ATTENMIA: READING MEMBERSHIP OFF THE ATTENTION MAPS OF LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) can reveal through attacks such as Membership Inference Attacks (MIAs) whether a sequence appeared in their training data, creating privacy risks. Many existing MIAs for LLMs derive their signal from model outputs. Motivated by recent research showing the important role attention plays in information flow within a model, in this paper, we instead ask whether membership signatures are visible in the model's attention patterns. We introduce AttenMIA, a white-box MIA with two variants: Transitional features capture cross-layer attention changes, while Perturbation features measure attention shifts under input edits. Each variant uses a lightweight supervised classifier to convert its features into membership scores. We evaluate on MIMIR, a benchmark constructed to reduce known distributional pitfalls in membership evaluation. Across seven domains and three Pythia model sizes, AttenMIA (Perturbation) achieves the highest ROC AUC in 20 of 21 settings, with domain-averaged AUCs of 0.83–0.85, compared with 0.71 for the strongest evaluated baseline. When each domain is held out from attack training and selection, AttenMIA (Transitional) performs best among the evaluated methods across all three model sizes, averaging 0.73 ROC AUC and 26.7% TPR@1%FPR across domains and models. Under the same protocol, a supervised hidden-state control achieves an average AUC of 0.54, showing that the advantage arises from attention features. GPT-2 fine-tuning experiment shows that cross-layer transitional attention features, and perturbation-response features have different attention gradients for members and non-members. These results substantially improve on SoTA membership inference attacks, and demonstrate that self-attention provides a useful signal for training-membership inference that transfers across the evaluated domains.

1 INTRODUCTION

Membership inference attacks (MIAs) threaten training-data privacy by revealing whether a candidate example was used to train a model (Shokri et al., 2017; Hu et al., 2022; Carlini et al., 2019; 2021; Lukas et al., 2023). Most of existing MIAs for large language models (LLMs) rely on output-based signals, such as sequence likelihood or its changes under input perturbations (Mattern et al., 2023). White-box attacks have also demonstrated that internal features and gradients can reveal membership information beyond model predictions (Nasr et al., 2019; Leino & Fredrikson, 2020). These findings motivate a more specific question: which aspects of internal computation provide useful membership signals? We investigate self-attention, whose weights govern how information from different token positions is combined within each attention head (Vaswani et al., 2017). Our motivation draws on the relationship between attention and memory: in human visual search, learned contextual associations can guide attention even when explicit recognition is at chance (Chun & Jiang, 1998). A computational connection also exists: transformer attention can be formulated as an associative-memory update in modern Hopfield networks (Ramsauer et al., 2020). These observations motivate us to investigate whether training exposure leaves detectable traces in how attention distributions evolve across layers and respond to input changes. Our central question is:

Can structured attention behavior provide more accurate and reliable membership inference signal that generalizes to unseen domains?

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We introduce ATTENMIA, a white-box membership inference attack that leverages the model's behaviour in the attention space to infer membership. To effectively exploit the attention signal, we propose two variations based on two hypotheses on structured representations of self-attention with regards to memorization, which we detail in Section 2.2. **(i)** *Transitional* features capture how the attention distribution of heads changes between adjacent transformer layers. **(ii)** *Perturbation* features capture how these distributions respond to controlled input edits: prefix insertion, word dropping, and word replacement. Both representations exploit measurements across layers and heads, allowing the attack to use patterns distributed throughout the network. For each representation, we train a separate lightweight classifier on member and non-member examples to produce membership scores. Feature extraction requires only forward passes through the target model, without modifying its parameters. An overview of our approach is presented in Figure 1.

In our experimental setup, we prioritize MIMIR, whose construction reduces known distributional confounds in membership evaluation (Duan et al., 2024). We evaluate across seven domains and three Pythia model sizes (Biderman et al., 2023). In the in-domain evaluation, ATTENMIA (Perturbation) achieves the highest ROC AUC in 20 of 21 model–domain settings, with domain-averaged AUCs of 0.83–0.85, compared with 0.71 for the strongest evaluated baseline. We next test whether these attention-based signals generalize to unseen domains, where held-out MIMIR domain is excluded from training data of MIA classifier. ATTENMIA (Transitional) achieves the strongest domain-averaged performance among the evaluated methods at all three Pythia sizes, with an overall average ROC AUC of 0.73 and a TPR of 26.7% at 1% FPR. Under the same protocol, a supervised control based on hidden-state activation statistics achieves an average AUC of 0.54. These results show that the proposed attention features support cross-domain membership inference more effectively than the evaluated hidden-state representation.

As a complementary analysis, we examine the association between ATTENMIA scores and source overlap in generated text. On a selected subset of high- and low-overlap GPT-2 generations, these scores correlate more strongly with source-text ROUGE-L ($r = 0.48$) than the evaluated likelihood and compression baselines (best $r = 0.32$).

Our contributions are:

- We introduce ATTENMIA, a white-box MIA with two structured attention representations: cross-layer transitions and responses to controlled input perturbations. Our approach proves the existence of a reliable membership signal in the attention space.
- We demonstrate strong in-domain performance on MIMIR and target-blind cross-domain transfer relative to the evaluated output-based baselines, and outperform the supervised hidden-state control in average cross-domain ROC AUC.
- We provide evidence of membership-associated differences in the sensitivities of prediction loss and the proposed attention features.

2 ATTENMIA: SETUP AND METHOD

AttenMIA is a supervised membership inference attack that uses the self-attention of a target transformer language model. We first define the setting, then motivate and describe two feature families: cross-layer attention transitions and attention responses to perturbations.

2.1 THREAT MODEL AND ATTACK SETTING

For a candidate sequence x, let $m(x) \in \{0, 1\}$ indicate whether x belongs to the training data. The adversary produces a continuous membership score $s(x)$, with larger values indicating stronger evidence of membership. We evaluate these scores using ROC AUC and TPR at 1% FPR.

Adversary access. We consider white-box membership inference, as studied in prior work (Nasr et al., 2019; Leino & Fredrikson, 2020; Sablayrolles et al., 2019). Our adversary can run the target model and inspect its forward-pass attention maps. For an input $x = (x_1, \ldots, x_T)$, a model with L layers and H attention heads per layer provides $\{A^{\ell,h}(x)\}_{\ell=1,h=1}^{L,H}$, where $A^{\ell,h}(x) \in \mathbb{R}^{T\times T}$ is the normalized attention matrix of head h in layer ℓ. Each row describes how a token distributes

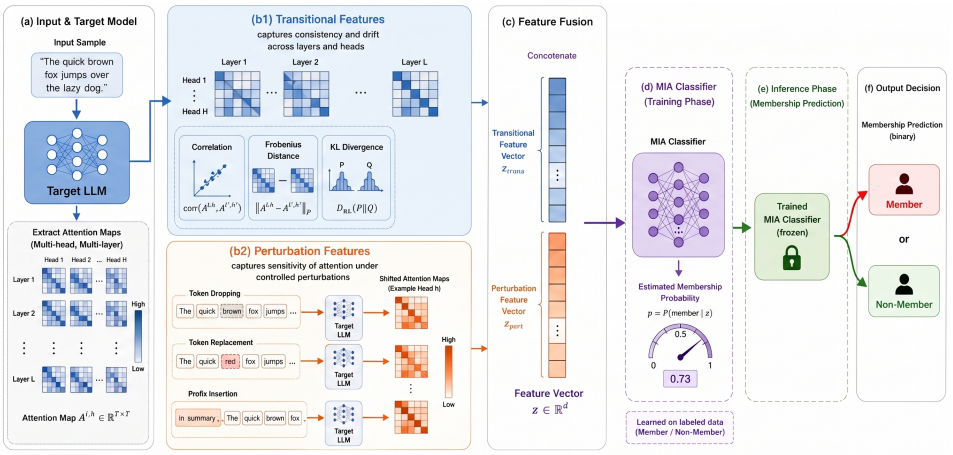


Figure 1: Overview of ATTENMIA. *Transitional* features compare attention distributions between same-index heads in adjacent layers, while *Perturbation* features measure attention changes under controlled input edits. Each representation is used by a separate lightweight classifier, trained on member and non-member examples, to produce a membership score. The resulting variants are ATTENMIA (Transitional) and ATTENMIA (Perturbation).

attention over itself and preceding tokens; padding positions are excluded. ATTENMIA extracts these maps through forward passes without modifying or fine-tuning the target model.

Attack knowledge. We assume access to labeled examples whose membership in the target model's training data is known. These examples are used to train the membership classifier and are separate from the candidates used to evaluate it. Their availability is an assumption of our supervised attack setting. In the in-domain evaluation, attack-training and test examples come from the same domain and are separated by held-out folds. In the cross-domain evaluation, labeled examples come from source domains, while the held-out target domain is excluded from attack preprocessing, training, and model selection.

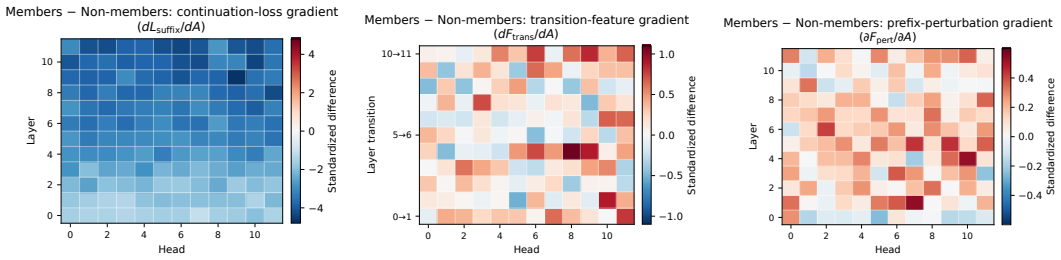
2.2 MOTIVATION AND EMPIRICAL OBSERVATIONS

Why should attention reveal training membership? And what tool should we use to exploit it? We hypothesize that training exposure affects how attention distributions change across layers and how attention patterns respond to input edits, discriminately between members and non-members. Specifically, our intuition is based on the following hypotheses:

H1: Cross-layer transitions. Structural relationships between attention distributions in adjacent layers contain discriminative information about training membership. We expect training exposure to systematically alter these transitions

H2: Perturbation response. Changes in attention distributions under controlled input edits contain information about training membership; Members' attention patterns should be more impacted by perturbations than non-members.

