

GI-MIA: Group-Invariant Membership Inference Attacks Against Large Language Models

Zichen Song

Sungkyunkwan University
Suwon, Republic of Korea
songzichen894@gmail.com

Farman Ali*

Sungkyunkwan University
Suwon, Republic of Korea
farman0977@skku.edu

Tamer Abuhmed*

Sungkyunkwan University
Suwon, Republic of Korea
tamer@skku.edu

Abstract

Membership inference attacks (MIAs) assess whether a sample was used to train a model, posing privacy risks for large language models (LLMs). Existing MIAs typically rely on single loss or confidence signals, which are often unstable and easily confounded by noise, leading to limited robustness and generalization. In this paper, we propose Group-Invariant Membership Inference Attack (GI-MIA) that explicitly exploits semantic-preserving transformations to capture stability differences between members and non-members. Our method constructs an orbit of paraphrased variants for each input using LLM-driven transformations, and extracts invariant statistics (mean, maximum, and variance of normalized losses) across the orbit. A simple decision rule based on these statistics effectively discriminates members from non-members by leveraging the model’s inherent invariance on training data. We conducted extensive experiments on the Pythia family of LLMs (different model parameter sizes) across six datasets. Results show that GI-MIA consistently outperforms strong baselines, achieving state-of-the-art performance while maintaining robustness against semantic-preserving transformations. Our study highlights new directions for MIAs and calls for further research into LLM privacy protection, emphasizing the need for developing reliable models with better privacy resilience.

1 Introduction

Large language models (LLMs) have achieved remarkable progress across a wide range of natural language processing tasks, yet their growing scale and adoption raise serious privacy concerns (Hayes et al., 2019). One fundamental threat is the membership inference attack (MIA), which aims to determine whether a given sample was included in a model’s training set (Groeneveld et al., 2024). Successful MIAs compromise individual privacy and

undermine the trustworthiness of deploying LLMs in sensitive applications such as healthcare, finance, and education (Soldaini et al., 2024). Recent research has developed a variety of MIAs (Liu et al., 2024), most of which rely on single-point statistics such as prediction loss (Buehler and Buehler, 2024), confidence scores, or gradient norms (Nam et al., 2022; Zhang et al., 2023). While effective in certain cases, these signals are inherently unstable: they are highly sensitive to noise (Colombo et al., 2024), dataset shifts, and model scale, which leads to limited robustness and poor generalization (Alves et al., 2024). More advanced attack methods, such as perturbation-based approaches in parameter space or sample space (Labrak et al., 2024), attempt to enhance signal strength (e.g., confidence scores, gradient norms) but often introduce additional variance and computational overhead (Hu et al., 2024). As a result, existing MIAs remain fragile, and their reliability across diverse models and datasets is far from guaranteed (Maini et al., 2024). In this paper, we argue that a key property of memorization in LLMs (stability under semantic-preserving transformations) has been largely overlooked (Brown et al., 2022). Intuitively, when a model has encountered a training sample, its prediction should remain stable even if the input is paraphrased or reordered while preserving meaning (Magar and Schwartz, 2022). In contrast, non-member samples tend to exhibit larger fluctuations under such transformations (Li, 2023). Capturing this invariance provides a principled way to distinguish members from non-members beyond conventional loss-based signals (Oren et al., 2024). To this end, we propose GI-MIA (Group-Invariant Membership Inference Attack), a novel framework that leverages LLM-driven paraphrasing to construct an orbit of semantically equivalent variants for each input. Rather than summarizing the orbit with simple statistics, GI-MIA leverages the full loss trajectory across paraphrased variants (a sequence of

*Corresponding author

cross-entropy losses computed for each semantically equivalent paraphrase of the same input) to reveal the inherent stability of member samples, which tend to exhibit low variance and flat curves. A simple decision rule based on these statistics enables reliable inference without complex training or auxiliary models. GI-MIA hypothesizes that member samples have more stable predictions under paraphrasing than non-members. We evaluate GI-MIA on the Pythia family of LLMs with different parameter sizes (1.4B, 2.8B, and 6.9B) across six diverse datasets: Arxiv, Wikipedia, PubMed Central, DM Mathematics, Pile-CC, and HackerNews. Experiments show that GI-MIA consistently outperforms strong baselines, including loss-based, perturbation-based, and embedding-based attacks, achieving state-of-the-art results while remaining robust to noise. Our contributions are threefold:

- **New perspective on membership inference.** We highlight group invariance as an overlooked yet powerful signal for distinguishing training members from non-members in large language models.
- **A novel attack framework.** We introduce GI-MIA, which leverages LLM-driven paraphrasing to construct semantic orbits and extract invariant statistics, enabling more robust membership inference than existing loss-based or perturbation-based methods.
- **Experiments.** We evaluate on Pythia models (1.4B, 2.8B, 6.9B) and six datasets, showing GI-MIA outperforms strong baselines with state-of-the-art robustness.

2 Motivation

Membership inference attacks (MIAs) attempt to determine whether a sample x was part of a model’s training set. While numerous methods have been proposed, most rely on single-point loss or confidence signals, which are often unstable and poorly generalize across datasets and model scales. We identify two key motivations for moving beyond conventional approaches (Li and Flanigan, 2024).

