.573	0.946	0.498
Strong MLP	comparison repr.	11	0.963	0.556	0.952	0.496

Table 21: Linear probing of the learned representations. DeMIA-Zmem and DeMIA-Zmem-8d denote auxiliary compact-channel variants, while DeMIA denotes the main configuration. Strong MLP is included as a same-feature supervised reference without explicit membership and generalization channels. The probes measure residual linear recoverability of membership, length, and base-side difficulty information from each representation.

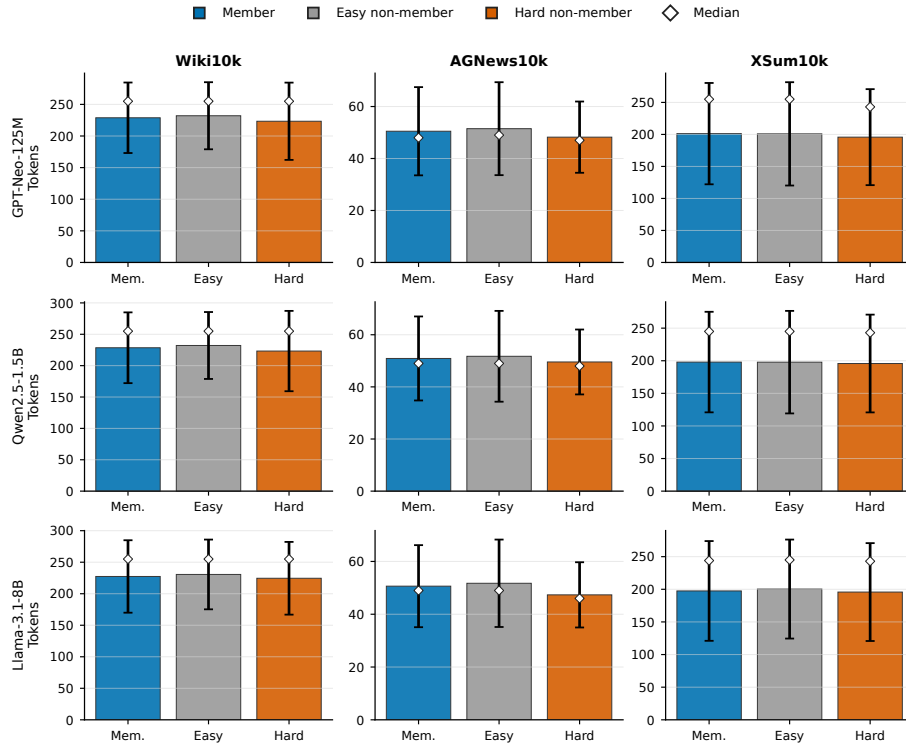


Figure 5: Token-length distributions across members, easy non-members, and hard non-members. Bars indicate group means with standard-deviation error bars, and diamond markers indicate medians.