These two forms of attention behavior motivate our Transitional and Perturbation features, respectively. To investigate this hypothesis, we fine-tune pretrained GPT-2 (Radford et al., 2019) on synthetic canaries and compare them with held-out canaries drawn from the same distribution. Membership in this experiment refers to inclusion in the controlled fine-tuning corpus. We examine three sensitivity measurements: the gradient of the mean continuation-token loss, the gradient of the cross-layer Transitional KL statistic, and the gradient of the prefix-insertion Perturbation KL statistic. The first two are computed through the model's computation graph. For the third, we treat the original and perturbed attention maps as independent inputs and differentiate the KL statistic directly. We summarize gradients using their root-mean-square (RMS) magnitude over valid causal attention en-



(a) (b) (c)

Figure 2: Attention-map sensitivity differences between synthetic fine-tuning members and non-members in GPT-2. Heatmaps show standardized member-minus-non-member differences in gradient-magnitude summaries for (a) the mean continuation-loss, (b) the adjacent-layer Transitional KL statistic, and (c) the prefix-insertion Perturbation KL statistic. (a, b) use gradients through the model's computation graph; panel (c) uses direct derivatives with respect to detached original and perturbed attention maps. Gradient magnitudes are summarized by RMS over valid causal entries, combining both maps for (b, c). These measurements are diagnostic and are not attack features.

tries, combining entries from both maps for the feature-based measurements. These gradients are used only for diagnosis; ATTENMIA itself extracts features using forward passes.

Figure 2 compares these sensitivities between members and non-members across layers and heads. The continuation-loss gradients show differences in how the loss locally depends on attention. Differences also appear in the sensitivities of the cross-layer transition statistic and the prefix-perturbation statistic, connecting the diagnostic analysis to both proposed feature families. The observed patterns vary across layers and heads, motivating representations that retain these individual measurements. These observations provide supporting evidence for investigating structured attention behavior; the predictive value of the resulting features is evaluated through feature-distribution comparisons and held-out membership inference.

Individual attention statistics also show member/non-member separation. On the MIMIR GitHub domain with Pythia-1.4B, the best-performing individual Transitional feature reaches a Hellinger distance of 0.76, compared with 0.41–0.52 for the evaluated output-based scores (Appendix C). This comparison is descriptive and reports the maximum separation among the examined Transitional features; held-out attack performance is evaluated separately. Figure 3 illustrates member/non-member distributions for selected Transitional and Perturbation features.

The following subsections define the corresponding Transitional and Perturbation representations and explain how each is used to produce membership scores.

2.3 TRANSITIONAL FEATURES

Transitional features characterize differences between attention distributions in adjacent layers. For each layer pair $(\ell, \ell+1)$, we compare head h with the same-index head in the next layer. We do not assume that paired heads perform the same function or apply semantic head matching. Our primary Transitional representation uses mean row-wise KL divergence (Shlens, 2014):

$$\Delta_{\mathrm{KL}}^{\ell,h}(x) = \frac{1}{T}\sum_{i=1}^{T} D_{\mathrm{KL}}\left(A_{i,:}^{\ell,h}(x) \,\middle\|\, A_{i,:}^{\ell+1,h}(x)\right). \tag{1}$$

This statistic measures how token-level attention distributions differ between the paired heads. We retain a separate value for every adjacent-layer pair and head, rather than averaging across the network (Clark et al., 2019). Padding positions are excluded, and numerical stabilization is described in Appendix A. For diagnostic and feature-separability analyses, we also consider Pearson correlation between vectorized attention maps and normalized Frobenius distance:

$$\mathrm{Corr}^{\ell,h}(x) = \mathrm{corr}\big(\mathrm{vec}(A^{\ell,h}(x)), \mathrm{vec}(A^{\ell+1,h}(x))\big)\,, \tag{2}$$

$$\Delta_{\mathrm{F}}^{\ell,h}(x) = \frac{\big\|A^{\ell+1,h}(x) - A^{\ell,h}(x)\big\|_{\mathrm{F}}}{T^2}. \tag{3}$$

Correlation measures similarity between the overall attention patterns, while Frobenius distance measures the magnitude of their difference. The Transitional attack in our main evaluations uses only the KL features; these alternative statistics are used for diagnostic analyses.

2.4 Perturbation Features

Whereas Transitional features compare layers for a fixed input, Perturbation features compare attention before and after controlled input edits. Prior MIAs examine changes in model likelihood under neighboring texts or added prefixes (Mattern et al., 2023; Xie et al., 2024). We instead measure changes in attention distributions.

We apply three strategies independently to the original input: *prefix insertion* prepends a synthetic natural-language prefix; *word dropping* removes five randomly selected words; and *word replacement* replaces five randomly selected words with sampled tokenizer-vocabulary entries. Perturbation generation, validity constraints, and other details are specified in Appendix A. For strategy s, let x'_s denote the edited input. Original and edited inputs are tokenized with the same maximum length. We exclude padding and set

$$T_s = \min\big(T(x), T(x'_s)\big). \tag{4}$$

We then crop both attention maps to their top-left $T_s \times T_s$ regions and normalize each row. Let $\widetilde{A}_s^{\ell,h}(x)$ and $\widetilde{A}_s^{\ell,h}(x'_s)$ denote the resulting matrices. The perturbation feature is

$$\Delta_s^{\ell,h}(x) = \frac{1}{T_s}\sum_{i=1}^{T_s} D_{\mathrm{KL}}\Big(\widetilde{A}_{s,i,:}^{\ell,h}(x) \,\Big\|\, \widetilde{A}_{s,i,:}^{\ell,h}(x'_s)\Big)\,. \tag{5}$$

2.5 Feature Aggregation and Membership Scoring

We concatenate the attention statistics into feature vectors, retaining a separate coordinate for each layer–head measurement. For the Transitional variant,

$$z_{\mathrm{trans}}(x) = \mathrm{concat}_{\ell=1,\ldots,L-1;\, h=1,\ldots,H}\left[\Delta_{\mathrm{KL}}^{\ell,h}(x)\right] \in \mathbb{R}^{(L-1)H}. \tag{6}$$

For the Perturbation variant,

$$z_{\mathrm{pert}}(x) = \mathrm{concat}_{s\in\mathcal{P};\, \ell=1,\ldots,L;\, h=1,\ldots,H}\left[\Delta_s^{\ell,h}(x)\right] \in \mathbb{R}^{|\mathcal{P}|LH}, \tag{7}$$

where $\mathcal{P} = \{\mathrm{prefix}, \mathrm{drop}, \mathrm{replace}\}$ for the primary MIMIR and cross-domain evaluations.

We train a separate binary multilayer perceptron (MLP) on each representation using labeled member and non-member examples. This yields AttenMIA (Transitional) and AttenMIA (Perturbation), which are evaluated separately in our membership-inference experiments. At inference time, we extract the selected feature vector, apply the preprocessing fitted during attack training, and use the classifier's sigmoid output as the membership score $s(x)$. Protocol-specific preprocessing, architectures, and optimization details appear in Appendix A.

3 Experimental Setup

Our evaluation examines the effectiveness of attention-based membership inference within domains, its transfer to unseen domains, and its performance across target models and feature configurations.

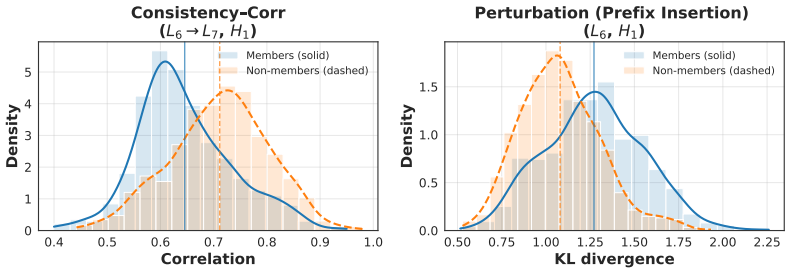


Figure 3: Distributions of individual attention features for members and non-members on MIMIR GitHub with Pythia-1.4B. Left: Transitional correlation between head H1 in layers L6 and L7. Right: mean row-wise KL divergence between original and prefix-perturbed attention maps at layer L6, head H1. The KDE illustrate member/non-member separation for these selected features.

Datasets and target models. Our primary evaluation uses MIMIR, a benchmark for studying membership inference on language-model pretraining data (Duan et al., 2024). We evaluate seven domains: Wikipedia, GitHub, Pile CC, PubMed Central, arXiv, DM Mathematics, and HackerNews. The target models are Pythia-1.4B, Pythia-2.8B, and Pythia-6.9B (Biderman et al., 2023), yielding 21 model–domain settings. These target models are evaluated without additional fine-tuning. We additionally evaluate on WikiMIA to examine performance across a broader set of target models and input lengths (Shi et al., 2023). We consider Pythia-6.9B, GPT-NeoX-20B (Black et al., 2022), OPT-6.7B (Zhang et al., 2022), and LLaMA models with 13B and 30B parameters (Touvron et al., 2023), using the benchmark's length settings of 32, 64, and 128. WikiMIA constructs membership labels using Wikipedia texts created before and after model training. Such temporal separation can introduce distributional differences unrelated to membership (Duan et al., 2024). We therefore treat it as complementary evidence rather than the primary benchmark.