Threat Model. We consider a black-box threat model where the adversary has query access to the target LLM and can obtain output probabilities for arbitrary inputs. The attacker aims to infer whether a target input was part of the model’s training data. We assume the attacker can generate semantically equivalent variants (via paraphrasing APIs) and can

compute supervised losses using known or inferred labels. No access to model internals, training data, or gradient information is required. The core assumption is that member samples exhibit stable loss trajectories under semantic-preserving transformations, while non-members display higher variability. Our method exploits this discrepancy to infer membership without relying on shadow models or architecture-specific knowledge.

2.1 Limitation of Single-Point Loss Signals

Let $x \in \mathcal{X}$ denote an input with label y , and define its supervised loss under model f_θ as

$$\ell(x) = -\log p_\theta(y \mid x). \quad (1)$$

Traditional MIAs assume that member samples ($x \in D_{\text{in}}$) yield smaller losses than non-members ($x \in D_{\text{out}}$). Formally, the loss distributions can be expressed as

$$\ell(x) \sim \begin{cases} \mathbb{P}_{\text{in}}(\ell), & x \in D_{\text{in}}, \\ \mathbb{P}_{\text{out}}(\ell), & x \in D_{\text{out}}. \end{cases} \quad (2)$$

The Bayes-optimal classifier is a threshold rule

$$y = \mathbb{1}\{\ell(x) < \tau\}, \quad (3)$$

where τ minimizes classification error. However, when $\mathbb{P}_{\text{in}}$ and $\mathbb{P}_{\text{out}}$ overlap significantly, the decision boundary becomes highly unstable. As shown by Shokri et al. (2017), this overlap leads to high variance in attack success (Dong et al., 2024). More formally, Sablayrolles et al. (2019) demonstrate that the generalization gap bounds the distinguishability between $\mathbb{P}_{\text{in}}$ and $\mathbb{P}_{\text{out}}$, implying that single-point loss signals are fundamentally limited when the distributions are close. This motivates the search for more robust features.

2.2 Stability Under Semantic-Preserving Transformations

A less explored but fundamental property of memorization in large language models is *stability*. Intuitively, if the model has encountered a sample x during training, its prediction should remain consistent even when x is paraphrased or reordered while preserving semantics. Let $O(x) = \{\tilde{x}^{(1)}, \dots, \tilde{x}^{(K)}\}$ denote an orbit of x generated by semantic-preserving transformations $g \in G$. For each variant $\tilde{x}^{(k)}$, the supervised loss is

$$\ell_k(x) = -\log p_\theta(y \mid \tilde{x}^{(k)}). \quad (4)$$

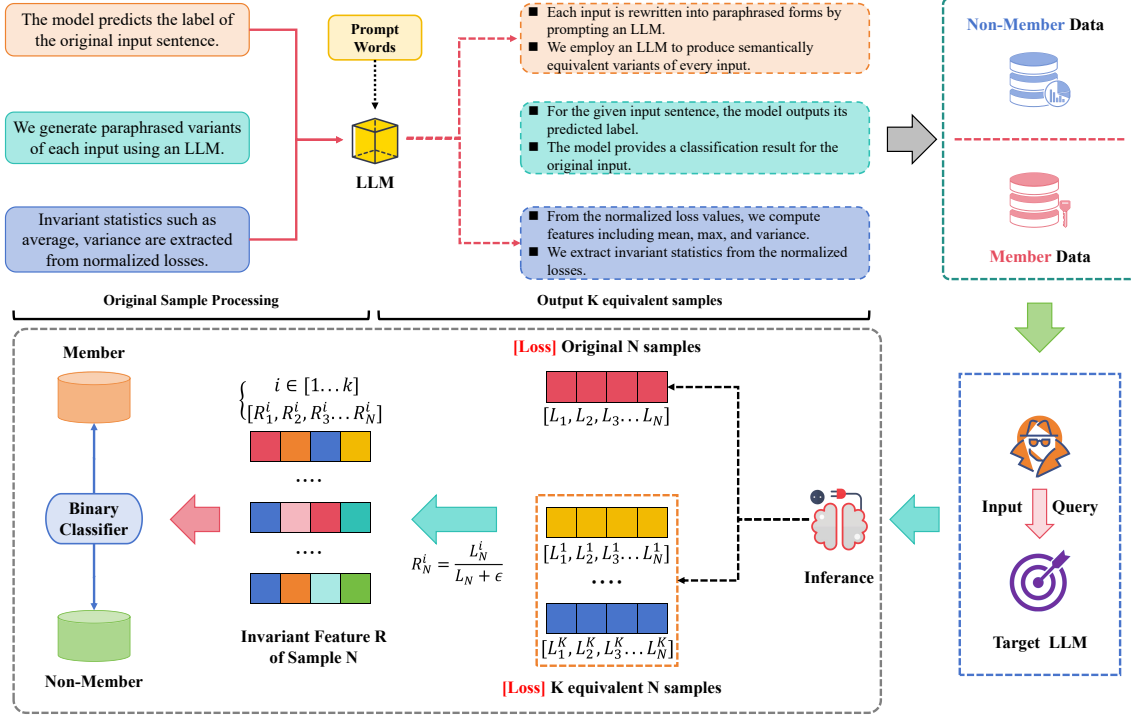


Figure 1: Overview of the proposed **GI-MIA** attack framework. In the first stage (*Original Sample Processing*), the input is fed into an LLM to obtain the supervised loss. In the second stage (*Output K equivalent samples*), we use LLM-driven paraphrasing to generate semantically equivalent variants of the input and compute their losses. Normalized losses are then aggregated into invariant statistics to form feature representations. Finally, a binary classifier leverages these features to distinguish *member* from *non-member* data.

We hypothesize that the variance of orbit losses provides a discriminative signal:

$$\mathbb{V}_k[\ell_k(x)] \text{ is small, } x \in D_{\text{in}}, \quad (5)$$

$$\mathbb{V}_k[\ell_k(x)] \text{ is large, } x \in D_{\text{out}}. \quad (6)$$

This claim is supported by prior work on memorization and extractability. show that LLMs exhibit strong recall of training samples even under paraphrases, while Maini et al. (2021) argue that training data induces smoother decision boundaries in its neighborhood. Similarly, Hu et al. (2022) observe that MIA success correlates with robustness of predictions to input perturbations. To operationalize this idea, we normalize orbit losses by the base loss $\ell_{\text{base}}(x)$:

$$r_k(x) = \frac{\ell_k(x)}{\ell_{\text{base}}(x) + \epsilon}, \quad (7)$$

where ϵ ensures numerical stability. Each sample can thus be represented by its invariance profile

$$R(x) = [r_1(x), \dots, r_K(x)]. \quad (8)$$

Unlike single-point loss values, $R(x)$ captures distributional behavior under semantic transforma-

tions, providing a more robust and theoretically grounded signal for membership inference.

Algorithm 1 GI-MIA: Group-Invariant Membership Inference Attack

Input: Target model f , input sample x , paraphraser $\mathcal{T}$, number of transformations K , thresholds $\tau_{\text{avg}}, \tau_{\text{std}}$

Output: Membership prediction $y \in \{\text{member}, \text{non-member}\}$

```

1: Initialize empty list  $\mathcal{S}$ 
2: for  $k = 1$  to  $K$  do
3:    $x^{(k)} \leftarrow \mathcal{T}(x)$ 
4:    $s^{(k)} \leftarrow \text{score}(f, x^{(k)})$ 
5:   Append  $s^{(k)}$  to  $\mathcal{S}$ 
6: end for
7: Compute  $\mu \leftarrow \text{mean}(\mathcal{S})$ 
8: Compute  $\sigma \leftarrow \text{std}(\mathcal{S})$ 
9: if  $\mu > \tau_{\text{avg}}$  and  $\sigma < \tau_{\text{std}}$  then
10:   $y \leftarrow \text{member}$ 
11: else
12:   $y \leftarrow \text{non-member}$ 
13: end if
14: return  $y$ 

```

3 Method

We propose **GI-MIA**, a framework that leverages semantic-preserving transformations to model stability in large language models (LLMs). Unlike prior MIAs that rely on a single loss or confidence