Baselines. We evaluate AttenMIA (Transitional) and AttenMIA (Perturbation) as separate attacks. Our output-based baselines include perplexity (PPL), compression-normalized scoring (Zlib), and reference-model calibration (Ref) (Carlini et al., 2021; Duan et al., 2024); neighborhood comparison (Neighbor) (Mattern et al., 2023); Min-K% (Shi et al., 2023); Min-K%++ (Zhang et al., 2025); RECALL (Xie et al., 2024); and PETAL (He et al., 2025). In the in-domain evaluation, these baselines use their native scalar scores; Min-K%++ averages the lowest 20% of token-wise standardized log-probabilities. They therefore do not receive the supervised classifier training used by AttenMIA. We also include *Hidden*, a supervised classifier using hidden-state activation statistics, to compare our attention features with another internal representation. Baseline scoring and calibration details are provided in Appendix A.

Evaluation protocol and metrics. For in-domain evaluation, we use stratified five-fold cross-validation, fitting preprocessing and the membership classifier using only the training folds and evaluating on the corresponding held-out fold. For cross-domain evaluation, we hold out each MIMIR domain in turn and train using the remaining six source domains. Feature transformations are selected through an inner leave-one-source-domain-out procedure. Subsequent preprocessing, classifier training, validation, and checkpoint selection use source-domain data only. The held-out target domain is used only for final evaluation. We repeat this evaluation with five random seeds. We report ROC AUC and the TPR@1%FPR (Carlini et al., 2022). For cross-domain results, we first average each metric across the seven held-out domains within each seed, then report the mean across seeds. Additional protocol details are provided in Appendix A.

4 Results and Analysis

We first evaluate membership inference on MIMIR, then examine cross-domain transfer and complementary results on WikiMIA. We analyze the contributions of individual perturbations and model layers, and examine the association between attack scores and source overlap in generated text.

Table 1: In-domain ROC AUC on MIMIR. Sub-columns correspond to Pythia-1.4B, 2.8B, and 6.9B. The Average columns report means across the seven domains for each model size. Trans. and Pert. denote the Transitional and Perturbation variants of ATTENMIA. Bold green entries indicate the highest value, and underlined entries indicate the second-highest distinct value within each column, including ties.

Method	Wikipedia			GitHub			Pile CC			PubMed			arXiv			DM Math			HackerNews			Average		
	1.4	2.8	6.9	1.4	2.8	6.9	1.4	2.8	6.9	1.4	2.8	6.9	1.4	2.8	6.9	1.4	2.8	6.9	1.4	2.8	6.9	1.4	2.8	6.9
PPL	.64	.65	.67	.86	.86	.87	.53	.54	.55	.76	.76	.76	.75	.74	.76	.85	.86	.75	.59	.59	.60	.71	.71	.71
Ref	.58	.59	.60	.51	.50	.51	.53	.53	.54	.65	.64	.65	.63	.61	.64	.82	.78	.77	.51	.51	.52	.60	.59	.60
Zlib	.57	.59	.60	.88	.89	.90	.52	.53	.54	.75	.75	.75	.76	.76	.77	.70	.73	.64	.58	.59	.59	.68	.69	.68
Neighbor	.58	.58	.61	.76	.76	.77	.52	.55	.54	.69	.68	.69	.58	.56	.56	.76	.73	.83	.53	.52	.52	.63	.64	.65
Min-K%	.61	.62	.65	.84	.85	.86	.52	.53	.54	.74	.74	.74	.67	.67	.69	.58	.61	.53	.55	.56	.57	.64	.65	.65
Min-K%++	.59	.60	.65	.62	.64	.66	.50	.51	.53	.38	.36	.38	.48	.44	.46	.42	.44	.39	.55	.55	.57	.50	.51	.52
PETAL	.61	.61	.61	.85	.85	.87	.53	.54	.54	.72	.72	.73	.61	.60	.61	.87	.88	.87	.56	.58	.58	.68	.68	.69
RECALL	.58	.59	.59	.81	.81	.82	.52	.53	.54	.67	.68	.61	.76	.75	.70	.54	.58	.67	.59	.58	.59	.64	.65	.65
Hidden	.61	.53	.50	.91	.91	.90	.53	.50	.50	.69	.58	.50	.74	.75	.74	.90	.81	.54	.53	.51	.50	.70	.66	.60
AttenMIA (Trans.)	.63	.64	.62	.89	.91	.87	.59	.57	.57	.86	.85	.84	.81	.79	.78	.94	.94	**.95**	.60	.59	.58	.76	.76	.75
AttenMIA (Pert.)	**.70**	**.72**	**.72**	**.99**	**.99**	**.99**	**.66**	**.68**	**.70**	**.93**	**.92**	**.95**	**.92**	**.90**	**.94**	**.96**	**.95**	.94	**.70**	**.67**	**.68**	**.84**	**.83**	**.85**

4.1 MAIN RESULTS

MIMIR. Table 1 reports in-domain ROC AUC across seven domains and three Pythia model sizes. ATTENMIA (Perturbation) achieves the highest AUC in 20 of the 21 model–domain settings. Its domain-averaged AUC ranges from 0.83 to 0.85, compared with 0.71 for PPL, the strongest evaluated output-based baseline by average AUC. The supervised Hidden control achieves average AUCs of 0.70, 0.66, and 0.60 at the three respective model sizes, below those of both ATTENMIA variants. At 1% FPR, at least one ATTENMIA variant achieves the best or tied-best TPR in all 21 settings, with Perturbation leading or tying in 19 (Table 4). Its TPR reaches 91.8–95.0% on GitHub. However, the gains vary across domains: on Pile CC with Pythia-1.4B, Perturbation and Zlib both achieve 5.1% TPR.

Generalization to Unseen Domains. We next evaluate whether the learned membership classifier transfers to domains excluded from attack development. Table 2 reports target-blind leave-one-domain-out results, averaged across the seven held-out MIMIR domains. ATTENMIA (Transitional) achieves the highest domain-averaged ROC AUC and TPR@1%FPR among the evaluated methods at all three Pythia sizes. Averaged across domains and model sizes, it reaches 0.73 AUC and 26.7% TPR@1%FPR, followed by ATTENMIA (Perturbation) at 0.69 and 17.8%. Among the evaluated output-based baselines, PPL achieves the highest average AUC (0.64), while Zlib achieves the highest average TPR@1%FPR (15.5%). The supervised Hidden control reaches 0.54 AUC and 4.3% TPR@1%FPR. These results demonstrate transfer to held-out domains and show that the evaluated hidden-state representation does not match the transfer performance of the proposed attention features. The variants also exhibit different strengths: Perturbation performs better in-domain, whereas Transitional performs better under cross-domain evaluation.

WikiMIA. Table 3 reports results across five target models and three input-length settings. ATTENMIA (Transitional) achieves higher ROC AUC than the evaluated output-based baselines and the Hidden control in 14 of 15 model–length settings, and in the other setting, ATTENMIA (Perturbation) is the best. Its model-averaged AUC is approximately 0.99, 0.99, and 0.97 at lengths 32, 64, and 128, respectively. The corresponding average TPR@1%FPR values are 86.8%, 78.4%, and 54.5%. RECALL nevertheless achieves higher TPR@1%FPR than Transitional in three settings. These results extend the evaluation across additional models and input lengths, and we interpret this evaluation as complementary evidence.

4.2 ABLATIONS AND FEATURE ANALYSIS

Contribution of perturbation strategies. We evaluate all three strategies separately on MIMIR with Pythia-1.4B, and compare them with the case that concatenates features from all three strategies (Table 6). The individual strategies achieve similar domain-averaged AUCs of approximately 0.77.

Table 2: Cross-domain evaluation on MIMIR. Each model's columns report equally weighted averages across the seven held-out target domains. The final Average columns additionally average across the three Pythia model sizes. TPR@1%FPR is expressed as a percentage.

Method	Pythia-1.4B		Pythia-2.8B		Pythia-6.9B		Average	
	AUC	TPR@1%FPR	AUC	TPR@1%FPR	AUC	TPR@1%FPR	AUC	TPR@1%FPR
PPL	0.64	14.1	0.63	13.6	0.65	13.1	0.64	13.6
Zlib	0.62	15.0	0.63	15.7	0.65	15.8	0.63	15.5
RECALL	0.62	13.2	0.62	14.3	0.63	15.0	0.62	14.2
Min-K%	0.59	10.1	0.60	10.1	0.60	10.9	0.60	10.4
Min-K%++	0.54	3.6	0.55	3.9	0.58	4.2	0.56	3.9
Neighbor	0.56	5.3	0.54	3.7	0.56	4.6	0.55	4.5
Ref	0.52	2.3	0.51	1.6	0.50	2.1	0.51	2.0
Hidden	0.55	3.7	0.55	5.1	0.54	4.2	0.54	4.3
AttenMIA (Perturbation)	0.68	18.3	0.68	16.7	0.70	18.4	0.69	17.8
AttenMIA (Transitional)	**0.73**	**26.6**	**0.73**	**24.3**	**0.73**	**29.2**	**0.73**	**26.7**

Table 3: In-domain results on WikiMIA. Sub-columns indicate the sequence lengths. Each entry reports ROC AUC, followed by TPR@1%FPR in parentheses. The Average columns report means across the five target models. Transitional and Perturbation denote the two ATTENMIA variants. Bold green entries indicate the highest value, and underlined entries indicate the second-highest distinct value, separately for each metric within each column, including ties.