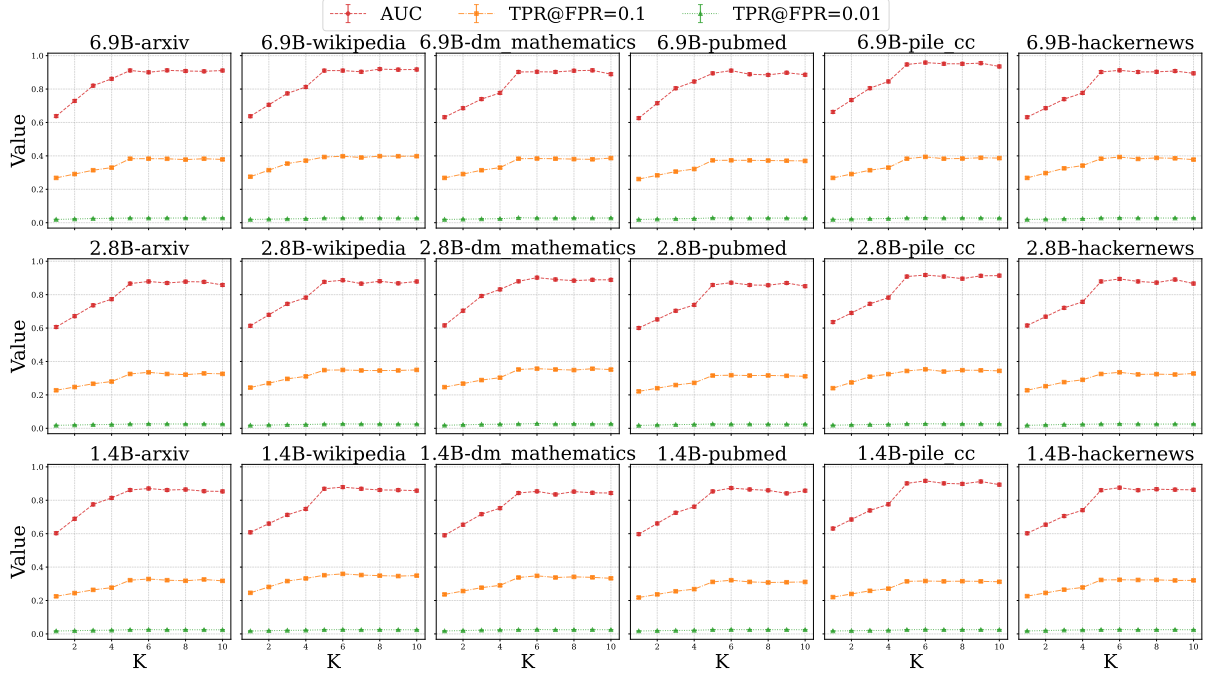


Figure 2: Performance of our proposed GI-MIA method across different model sizes (1.4B, 2.8B, and 6.9B) and datasets (Arxiv, Wikipedia, DM Mathematics, PubMed, Pile-CC, and HackerNews). The plots show AUC, TPR@FPR=0.1, and TPR@FPR=0.01 as functions of the number of paraphrases K . Our method consistently achieves higher AUC and TPR values as K increases, demonstrating the effectiveness of semantic-preserving transformations for distinguishing members from non-members.

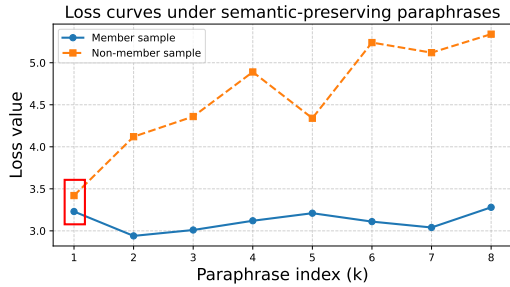


Figure 3: Two samples (one member, one non-member) misclassified by a loss-based attack. Their raw losses (red box) are indistinguishable, but after semantic-preserving transformations, the orbit curves diverge. The member remains stable while the non-member fluctuates, enabling accurate classification with GI-MIA (arXiv dataset).

score, GI-MIA constructs paraphrase-based orbits and derives discriminative features that are subsequently classified by a support vector machine. This design allows us to move beyond handcrafted thresholds and to exploit richer decision boundaries in the feature space.

3.1 Orbit Sampling

Let $D = D_{\text{in}} \cup D_{\text{out}}$ denote the dataset containing member samples (D_{in}) and non-member samples

(D_{out}). Each input $x_i \in D$ is expanded into an orbit $O(x_i)$, defined as the collection of semantically equivalent variants generated by a transformation group G . In this work, G corresponds to paraphrasing operations, such as synonym substitution or word-order reorganization, that preserve meaning but alter surface realization. Formally, the orbit is defined as:

$$O(x_i) = \{\tilde{x}_i^{(1)}, \tilde{x}_i^{(2)}, \dots, \tilde{x}_i^{(K)}\}. \quad (9)$$

This step captures the intuition that memorized samples should be robust to semantic-preserving perturbations (Hilprecht et al., 2019). If the model has encountered x_i during training, its prediction should remain stable even under reformulated inputs (Stadler et al., 2022). By contrast, for non-member samples, the model may respond inconsistently across paraphrases (Shejwalkar et al., 2021; Ye et al., 2022). The orbit construction therefore provides a natural basis for probing the invariance of model predictions (Shilov et al., 2026).

3.2 Base Loss

For the canonical input x_i with label y_i , we compute the supervised loss of the target model f_θ . The

Table 1: AUC scores of different MIA methods on Pythia-6.9B across six datasets.(BOW is bag of words)

MIA Method	ArXiv	DM-Math	HackerNews	Pile-CC	PubMed	Wikipedia
EM-MIA	0.6017 ± 0.0045	0.6001 ± 0.0039	0.6016 ± 0.0044	0.6124 ± 0.0045	0.5914 ± 0.0042	0.6121 ± 0.0043
MOPE	0.6993 ± 0.0133	0.6949 ± 0.0132	0.6993 ± 0.0133	0.7194 ± 0.0134	0.6874 ± 0.0133	0.7002 ± 0.0132
SPV-MIA	0.5059 ± 0.0072	0.5055 ± 0.0071	0.5146 ± 0.0072	0.5259 ± 0.0072	0.5064 ± 0.0072	0.5231 ± 0.0072
BOW	0.5012 ± 0.0141	0.5012 ± 0.0145	0.4881 ± 0.0143	0.5101 ± 0.0143	0.4902 ± 0.0143	0.5011 ± 0.0143
Loss	0.4929 ± 0.0118	0.4923 ± 0.0112	0.5126 ± 0.0118	0.5129 ± 0.0118	0.4821 ± 0.0118	0.5069 ± 0.0116
Lower	0.4994 ± 0.0168	0.4991 ± 0.0162	0.4993 ± 0.0168	0.5093 ± 0.0168	0.4991 ± 0.0165	0.5093 ± 0.0168
Min-K%++	0.5161 ± 0.0058	0.5164 ± 0.0056	0.5107 ± 0.0058	0.5263 ± 0.0059	0.5052 ± 0.0058	0.5061 ± 0.0059
Min-K% Prob	0.5087 ± 0.0141	0.5086 ± 0.0141	0.5185 ± 0.0141	0.5187 ± 0.0141	0.5042 ± 0.0141	0.5132 ± 0.0141
Neighborhood	0.4935 ± 0.0156	0.4933 ± 0.0158	0.4869 ± 0.0158	0.5032 ± 0.0157	0.4911 ± 0.0158	0.5035 ± 0.0158
PAC	0.5012 ± 0.0063	0.5011 ± 0.0063	0.4811 ± 0.0063	0.5011 ± 0.0063	0.5011 ± 0.0063	0.4913 ± 0.0063
SaMIA	0.5023 ± 0.0091	0.5021 ± 0.0092	0.5124 ± 0.0092	0.5123 ± 0.0092	0.5001 ± 0.0091	0.5021 ± 0.0091
SURP	0.5072 ± 0.0075	0.5074 ± 0.0074	0.5217 ± 0.0075	0.5121 ± 0.0075	0.5023 ± 0.0075	0.5062 ± 0.0075
Zlib	0.5123 ± 0.0033	0.5124 ± 0.0034	0.5134 ± 0.0034	0.5132 ± 0.0034	0.5083 ± 0.0033	0.5134 ± 0.0034
Our (GI-MIA)	0.9120 ± 0.0087	0.9033 ± 0.0088	0.9121 ± 0.0089	0.9584 ± 0.0088	0.9104 ± 0.0086	0.9195 ± 0.0088

Table 2: Comparison of GI-MIA (ours) and baseline membership inference attacks across key dimensions. ✓ denotes support, ✗ denotes lack of support, and N/A indicates not applicable. "Shadow" refers to whether the method uses a shadow model; "Signal" indicates the main inference signal; "Paraphr." reflects the use of paraphrased or transformed inputs; "Group-Inv." indicates group-invariant prediction consistency; "Low FPR" measures robustness under strict false positive constraints; "LLM Scal." refers to applicability to large language models; "Cross-Domain" assesses generalization across diverse datasets.

MIA Method	Shadow	Signal	Paraphr.	Group-Inv.	Low FPR	LLM Scal.	Cross-Domain
EM-MIA (Song et al., 2025)	✗	Loss	✗	✗	✗	✓	✗
MOPE (Li et al., 2023)	✗	Entropy	✓	✗	✗	✓	✓
SPV-MIA (Fu et al., 2024)	✓	Vector dist	✓	✗	✗	✗	✗
BOW (Meeus et al., 2025)	✗	N-gram	N/A	✗	✗	✗	✗
Loss (Yeom et al., 2018)	✗	Loss	✗	✗	✗	✓	✗
Lower (Carlini et al., 2021)	✗	Token match	✗	✗	✗	✗	✗
Min-K%++ (Zhang et al., 2025)	✗	Likelihood	✗	✗	✗	✓	✗
Min-K% Prob (Shi et al., 2024)	✗	Prob score	✓	✗	✗	✓	✗
Neighborhood (Mattern et al., 2023)	✗	Embedding	N/A	✗	✗	✓	✗
PAC (Ye et al., 2024)	✗	Conf. gap	✓	✗	✗	✗	✗
SaMIA (Kaneko et al., 2025)	✓	ROUGE sim	✓	✗	✗	✗	✗
SURP (Zhang and Wu, 2024)	✗	Surprisal	✓	✗	✗	✗	✗
Zlib (Carlini et al., 2021)	✗	Compress.	✗	✗	✗	✗	✗
Our (GI-MIA)	✗	Loss var.	✓	✓	✓	✓	✓

negative log-likelihood (NLL) serves as the base measure:

$$\ell_{\text{base}}^{(i)} = -\log p_{\theta}(y_i | x_i). \quad (10)$$

This base loss serves as an anchor against which the losses of the paraphrased inputs are compared. Intuitively, $\ell_{\text{base}}^{(i)}$ reflects how well the model fits the unaltered representation of the sample. If x_i is a member, the model should have assigned it a relatively high probability during training, resulting in a lower base loss. While base loss alone is insufficient for reliable inference, it provides a crucial reference point for normalizing orbit losses (Villalobos et al., 2024).

3.3 Supervised Loss and Label Construction

This section clarifies the construction of the supervised loss used by GI-MIA, the source of labels in

black-box settings, and how this protocol ensures fair comparison with baseline attacks.

Supervised loss definition. For a target LLM f_{θ} and an input x , GI-MIA evaluates the negative log-likelihood (NLL) of a label y under the model:

$$\ell(x, y) = -\log p_{\theta}(y | x). \quad (11)$$

For each semantic-preserving variant $\tilde{x}^{(k)}$ in the orbit $O(x)$, the corresponding loss is

$$\ell_k(x) = -\log p_{\theta}(y | \tilde{x}^{(k)}). \quad (12)$$

To remove scale effects due to tokenization, label length, or model calibration, GI-MIA uses normalized loss ratios

$$r_k(x) = \frac{\ell_k(x)}{\ell_{\text{base}}(x) + \varepsilon}, \quad (13)$$

where $\ell_{\text{base}}(x) = \ell(x, y)$ and ε is a small constant for numerical stability. All invariant features and

Table 3: TPR at FPR = 0.1 for different MIA methods on Pythia-6.9B across six datasets.