Method	Pythia-6.9B			LLaMA-13B			NeoX-20B			LLaMA-30B			OPT-6.7B			Average		
	32	64	128	32	64	128	32	64	128	32	64	128	32	64	128	32	64	128
PPL	.72 (23.9)	.69 (13.2)	.71 (8.8)	.72 (28.6)	.66 (11.2)	.64 (8.8)	.77 (28.2)	.75 (19.6)	.79 (11.9)	.77 (35.9)	.73 (16.8)	.74 (14.4)	.71 (17.2)	.67 (11.2)	.68 (13.3)	.74 (26.8)	.70 (14.4)	.71 (11.4)
Ref	.66 (6.2)	.65 (8.6)	.64 (8.4)	.69 (3.5)	.62 (4.9)	.66 (1.8)	.79 (21.4)	.81 (22.9)	.81 (16.1)	.80 (7.8)	.81 (16.8)	.88 (28.4)	.63 (0.8)	.65 (6.3)	.65 (6.0)	.71 (7.9)	.71 (11.9)	.73 (12.1)
Zlib	.74 (17.3)	.70 (16.2)	.74 (10.5)	.73 (17.1)	.67 (12.3)	.66 (9.5)	.79 (28.5)	.76 (22.2)	.81 (13.3)	.78 (23.0)	.74 (17.8)	.76 (14.7)	.73 (14.5)	.69 (11.9)	.70 (14.3)	.75 (20.1)	.71 (16.1)	.73 (12.5)
Neighbor	.68 (10.2)	.67 (9.2)	.63 (3.5)	.66 (6.0)	.65 (5.2)	.60 (3.9)	.71 (16.9)	.73 (13.3)	.70 (8.4)	.67 (3.7)	.68 (9.0)	.69 (7.7)	.57 (2.2)	.58 (4.0)	.58 (4.2)	.66 (7.8)	.66 (8.1)	.64 (5.5)
Min-K%	.76 (21.4)	.73 (15.6)	.76 (9.8)	.73 (26.7)	.68 (12.9)	.67 (8.4)	.83 (29.9)	.79 (21.8)	.83 (15.8)	.78 (33.8)	.75 (19.1)	.78 (16.8)	.80 (26.7)	.75 (19.8)	.75 (12.6)	.78 (27.7)	.74 (17.8)	.76 (12.7)
Min-K%++	.76 (29.3)	.75 (18.8)	.77 (16.1)	.86 (48.9)	.80 (20.5)	.77 (9.5)	.83 (31.2)	.80 (22.4)	.84 (28.6)	.86 (50.1)	.82 (24.7)	.82 (19.6)	.76 (25.3)	.75 (18.9)	.77 (16.4)	.81 (37.0)	.78 (21.1)	.79 (18.0)
PETAL	.71 (26.7)	.69 (17.9)	.65 (11.5)	.73 (23.1)	.67 (12.0)	.62 (11.2)	.74 (26.8)	.73 (17.8)	.73 (13.6)	.75 (28.4)	.74 (27.0)	.71 (25.8)	.65 (8.1)	.62 (8.0)	.61 (8.4)	.72 (22.6)	.69 (16.5)	.66 (14.1)
RECALL	.97 (44.4)	.97 (58.9)	.92 (**44.6**)	.68 (2.8)	.71 (3.4)	.76 (10.5)	.98 (53.4)	.98 (72.2)	.93 (**50.1**)	.98 (73.2)	.98 (**92.0**)	.74 (6.0)	.83 (16.4)	.69 (4.7)	.61 (0.7)	.89 (38.0)	.87 (46.2)	.79 (22.4)
Hidden	.80 (0.0)	.70 (9.7)	.73 (10.8)	.81 (0.0)	.75 (5.0)	.78 (14.7)	.81 (4.4)	.70 (9.9)	.71 (5.9)	.83 (0.0)	.78 (5.6)	.77 (15.7)	.78 (3.1)	.70 (7.7)	.71 (16.4)	.81 (1.5)	.73 (7.6)	.74 (12.7)
Transitional	**.99** (**81.5**)	**.98** (**68.6**)	.93 (37.5)	**.99** (86.4)	**.99** (**84.0**)	**.99** (65.9)	**.99** (**84.8**)	**.99** (78.4)	**.96** (45.2)	**.99** (**93.7**)	**.99** (87.5)	**.99** (67.4)	**.99** (**87.7**)	**.99** (**73.6**)	**.98** (56.3)	**.99** (**86.8**)	**.99** (**78.4**)	**.97** (**54.5**)
Perturbation	**.99** (81.3)	.88 (31.5)	**.95** (42.5)	**.99** (**91.4**)	**.99** (79.4)	**.99** (**68.4**)	**.99** (55.5)	**.99** (**86.8**)	.85 (15.0)	**.99** (76.0)	**.99** (70.4)	**.99** (**70.5**)	**.99** (63.3)	**.99** (72.2)	**.98** (**66.9**)	**.99** (73.5)	.97 (68.1)	.95 (52.7)

Combining them increases average AUC to 0.84 and TPR@1%FPR to 45.5%, compared with 32.9–35.8% for the individual strategies. The combined representation therefore performs best on average, although it does not achieve the highest TPR in every domain. These results support using the three perturbation strategies together in our primary evaluation.

Contribution of different layers. We examine how the choice of layers affects membership inference using Perturbation features from Pythia-1.4B on the MIMIR PubMed Central domain. We compare representations containing features from individual layers, groups of eight consecutive layers (lower, middle, and upper), and all 24 layers. Figure 4 shows that individual layers provide varying levels of predictive information, while combining features across layers improves performance. The representation using all layers achieves the highest ROC AUC among the evaluated configurations. These results support why we extract the MIA features from all heads/layers.

4.3 ASSOCIATION WITH SOURCE-TEXT OVERLAP

As a complementary analysis, we examine whether AttenMIA scores track source-text overlap in GPT-2 generations. We prompt GPT-2 with 5–10 token prefixes sampled from Common Crawl and generate 256-token continuations. From $14{,}000$ continuations, we select the $1{,}000$ highest and $1{,}000$ lowest ROUGE-L examples and evaluate all methods on them. AttenMIA achieves a Pearson correlation of $r = 0.48$ with ROUGE-L, compared with $r = 0.32$ for Zlib/XL, the strongest evaluated baseline by absolute correlation. Appendix E provides the setup, complete results, and examples.

5 RELATED WORK

Membership inference. Membership inference attacks determine whether a candidate example was used for training (Shokri et al., 2017; Yeom et al., 2018). For LLMs, existing attacks use se-

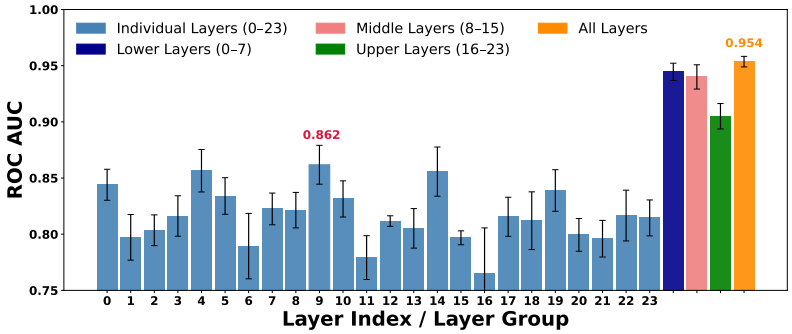


Figure 4: Effect of layer selection on MIA using Perturbation features from Pythia-1.4B on PubMed Central. We compare individual layers, lower (0–7), middle (8–15), and upper (16–23) layer groups, and all 24 layers. All layers achieves the highest ROC AUC among the evaluated configurations.

quence likelihood (Yeom et al., 2018), compression, or reference-model calibration (Carlini et al., 2021), low-probability token statistics (Shi et al., 2023; Zhang et al., 2025), and likelihood changes under neighboring texts or added prefixes (Mattern et al., 2023; Xie et al., 2024). PETAL estimates token likelihoods from semantic similarities between generated and candidate tokens, enabling MIA without access to target-model output probabilities (He et al., 2025). Our approach instead characterizes cross-layer attention transitions and attention responses to input edits.

Internal signals and attention-based security. White-box MIAs already exploit model internals, including gradients and learned features (Nasr et al., 2019; Leino & Fredrikson, 2020; Sablayrolles et al., 2019; Li et al., 2024). Our contribution is the design and evaluation of structured attention representations for membership inference. The supervised hidden-state control compares these representations with an alternative source of internal features under the evaluated protocols. Attention has also been studied in other security settings. Attention Eclipse manipulates attention relationships within prompts to strengthen jailbreak attacks (Zaree et al., 2025). Pandya et al. (Pandya et al., 2025) develop attention-based attacks against fine-tuned prompt-injection defenses, while Pu et al. (Pu et al., 2025) use attention distributions to guide jailbreak attacks and detect them. These studies examine adversarial behavior and defenses; ATTENMIA investigates whether cross-layer attention transitions and perturbation responses reveal training membership.

Memorization and extraction. Language models can reproduce training text, enabling attacks that extract memorized sequences (Carlini et al., 2021). Prior work finds that verbatim memorization increases with model capacity, repeated training examples, and prompt context length (Carlini et al., 2023), while training-data deduplication reduces susceptibility to extraction attacks (Kandpal et al., 2022). Counterfactual memorization instead measures how a model's predictions change when a document is excluded from training (Zhang et al., 2023). Membership inference asks whether a candidate was used in training, without requiring its verbatim reproduction. Our analysis examines membership-associated attention sensitivities, while our complementary generation experiment measures score–source overlap correlation.

6 CONCLUSION

We introduced ATTENMIA, a membership inference attack that leverages the membership signal in the attention space. We use two complementary views of self-attention: cross-layer transitions and responses to controlled input perturbations. A supervised hidden-state control performs substantially worse in cross-domain evaluation, supporting the effectiveness of our attention features relative to the evaluated internal representation. Fine-tuning experiments further show membership-associated differences in the sensitivities of prediction loss and attention features to changes in attention maps. ATTENMIA scores also correlate more strongly with source-text overlap than the evaluated baselines on generated continuations. Together, these results identify self-attention as a useful privacy signal beyond model outputs, motivating both stronger internal-state auditing and defenses against leakage through model representations.

Ethics Statement

This work studies a privacy vulnerability, so we take its dual-use nature seriously. Our aim is diagnostic: we characterize a leakage channel inside transformers so that model builders can measure and defend against it, not so that it can be exploited. Our experiments use publicly released, open-weight models, public benchmarks, and synthetic examples. We do not target proprietary models, seek to recover personal data, or conduct studies involving human subjects. We view membership auditing as a tool that, like the attacks it mirrors, is most useful to the people trying to protect training data. We encourage practitioners to treat access to attention weights as part of the privacy surface of a released model.

Reproducibility Statement

We describe our MIA feature families, computed statistics, and membership classifiers in Section 2 (Equations 1–7). Section 3 describes the target models, datasets, evaluation metrics, in-domain cross-validation, and target-blind cross-domain protocol. Additional implementation details, tables, figures, and analyses are provided in the Appendix. We release the code needed to reproduce AttenMIA, including attention-feature extraction, membership-classifier training and evaluation, dataset preprocessing, baseline evaluation, and the MIMIR and WikiMIA pipelines. An anonymized repository is available for review: https://anonymous.4open.science/r/AttenMIA... Click on it..

Experiments are implemented in PyTorch and Hugging Face Transformers and were run primarily on NVIDIA A100 40GB GPUs.

AI Use Statement

Generative AI tools were used to assist with language editing and manuscript organization. All research decisions, experimental protocols, implementations, analyses, numerical results, figures, and claims were reviewed and verified by the authors. The authors take full responsibility for the correctness and content of the paper.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pp. 2397–2430. PMLR, 2023.

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, et al. Gpt-neox-20b: An open-source autoregressive language model. *arXiv preprint arXiv:2204.06745*, 2022.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX security symposium (USENIX security 19)*, pp. 267–284, 2019.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

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

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models, 2023. *URL https://arxiv.org/abs/2202.07646*, 2202, 2023.

Marvin M Chun and Yuhong Jiang. Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive psychology*, 36(1):28–71, 1998.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does bert look at? an analysis of bert's attention. In *Proceedings of the 2019 ACL workshop BlackboxNLP: analyzing and interpreting neural networks for NLP*, pp. 276–286, 2019.

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

Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

Yu He, Boheng Li, Liu Liu, Zhongjie Ba, Wei Dong, Yiming Li, Zhan Qin, Kui Ren, and Chun Chen. Towards label-only membership inference attack against pre-trained large language models. In *USENIX Security*, 2025.

Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S Yu, and Xuyun Zhang. Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)*, 54 (11s):1–37, 2022.

Nikhil Kandpal, Eric Wallace, and Colin Raffel. Deduplicating training data mitigates privacy risks in language models. In *International Conference on Machine Learning*, pp. 10697–10707. PMLR, 2022.

Klas Leino and Matt Fredrikson. Stolen memories: Leveraging model memorization for calibrated {White-Box} membership inference. In *29th USENIX security symposium (USENIX Security 20)*, pp. 1605–1622, 2020.

Chenxi Li, Abhinav Kumar, Zhen Guo, Jie Hou, and Reza Tourani. Unveiling the unseen: Exploring whitebox membership inference through the lens of explainability. *arXiv preprint arXiv:2407.01306*, 2024.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

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

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

Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE symposium on security and privacy (SP)*, pp. 739–753. IEEE, 2019.

Mikhail S Nikulin et al. Hellinger distance. *Encyclopedia of mathematics*, 78, 2001.

Nishit V Pandya, Andrey Labunets, Sicun Gao, and Earlence Fernandes. May i have your attention? breaking fine-tuning based prompt injection defenses using architecture-aware attacks. *arXiv preprint arXiv:2507.07417*, 2025.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

Rui Pu, Chaozhuo Li, Rui Ha, Zejian Chen, Litian Zhang, Zheng Liu, Lirong Qiu, and Zaisheng Ye. Feint and attack: Jailbreaking and protecting llms via attention distribution modeling. In *IJCAI*, pp. 493–501, 2025.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, et al. Hopfield networks is all you need. *arXiv preprint arXiv:2008.02217*, 2020.

Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Yann Ollivier, and Hervé Jégou. White-box vs black-box: Bayes optimal strategies for membership inference. In *International Conference on Machine Learning*, pp. 5558–5567. PMLR, 2019.

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

Jonathon Shlens. Notes on kullback-leibler divergence and likelihood. *arXiv preprint arXiv:1404.2000*, 2014.

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

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

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

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

Pedram Zaree, Md Abdullah Al Mamun, Quazi Mishkatul Alam, Yue Dong, Ihsen Alouani, and Nael Abu-Ghazaleh. Attention eclipse: Manipulating attention to bypass llm safety-alignment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 16659–16679, 2025.

Chiyuan Zhang, Daphne Ippolito, Katherine Lee, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. Counterfactual memorization in neural language models. *Advances in Neural Information Processing Systems*, 36:39321–39362, 2023.

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

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

A Reproducibility and Implementation Details

This appendix provides implementation details for attention feature extraction, classifier training, and baseline evaluation.

A.1 Attention Feature Extraction

Transitional features. For each input, we compare the attention map of head h in layer ℓ with the same-index head in layer $\ell+1$, without learned or semantic head matching. The Transitional attacks in the main MIMIR, cross-domain, and WikiMIA evaluations use mean row-wise KL divergence (Equation 1). Pearson correlation and normalized Frobenius distance (Equations 2 and 3) are used for diagnostic and feature-separability analyses.

Padding positions are excluded using the input attention mask. Zeros introduced by causal masking remain in the attention matrices. For KL computation, each valid row is normalized and its entries are clamped to $[10^{-12}, 1]$. The divergence is computed in the direction $D_{\mathrm{KL}}(A^{\ell,h}_{i,:} \| A^{\ell+1,h}_{i,:})$ and averaged over valid query rows.

We retain one KL statistic per adjacent-layer pair and head, giving $(L-1)H$ features per input. This produces 368 features for Pythia-1.4B and 992 features for each of Pythia-2.8B and Pythia-6.9B.

Perturbation features. For the primary MIMIR and cross-domain evaluations, we independently apply three perturbations to each original input: prefix insertion, word dropping, and word replacement. Prefix insertion prepends one of ten fixed synthetic strings before tokenization. Word dropping removes five randomly selected whitespace-delimited words. Word replacement replaces five such words with sampled tokenizer-vocabulary entries. Inputs containing at most five whitespace-delimited words are left unchanged for dropping and replacement.

Replacement candidates must have token IDs of at least 100, decode to nonempty text, and not begin with <. GPT-style and SentencePiece-style word-boundary markers are stripped from sampled replacements. Perturbations are deterministic given seed 42 and the example index, and are applied independently rather than cumulatively.

Original and edited inputs are tokenized using the same maximum length. After excluding padding, we crop both attention maps to their top-left $T_s \times T_s$ regions, where $T_s = \min(T(x), T(x'_s))$. Each row is normalized, entries are clamped to $[10^{-12}, 1]$, and the mean row-wise KL divergence is computed from the original to the edited map (Equation 5). Non-finite divergence values are mapped to zero.

We concatenate the three perturbation blocks across all layers and heads, producing $3LH$ features per input: 1,152 for Pythia-1.4B and 3,072 for each of Pythia-2.8B and Pythia-6.9B.

A.2 Classifier and Preprocessing Details

In-domain MIMIR. We evaluate each AttenMIA variant separately using five-fold cross-validation. Non-finite feature values are replaced with finite constants before classification. Within each fold, a `StandardScaler` (Pedregosa et al., 2011) is fitted on the attack-training examples and applied to both the training and held-out examples.

The classifier is an MLP with one hidden layer containing 128 ReLU units, dropout with probability 0.1, and a single output logit. It is trained for 50 epochs using AdamW (Loshchilov, 2017) and binary cross-entropy with logits. At evaluation time, the sigmoid of the output logit provides the membership score. Held-out examples are excluded from preprocessing fitting and classifier training.

WikiMIA. The WikiMIA Transitional classifier uses two hidden layers with 512 and 128 ReLU units, dropout with probability 0.3, and a single output logit. Features are used without standardization. We evaluate the classifier using five-fold cross-validation, training on four folds and evaluating on the remaining fold.

Target-blind cross-domain evaluation. For each held-out MIMIR domain, the remaining six domains supply all attack-development data. We first select a feature transformation from {`raw`, `group_robust_z`, `group_rank`, `group_robust_rank`} through an inner leave-one-source-domain-out evaluation. Transformations are ranked by mean TPR@1%FPR across the inner held-out domains, with mean ROC AUC breaking ties.

After transformation selection, the pooled source examples are split into 80% training and 20% validation data. A `StandardScaler` is fitted on the source-training partition and applied to validation and target examples. The classifier has two hidden layers with 128 and 64 ReLU units, dropout with probability 0.1, and a single output logit.

We train for up to 50 epochs using AdamW with learning rate 10^{-3}, weight decay 10^{-4}, and batch size 64. The objective combines binary cross-entropy with logits and a source-only pairwise tail-ranking term with weight 0.5 and tail fraction 0.2. Checkpoint selection uses source-validation performance, prioritizing partial AUC at 5% FPR, followed by TPR@5%FPR and overall ROC AUC as tie-breakers. The target domain is excluded from transformation selection, preprocessing fitting, classifier training, validation, and checkpoint selection. We repeat the evaluation with seeds 42–46.