MIA Method	ArXiv	DM-Math	HackerNews	Pile-CC	PubMed	Wikipedia
EM-MIA	0.1861 ± 0.0164	0.1864 ± 0.0166	0.1862 ± 0.0166	0.1872 ± 0.0166	0.1854 ± 0.0166	0.2061 ± 0.0166
MOPE	0.1957 ± 0.0103	0.1957 ± 0.0103	0.1955 ± 0.0103	0.1964 ± 0.0103	0.1853 ± 0.0103	0.2057 ± 0.0103
SPV-MIA	0.0807 ± 0.0136	0.0806 ± 0.0136	0.0804 ± 0.0136	0.0817 ± 0.0136	0.0807 ± 0.0136	0.1002 ± 0.0136
BOW	0.0852 ± 0.0131	0.0852 ± 0.0131	0.0852 ± 0.0131	0.0851 ± 0.0131	0.0852 ± 0.0131	0.0953 ± 0.0131
Loss	0.0788 ± 0.0061	0.0788 ± 0.0061	0.0788 ± 0.0061	0.0788 ± 0.0061	0.0788 ± 0.0061	0.0821 ± 0.0061
Lower	0.1025 ± 0.0112	0.1025 ± 0.0112	0.1025 ± 0.0112	0.1025 ± 0.0112	0.1025 ± 0.0112	0.1121 ± 0.0112
Min-K%++	0.0985 ± 0.0131	0.0985 ± 0.0131	0.0985 ± 0.0131	0.0985 ± 0.0131	0.0985 ± 0.0131	0.1012 ± 0.0131
Min-K% Prob	0.0900 ± 0.0166	0.0901 ± 0.0166	0.0901 ± 0.0166	0.0901 ± 0.0166	0.0901 ± 0.0166	0.0901 ± 0.0166
Neighborhood	0.0830 ± 0.0183	0.0831 ± 0.0183	0.0831 ± 0.0183	0.0831 ± 0.0183	0.0831 ± 0.0183	0.0831 ± 0.0183
PAC	0.0929 ± 0.0061	0.0929 ± 0.0061	0.0929 ± 0.0061	0.0929 ± 0.0061	0.0929 ± 0.0061	0.0929 ± 0.0062
SaMIA	0.0732 ± 0.0036	0.0732 ± 0.0036	0.0732 ± 0.0036	0.0732 ± 0.0036	0.0732 ± 0.0036	0.0831 ± 0.0036
SURP	0.0972 ± 0.0163	0.0972 ± 0.0163	0.0972 ± 0.0163	0.0972 ± 0.0163	0.0972 ± 0.0163	0.0972 ± 0.0163
Zlib	0.1105 ± 0.0104	0.1105 ± 0.0105	0.1105 ± 0.0105	0.1105 ± 0.0105	0.1105 ± 0.0105	0.1122 ± 0.0105
Our (GI-MIA)	0.3832 ± 0.0063	0.3841 ± 0.0063	0.3931 ± 0.0063	0.3941 ± 0.0063	0.3735 ± 0.0063	0.3983 ± 0.0063

classifiers operate in these ratios rather than raw losses, making the attack insensitive to the magnitude of absolute loss.

Source of labels in black-box settings. GI-MIA does *not* require access to ground-truth labels. It only requires that the *same label* be used consistently across the original input and its paraphrased variants. In practice, the label y can be obtained in three standard ways:

1. From public datasets that provide ground-truth classification labels (e.g., topic or domain labels);
2. From the target model itself by pseudo-labeling:

$$\hat{y} = \arg \max_y p_\theta(y | x); \quad (14)$$

3. From weak or heuristic labelers commonly used in black-box MIA settings.

Once y (or $\hat{y}$) is fixed for x , the same label is used for all paraphrases $\tilde{x}^{(k)}$. Importantly, GI-MIA does not require the label to be correct; it only requires label consistency across the orbit, which is sufficient to compute stability statistics.

Consistency with baseline attacks. All baseline methods (e.g., EM-MIA, MOPE, Min-K%, PAC, Neighborhood) are evaluated using the *same* supervised loss definition, the *same* label source (ground-truth or pseudo-label), and the *same* member/non-member splits. Therefore, any performance differences reflect differences in the extracted signals rather than inconsistencies in the underlying oracle or labeling protocol. This protocol aligns with standard black-box MIA assumptions and ensures that GI-MIA and all baselines operate under a unified and fair threat model.

3.4 Orbit Losses

Once the orbit $O(x_i)$ constructed, the model computes supervised losses of each variant $\tilde{x}_i^{(k)}$:

$$\ell_k^{(i)} = -\log p_\theta(y_i | \tilde{x}_i^{(k)}), \quad k = 1, \dots, K. \quad (15)$$

These orbit losses reflect the model’s behavior on paraphrased yet semantically equivalent samples (Das et al., 2025). The key hypothesis is that member samples should induce more consistent losses across their orbit, while non-members exhibit larger fluctuations (Lee and Lemieux, 2010). This is a memorization effect: training samples are encoded more robustly, and the model generalizes their semantics more reliably (Das et al., 2025). By contrast, for unseen samples, even minor surface-level perturbations can destabilize predictions (Duarte et al., 2024). Thus, the collection $\{\ell_k^{(i)}\}$ provides a window into the stability profile of each input (Yao et al., 2024).