A.3 Baseline Scoring

For the in-domain evaluations, output-based baselines use their native scalar scores without an additional membership classifier. For Min-K%++ (Zhang et al., 2025), we compute

$$z_t = \frac{\log p(x_t \mid x_{<t}) - \mu_t}{\sigma_t}, \tag{8}$$

where μ_t and σ_t are the mean and standard deviation of log probability under the model's next-token distribution at position t. The membership score is the mean of the lowest 20% of these standardized token scores. No learned calibration layer is applied to the in-domain Min-K%++ scores.

For target-blind cross-domain evaluation, scalar baselines are additionally evaluated with supervised calibration using source-domain data only. The held-out target domain is excluded from calibration fitting and model selection. The Hidden baseline is a separate supervised using hidden-state activation statistics.

A.4 AttenMIA Algorithm

Algorithm 1 summarizes feature extraction, attack training, and inference. Both representations are used to train separate attacks. Preprocessing and classifier selection follow the evaluation-specific splits described above. For the primary MIMIR and cross-domain evaluations, $\mathcal{P} = \{\text{prefix}, \text{drop}, \text{replace}\}$.

A.5 Controlled GPT-2 Diagnostics

Experimental setup. We fine-tune pretrained GPT-2 on synthetic canaries and compare its attention behavior on training members and non-members. For each seed, 256 synthetic canaries are shuffled and divided into 128 members and 128 non-members. Both groups follow the same generation procedure: each canary contains a subject-specific prompt followed by a synthetic secret. Non-members are excluded from fine-tuning.

The training corpus contains 32 member canaries at each duplication level of 4, 8, 16, and 24 copies, together with 128 synthetic distractor sentences. These duplication levels specify corpus multiplicities, rather than the number of presentations during optimization. We fine-tune for 300 AdamW steps with learning rate 5×10^{-4}, weight decay 0.01, batch size 4, maximum sequence length 48, and FP32 computation.

Figure 2 compares members with non-members. Membership here refers to inclusion in the fine-tuning corpus.

Continuation-loss sensitivity. Let $x = (x_1, \ldots, x_T)$ be a tokenized canary, excluding padding, and let $\mathcal{S}(x)$ contain the positions of its continuation tokens. We compute the mean next-token cross-entropy over these positions:

$$\mathcal{L}_{\text{suffix}}(x) = -\frac{1}{|\mathcal{S}(x)|} \sum_{t \in \mathcal{S}(x)} \log p_M(x_t \mid x_{<t}), \tag{9}$$

Algorithm 1 ATTENMIA: training and inference

Require: Target language model M; labeled attack-development set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where $y_i = 1$ denotes a member and $y_i = 0$ a non-member; perturbation set $\mathcal{P}$; evaluation protocol
Ensure: Two trained classifiers, C_{trans} and C_{pert}, with their fitted preprocessing functions, P_{trans} and P_{pert}
Notation: $A^{\ell,h}(x)$ is the attention map from layer ℓ, head h, of M on input x. $z_{\text{trans}}(x)$ and $z_{\text{pert}}(x)$ are the Transitional and Perturbation feature vectors. Each classifier C_v returns a single logit; σ denotes the sigmoid function.
Feature extraction
for each $(x_i, y_i) \in \mathcal{D}$ **do**
  Run $M(x_i)$ and collect $\{A^{\ell,h}(x_i)\}_{\ell,h}$.
  Construct $z_{\text{trans}}(x_i)$ using Equation 6.
  for each perturbation strategy $s \in \mathcal{P}$ **do**
    Apply s to the original input: $x'_{i,s} \leftarrow s(x_i)$.
    Run $M(x'_{i,s})$ and collect its attention maps.
    Compute attention-change statistics $\{\Delta_s^{\ell,h}(x_i)\}_{\ell,h}$ using Equation 5.
  end for
  Construct $z_{\text{pert}}(x_i)$ using Equation 7.
end for
Attack training
for each variant $v \in \{\text{trans}, \text{pert}\}$ **do**
  Select and fit preprocessing P_v and train classifier C_v on $\{(z_v(x_i), y_i)\}_{i=1}^{N}$, following the protocol's training and validation splits.
  Exclude held-out evaluation examples from all fitting and selection.
  Save (P_v, C_v).
end for
Inference
Require: Candidate x and selected variant $v \in \{\text{trans}, \text{pert}\}$
Extract the corresponding feature vector $z_v(x)$.
Apply the saved preprocessing: $u \leftarrow P_v(z_v(x))$.
return Membership score $s_v(x) = \sigma(C_v(u))$.

where p_M is the target model's next-token distribution. Prompt tokens provide context but do not contribute directly to the loss.

For each layer ℓ and head h, we differentiate this loss with respect to the post-softmax attention probabilities used in the model's forward computation:

$$G_{\text{loss}}^{\ell,h}(x) = \frac{\partial \mathcal{L}_{\text{suffix}}(x)}{\partial A^{\ell,h}(x)}. \tag{10}$$

The derivative is computed through the model's computation graph.

Let $\mathcal{V}(x) = \{(i, j) : 1 \le j \le i \le T\}$ denote the valid causal attention entries. We summarize the gradient for each example by its root-mean-square magnitude:

$$g_{\text{loss}}^{\ell,h}(x) = \sqrt{\frac{1}{|\mathcal{V}(x)|} \sum_{(i,j)\in\mathcal{V}(x)} \left(G_{\text{loss},ij}^{\ell,h}(x)\right)^2}. \tag{11}$$

Although the loss includes only continuation tokens, this summary covers the full valid causal attention map. It measures gradient magnitude rather than signed gradient direction.

Transitional-feature sensitivity. For each adjacent-layer pair $(\ell, \ell + 1)$ and head h, we use the same mean row-wise KL statistic as the Transitional attack:

$$F_{\text{trans}}^{\ell,h}(x) = \Delta_{\text{KL}}^{\ell,h}(x), \tag{12}$$

as defined in Equation 1. We differentiate this scalar with respect to both attention maps involved:

$$G_{\text{earlier}}^{\ell,h}(x) = \frac{\partial F_{\text{trans}}^{\ell,h}(x)}{\partial A^{\ell,h}(x)}, \qquad G_{\text{later}}^{\ell,h}(x) = \frac{\partial F_{\text{trans}}^{\ell,h}(x)}{\partial A^{\ell+1,h}(x)}. \tag{13}$$

Both attention maps remain connected to the model's computation graph. Consequently, the earlier-map gradient includes the dependence of the later attention map on earlier computation; it is not a direct derivative with the two maps treated as independent inputs.

We combine the valid causal gradient entries from both maps before computing their RMS magnitude:

$$g^{\ell,h}_{\text{trans}}(x) = \sqrt{\frac{\sum_{(i,j)\in\mathcal{V}(x)} \left[\left(G^{\ell,h}_{\text{earlier},ij}(x)\right)^2 + \left(G^{\ell,h}_{\text{later},ij}(x)\right)^2\right]}{2|\mathcal{V}(x)|}}. \tag{14}$$

Here, $\mathcal{V}(x)$ is the valid causal region defined above. Each value summarizes one adjacent-layer pair and head.

Prefix-perturbation sensitivity. Let x' be the input obtained by prepending a synthetic prefix to x, and let $T_s = \min(T(x), T(x'))$. For each layer ℓ and head h, we detach the original and perturbed attention maps from the model's computation graph and treat them as independent differentiable inputs, denoted by B and B', respectively.

We compute the prefix-perturbation statistic $F^{\ell,h}_{\text{prefix}}(B, B')$ using the cropping, row normalization, and KL computation specified in Equation 5. The derivatives are

$$G^{\ell,h}_{\text{original}} = \frac{\partial F^{\ell,h}_{\text{prefix}}}{\partial B}, \qquad G^{\ell,h}_{\text{prefix}} = \frac{\partial F^{\ell,h}_{\text{prefix}}}{\partial B'}. \tag{15}$$

These are direct derivatives of the feature computation; they do not propagate through GPT-2.

Let $\mathcal{V}_s = \{(i,j) : 1 \le j \le i \le T_s\}$ denote the valid causal entries in the common cropped region. We combine entries from both gradients and compute

$$g^{\ell,h}_{\text{pert}}(x) = \sqrt{\frac{\sum_{(i,j)\in\mathcal{V}_s} \left[\left(G^{\ell,h}_{\text{original},ij}\right)^2 + \left(G^{\ell,h}_{\text{prefix},ij}\right)^2\right]}{2|\mathcal{V}_s|}}. \tag{16}$$

This diagnostic uses prefix insertion only, rather than the three-edit representation used by the primary Perturbation attack.

Heatmap standardization. For each heatmap cell, let $u(x)$ denote the per-example statistic used for plotting. Let μ_{M} and μ_{N} be its means over members and non-members, and let s^2_{M} and s^2_{N} be the corresponding within-group variances. We compute the standardized mean difference

$$d = \frac{\mu_{\text{M}} - \mu_{\text{N}}}{\sqrt{\left(s^2_{\text{M}} + s^2_{\text{N}}\right)/2}}. \tag{17}$$

Positive values indicate larger mean plotting statistics for members, while negative values indicate larger means for non-members. The sign describes the group difference, not the direction of the underlying gradient.

B Additional MIMIR Results

Table 4 reports the complete TPR@1%FPR results corresponding to the in-domain ROC AUC evaluation. At least one AttenMIA variant achieves the best or tied-best TPR in all 21 model–domain settings; the Perturbation variant leads or ties in 19. All TPR values are expressed as percentages.

C Additional Motivation and Feature Separability

We provide additional descriptive evidence of differences in attention between members and non-members. These analyses complement the attention-space gradient observations in Section 2.2 by examining attention concentration, responses to token dropping, and individual feature distributions.

Table 4: In-domain TPR at 1% FPR on MIMIR. Within each domain, sub-columns correspond to Pythia-1.4B, 2.8B, and 6.9B. The Average columns report equally weighted means across the seven domains for each model size. Trans. and Pert. denote the Transitional and Perturbation variants of ATTENMIA. Bold green entries indicate the highest value, and underlined entries indicate the second-highest distinct value within each column, including ties. Higher is better.

Method	Wikipedia			GitHub			Pile CC			PubMed Central			arXiv			DM Math			HackerNews			Average		
	1.4	2.8	6.9	1.4	2.8	6.9	1.4	2.8	6.9	1.4	2.8	6.9	1.4	2.8	6.9	1.4	2.8	6.9	1.4	2.8	6.9	1.4	2.8	6.9
PPL	9.8	10.0	10.7	23.6	25.8	28.4	4.8	6.0	6.6	12.0	9.2	11.4	19.0	18.2	19.8	47.1	45.7	34.9	3.1	3.4	3.7	17.1	16.9	16.5
Ref	4.2	4.2	4.3	3.4	3.0	4.1	3.2	4.0	4.7	2.0	1.2	1.6	6.0	6.2	5.8	49.4	50.6	47.2	2.6	1.4	2.0	10.2	10.1	10.0
Zlib	7.6	8.2	9.9	37.7	41.1	41.4	**5.1**	6.1	7.0	11.4	11.0	10.0	15.4	18.0	17.4	24.7	20.1	15.6	4.2	4.0	4.5	15.2	15.5	15.1
Neighbor	3.8	5.2	4.9	17.1	14.5	27.2	2.4	3.2	1.8	8.4	7.2	5.2	14.8	12.4	12.0	19.1	24.8	27.8	1.4	0.2	1.6	9.6	9.6	11.5
Min-K%	8.3	8.5	10.6	23.6	26.2	30.3	3.8	4.2	5.3	9.0	9.4	10.2	8.2	11.0	11.8	15.7	14.5	15.7	3.9	3.4	4.2	10.4	11.0	12.6
Min-K%++	6.0	6.8	8.0	7.9	5.6	8.6	2.4	2.3	4.4	1.6	1.0	2.0	2.4	2.2	2.0	2.2	2.2	1.1	2.2	3.4	3.6	3.5	3.4	4.2
PETAL	9.1	8.4	10.0	47.0	48.5	49.0	4.0	4.5	4.8	5.5	4.0	4.5	10.2	9.0	11.0	59.5	57.2	55.0	2.9	4.4	6.8	19.7	19.3	20.1
RECALL	2.4	3.8	5.4	14.9	5.9	17.2	3.4	3.2	5.7	0.4	0.6	1.0	11.6	15.2	12.0	4.5	14.7	21.4	2.6	3.3	2.9	5.7	6.7	9.4
Hidden	7.9	1.7	0.0	68.9	61.8	67.0	2.6	0.3	0.3	2.2	1.0	0.2	15.6	19.2	18.8	65.1	55.4	11.8	3.3	1.2	0.0	23.7	20.1	14.0
AttenMIA (Trans.)	14.7	7.5	7.2	59.4	45.8	41.7	2.6	4.7	2.6	29.5	23.6	13.0	29.2	15.4	16.8	**85.6**	84.4	**87.8**	5.0	3.7	5.0	32.3	26.4	24.9
AttenMIA (Pert.)	**19.0**	**16.0**	**16.9**	**92.3**	**91.8**	**95.0**	**5.1**	**9.2**	**7.2**	**47.6**	**53.1**	**47.1**	**60.0**	**41.2**	**55.2**	84.1	**86.4**	86.3	**9.3**	**10.3**	**9.6**	**45.5**	**44.0**	**45.3**

Attention concentration. On WikiMIA, we measure attention concentration using the Kullback–Leibler (KL) divergence from each attention row to a uniform distribution:

$$\kappa^{\ell,h}(x) = \frac{1}{T}\sum_{i=1}^{T} \mathrm{KL}\Big(A^{\ell,h}_{i,:}(x) \,\|\, U_T\Big), \tag{18}$$

where $A^{\ell,h}_{i,:}(x)$ is the attention distribution of token i in head h of layer ℓ, and U_T is uniform over the T token positions. We average these values across heads to obtain a layer-wise statistic. Larger values indicate more concentrated attention.

Figure 5 shows higher average divergence for members than non-members, with greater attention concentration in deeper layers. These group-level differences suggest that attention concentration may provide a useful membership signal; they do not, by themselves, establish memorization or individual-sample separability.

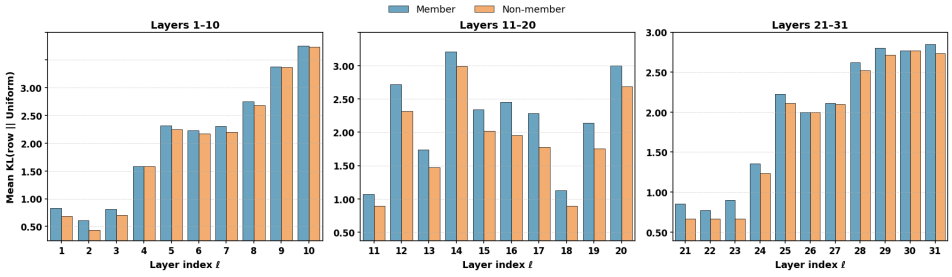


Figure 5: Layer-wise attention concentration on WikiMIA. Mean KL divergence from attention rows to a uniform distribution, averaged across heads and samples within each group. Larger values indicate more concentrated attention.

Response to token dropping. We remove seven tokens at fixed positions from each sample and measure the percentage change in attention at each layer. As shown in Figure 6, members generally exhibit larger changes than non-members. Both groups show substantial changes in early layers, while their responses remain distinguishable in deeper layers. This observation motivates using perturbation responses as an additional source of membership information.

Single-feature separability. We compare individual transitional features with baseline membership scores on the GitHub subset of MIMIR using Pythia-1.4B. We consider Consistency–Corr equation 2, Consistency–Frob equation 3, and Consistency–KL equation 1. Separability is measured using the Hellinger distance (HD) Nikulin et al. (2001) between member and non-member score distributions. HD ranges from 0 for identical distributions to 1 for distributions with disjoint support.

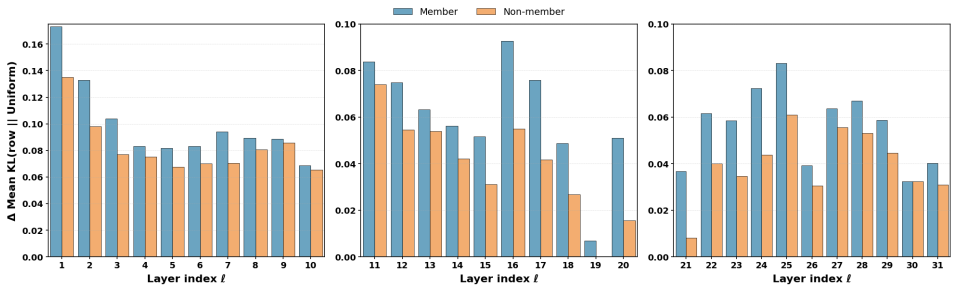


Figure 6: Layer-wise attention change after token removal on WikiMIA. Percentage change in attention after removing seven tokens at fixed positions, shown separately for members and non-members.

Table 5 reports the best-performing head-level feature from each consistency family alongside the baseline scores. These selected features achieve HD values of 0.70–0.76, compared with 0.41–0.52 for the baselines. This descriptive comparison suggests that individual attention features can provide useful membership information.

Table 5: Single-feature separability on MIMIR GitHub using Pythia-1.4B. Hellinger distance between member and non-member score distributions. Each consistency row reports the best-performing head-level feature in that family. Higher values indicate less overlap. Bold and underline indicate the highest and second-highest values, respectively.

Score	**Hellinger Distance (↑)**
LiRA	0.48
Loss	0.49
PPL	0.41
Zlib	0.52
Neighbor	0.50
Min-K%	0.51
Min-K%++	0.46
Consistency–Corr	0.70
Consistency–Frob	<u>0.75</u>
Consistency–KL	**0.76**

D Additional Analysis

D.1 Sequence Length

Figure 7 shows performance across WikiMIA sequence lengths, averaged over the evaluated target models. Membership separability generally decreases for longer sequences, particularly at TPR@1%FPR, although trends vary across models. Because the length settings contain different samples, this comparison does not isolate the effect of sequence length alone.

D.2 Model Scale

Figure 8 compares AttenMIA across model sizes on WikiMIA. TPR@1%FPR varies more noticeably than ROC AUC in the evaluated settings. These results describe membership detectability on this benchmark and do not directly measure how much training data a model memorizes.

D.3 Mixture-of-Experts and Lexical Shift

We evaluate AttenMIA (Transitional) on Qwen1.5-MoE-A2.7B (Bai et al., 2023) using standard WikiMIA and a lexically perturbed version. For the lexical-shift evaluation, we train the membership classifier on original WikiMIA samples and evaluate it on perturbed samples.

Table 6: Perturbation ablation on MIMIR with Pythia-1.4B. Each strategy is evaluated separately; All concatenate features from the three strategies. Average denotes the equally weighted mean across the seven domains. TPR@1%FPR is expressed as a percentage. Bold and underlined entries indicate the highest and second-highest distinct values within each column, respectively, including ties. Higher is better for both metrics.