3.5 Invariant Feature Extraction

Algorithm 1 and Figure 4 provide an intuitive low-dimensional visualization based on mean and variance statistics, while all experiments use SVC trained on the full invariant feature vector.

To quantify invariance, we normalize orbit losses with respect to the base loss:

$$r_k^{(i)} = \frac{\ell_k^{(i)}}{\ell_{\text{base}}^{(i)} + \epsilon}, \quad (16)$$

where ϵ is a small constant for numerical stability. This normalization highlights the relative deviation of each paraphrased variant compared to the canonical input. The resulting invariant feature vector is defined as:

$$R^{(i)} = [r_1^{(i)}, r_2^{(i)}, \dots, r_K^{(i)}]. \quad (17)$$

Table 4: TPR at FPR = 0.01 for different MIA methods on Pythia-6.9B across six datasets.

MIA Method	ArXiv	DM-Math	HackerNews	Pile-CC	PubMed	Wikipedia
EM-MIA	0.0142 ± 0.0031	0.0142 ± 0.0032	0.0141 ± 0.0031	0.0153 ± 0.0032	0.0141 ± 0.0031	0.0241 ± 0.0032
MOPE	0.0175 ± 0.0042	0.0175 ± 0.0041	0.0174 ± 0.0041	0.0183 ± 0.0042	0.0175 ± 0.0042	0.0175 ± 0.0042
SPV-MIA	0.0105 ± 0.0037	0.0104 ± 0.0037	0.0103 ± 0.0034	0.0112 ± 0.0037	0.0105 ± 0.0037	0.0105 ± 0.0037
BOW	0.0061 ± 0.0021	0.0061 ± 0.0021	0.0061 ± 0.0021	0.0068 ± 0.0022	0.0061 ± 0.0021	0.0101 ± 0.0021
Loss	0.0066 ± 0.0038	0.0066 ± 0.0037	0.0066 ± 0.0038	0.0066 ± 0.0038	0.0066 ± 0.0038	0.0106 ± 0.0038
Lower	0.0111 ± 0.0035	0.0112 ± 0.0036	0.0111 ± 0.0035	0.0111 ± 0.0035	0.0111 ± 0.0035	0.0111 ± 0.0036
Min-K%++	0.0054 ± 0.0015	0.0054 ± 0.0015	0.0054 ± 0.0015	0.0054 ± 0.0015	0.0054 ± 0.0015	0.0054 ± 0.0015
Min-K% Prob	0.0125 ± 0.0039	0.0125 ± 0.0039	0.0125 ± 0.0039	0.0125 ± 0.0039	0.0125 ± 0.0039	0.0125 ± 0.0039
Neighborhood	0.0052 ± 0.0033	0.0052 ± 0.0032	0.0052 ± 0.0033	0.0052 ± 0.0033	0.0052 ± 0.0033	0.0052 ± 0.0033
PAC	0.0083 ± 0.0034	0.0083 ± 0.0034	0.0083 ± 0.0034	0.0083 ± 0.0034	0.0083 ± 0.0034	0.0083 ± 0.0034
SaMIA	0.0058 ± 0.0031	0.0057 ± 0.0032	0.0058 ± 0.0031	0.0058 ± 0.0031	0.0058 ± 0.0031	0.0108 ± 0.0032
SURP	0.0066 ± 0.0038	0.0066 ± 0.0038	0.0066 ± 0.0038	0.0066 ± 0.0038	0.0066 ± 0.0038	0.0066 ± 0.0038
Zlib	0.0066 ± 0.0031	0.0065 ± 0.0032	0.0066 ± 0.0031	0.0066 ± 0.0031	0.0066 ± 0.0031	0.0106 ± 0.0032
Our (GI-MIA)	0.0282 ± 0.0017	0.0297 ± 0.0017	0.0288 ± 0.0017	0.0289 ± 0.0017	0.0287 ± 0.0017	0.0279 ± 0.0017

This representation encodes the stability of model predictions under semantic-preserving transformations. Unlike handcrafted statistics (e.g., mean, variance), which may discard information, the full ratio vector preserves richer structural signals. These signals allow downstream classifiers to learn non-linear decision boundaries that more effectively distinguish members from non-members.

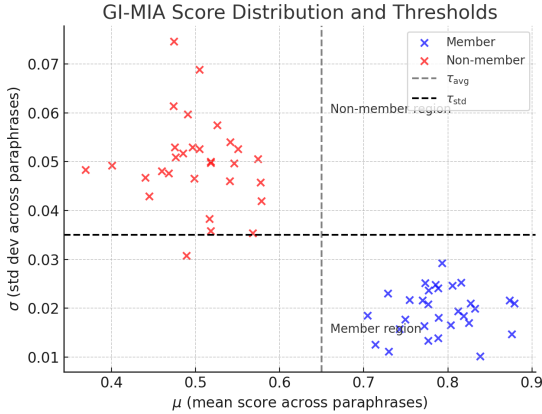


Figure 4: Visualization of GI-MIA’s decision mechanism in the 2D space of mean score (μ) and standard deviation (σ) across paraphrased queries. Samples in the lower-right region (high μ , low σ) are classified as members, while others are classified as non-members, using thresholds τ_{avg} and τ_{std} .