Perturbation	Wikipedia		GitHub		Pile CC		PubMed Central	
	AUC	TPR@1%FPR	AUC	TPR@1%FPR	AUC	TPR@1%FPR	AUC	TPR@1%FPR
Prefix Insertion	0.62	10.4	0.93	75.9	0.60	**7.3**	0.86	31.0
Word Dropping	0.62	11.3	0.93	74.5	0.61	7.1	0.85	34.5
Word Replacement	0.62	9.4	0.94	73.0	0.59	4.2	0.85	26.5
All	**0.70**	**19.0**	**0.99**	**92.3**	**0.66**	5.1	**0.93**	**47.6**
Perturbation	**arXiv**		**DM Math**		**HackerNews**		**Average**	
	AUC	TPR@1%FPR	AUC	TPR@1%FPR	AUC	TPR@1%FPR	AUC	TPR@1%FPR
Prefix Insertion	0.82	28.1	0.93	**84.2**	0.61	**10.3**	0.77	35.3
Word Dropping	0.83	30.8	0.94	82.9	0.62	9.5	0.77	35.8
Word Replacement	0.82	27.5	0.93	80.7	0.61	8.9	0.77	32.9
All	**0.92**	**60.0**	**0.96**	84.1	**0.70**	9.3	**0.84**	**45.5**

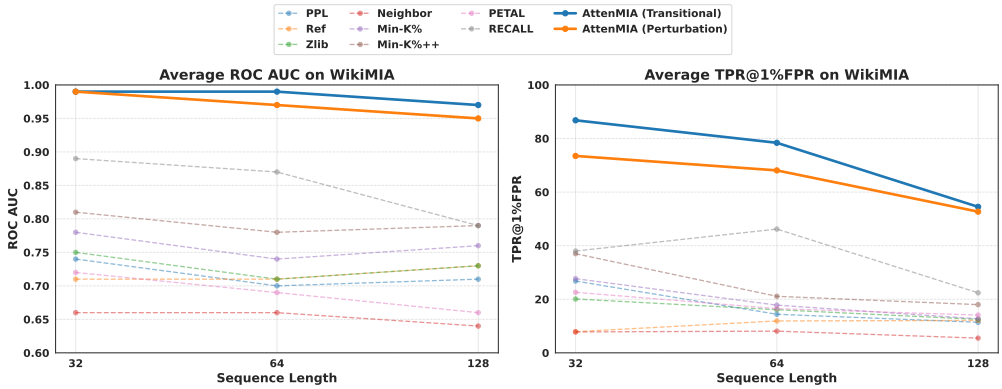


Figure 7: Sequence-length comparison on WikiMIA. ROC AUC and TPR@1%FPR averaged across the evaluated target models. Low-FPR performance generally decreases at longer sequence lengths.

Table 7 shows ROC AUC values of 0.92–0.99 on standard WikiMIA and 0.98–0.99 under lexical shift. TPR@1%FPR is lower under lexical shift at lengths 32 and 64, but higher at lengths 128 and 256. These results show that transitional attention features remain informative for this MoE model under the evaluated lexical transformation, while low-FPR performance depends on the evaluation setting.

E Association with Source-Text Overlap: Additional Results

This section supplements Section 4.3. From a pool of 14,000 GPT-2 generations, we select the 1,000 highest- and 1,000 lowest ROUGE-L examples and compute the Pearson correlation between each method's score and source-text overlap. All methods use the same 2,000 selected generations. The AttenMIA scorer combines Transitional and Perturbation features. ROUGE-L (Lin, 2004) measures textual overlap and does not independently establish training membership.

Score definitions and results. PPL-XL and PPL-Small denote perplexity under GPT-2-XL and GPT-2 Small, respectively; PPL-Lower denotes GPT-2-XL perplexity on lowercased text. Small/XL and Lower/XL divide the corresponding log perplexity by the original-text log perplexity under GPT-2-XL. Zlib/XL divides the compressed byte length by that same denominator.

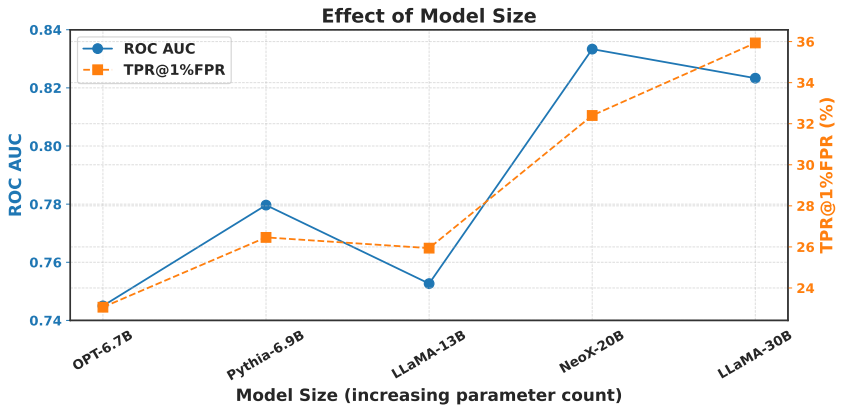


Figure 8: Model-scale comparison on WikiMIA. ROC AUC and TPR@1%FPR across evaluated model sizes. The observed differences reflect attack performance rather than a direct measure of memorization.

Table 7: MoE and lexical-shift results on WikiMIA. AttenMIA (Transitional) evaluated on Qwen1.5-MoE-A2.7B. Standard results use original WikiMIA samples; lexical-shift results train the membership classifier on original samples and evaluate it on lexically perturbed samples. TPR@1%FPR is reported as a percentage. Higher values are better.

Length	WikiMIA		WikiMIA → Perturbed	
	ROC AUC	TPR@1%FPR	ROC AUC	TPR@1%FPR
32	0.99	82.0	0.98	52.4
64	0.98	71.6	0.98	48.2
128	0.97	51.5	0.99	95.6
256	0.92	59.8	0.99	96.4

Table 8: Association between MIA scores and source-text overlap. Pearson correlation with ROUGE-L on the selected 2,000 generations. Correlations retain each score's original direction. Bold and underline indicate the largest and second-largest absolute correlations, respectively.

Score	Pearson r
Zlib/XL Ratio	<u>+0.32</u>
Lower/XL Ratio	+0.07
Small/XL Ratio	+0.05
Zlib Entropy	−0.09
PPL-XL	−0.21
PPL-Lower	−0.24
PPL-Small	−0.26
AttenMIA	**+0.48**

As shown in Table 8, AttenMIA has the largest observed absolute correlation ($r = 0.48$), followed by Zlib/XL ($r = 0.32$). The negative perplexity correlations indicate that lower perplexity is associated with greater source-text overlap in these samples.

Qualitative examples. Figure 9 shows two generations with high source-text overlap and high AttenMIA scores. They illustrate outputs that the scorer can surface for inspection, without independently demonstrating training membership.

Prefix: Egypt El Salvador Equatorial Guinea Eritrea

Generation: Estonia Ethiopia Falkland Islands (Malvinas) Fare Islands Fiji Finland France French Guiana French Polynesia French Southern Territories Georgia Germany Gibraltar ...

ROUGE-L: 0.90 **AttenMIA score:** 1.0

Prefix: how this page has evolved in the past.

Generation: "If you want to discuss contents of this page this is the easiest way to do it. View and manage ... General Wikidot.com documentation and help section. Wikidot.com Terms of Service ..."

ROUGE-L: 0.84 **AttenMIA score:** 0.78

Figure 9: Examples with high source-text overlap and high AttenMIA scores. Two generated continuations, with their prefixes, ROUGE-L values, and AttenMIA scores.

F ROC Curves Across MIMIR Domains

Figure 10 compares AttenMIA and the evaluated baselines across MIMIR domains using Pythia-1.4B. The curves show the trade-off between true-positive and false-positive rates across decision thresholds (Fawcett, 2006), complementing the ROC AUC and TPR@1%FPR results in the main text.

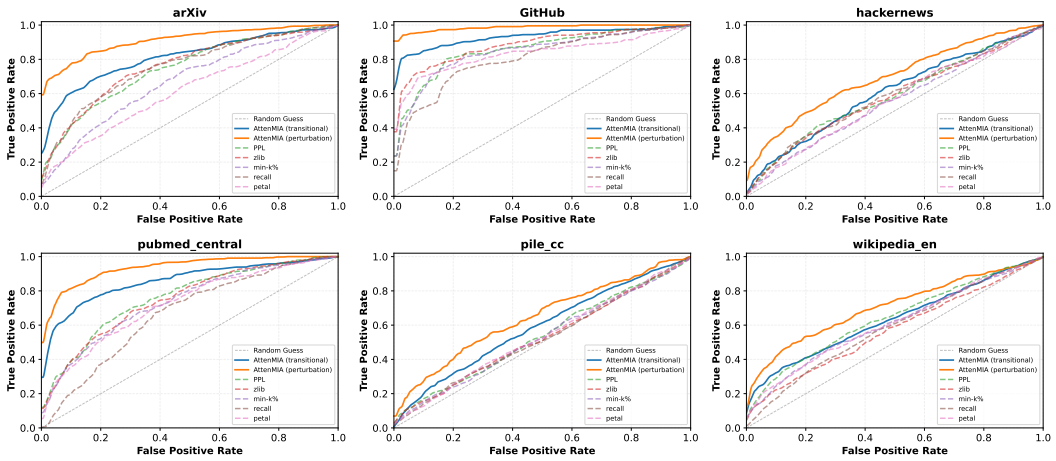


Figure 10: ROC curves across MIMIR domains. AttenMIA and baseline results using Pythia-1.4B as the target model. Curves are averaged over five evaluation folds.