3.6 Classification via SVC

Rather than using rule-based thresholds on average deviation or variance, we adopt a Support Vector Classifier (SVC) to separate members from non-members in the invariant feature space. Given training pairs $\{(R^{(i)}, y^{(i)})\}$, where $y^{(i)} = 1$ if $x_i \in D_{\text{in}}$ and 0 otherwise, the SVC optimizes:

$$\min_{w,b} \frac{1}{2} \|w\|^2 + C \sum_i \xi_i, \quad (18)$$

$$\text{s.t. } y^{(i)}(w^\top R^{(i)} + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad (19)$$

where C is a regularization parameter and ξ_i are slack variables. The decision function is then:

$$\hat{y}^{(i)} = \text{sign}(w^\top R^{(i)} + b). \quad (20)$$

This classifier-based approach has two advantages. First, it avoids reliance on manually chosen thresholds, which may be suboptimal across datasets and model scales. Second, it allows the decision boundary to adapt to the geometry of the invariant feature space, capturing more complex separations between members and non-members. As our experiments show, this significantly enhances robustness and generalization compared to rule-based methods.

4 Experiments

We conduct comprehensive experiments to evaluate GI-MIA’s effectiveness, robustness, and scalability across varying model sizes and diverse benchmark datasets. Our evaluation targets three main goals: (1) demonstrating GI-MIA’s superiority over prior membership inference methods; (2) analyzing how model scale influences attack efficacy; and (3) assessing performance under strict false positive constraints, crucial for privacy-critical applications. All experiments are conducted on a unified testbed using three Pythia models (1.4B, 2.8B, 6.9B) and six datasets, i.e., ArXiv, PubMed, Wikipedia, DM-Math, HackerNews, and Pile-CC, spanning formal, encyclopedic, and informal domains to comprehensively test scale-sensitive and domain-generalization behaviors.

4.1 Performance on Pythia-6.9B

We evaluate all methods on the largest model in our suite, Pythia-6.9B, which contains approximately

Table 5: GI-MIA performance across Pythia model sizes (1.4B, 2.8B, 6.9B). Reported as mean $\pm$ std for AUC, TPR@FPR=0.1, and TPR@FPR=0.01.

Dataset	Metric	1.4B	2.8B	6.9B
ArXiv	AUC	0.8705 $\pm$ 0.0088	0.8789 $\pm$ 0.0088	0.9120 $\pm$ 0.0087
	TPR@0.1	0.3286 $\pm$ 0.0063	0.3353 $\pm$ 0.0063	0.3832 $\pm$ 0.0063
	TPR@0.01	0.0259 $\pm$ 0.0017	0.0266 $\pm$ 0.0017	0.0282 $\pm$ 0.0017
DM-Math	AUC	0.8534 $\pm$ 0.0088	0.9019 $\pm$ 0.0088	0.9033 $\pm$ 0.0088
	TPR@0.1	0.3477 $\pm$ 0.0063	0.3571 $\pm$ 0.0063	0.3841 $\pm$ 0.0063
	TPR@0.01	0.0263 $\pm$ 0.0017	0.0281 $\pm$ 0.0012	0.0297 $\pm$ 0.0017
HackerNews	AUC	0.8754 $\pm$ 0.0089	0.8942 $\pm$ 0.0089	0.9121 $\pm$ 0.0089
	TPR@0.1	0.3241 $\pm$ 0.0063	0.3353 $\pm$ 0.0063	0.3931 $\pm$ 0.0063
	TPR@0.01	0.0264 $\pm$ 0.0017	0.0266 $\pm$ 0.0017	0.0288 $\pm$ 0.0017
Pile-CC	AUC	0.9162 $\pm$ 0.0088	0.9175 $\pm$ 0.0088	0.9584 $\pm$ 0.0088
	TPR@0.1	0.3167 $\pm$ 0.0063	0.3531 $\pm$ 0.0063	0.3941 $\pm$ 0.0063
	TPR@0.01	0.0265 $\pm$ 0.0017	0.0271 $\pm$ 0.0017	0.0289 $\pm$ 0.0017
PubMed	AUC	0.8732 $\pm$ 0.0086	0.8718 $\pm$ 0.0086	0.9104 $\pm$ 0.0086
	TPR@0.1	0.3214 $\pm$ 0.0063	0.3179 $\pm$ 0.0063	0.3735 $\pm$ 0.0063
	TPR@0.01	0.0261 $\pm$ 0.0017	0.0257 $\pm$ 0.0017	0.0287 $\pm$ 0.0017
Wikipedia	AUC	0.8788 $\pm$ 0.0088	0.8864 $\pm$ 0.0088	0.9195 $\pm$ 0.0088
	TPR@0.1	0.3594 $\pm$ 0.0063	0.3493 $\pm$ 0.0063	0.3983 $\pm$ 0.0063
	TPR@0.01	0.0259 $\pm$ 0.0017	0.0263 $\pm$ 0.0017	0.0279 $\pm$ 0.0017

6.9 billion parameters and thus represents a rigorous test of attack capability. As shown in Table 1, GI-MIA achieves substantially higher AUC across all six datasets, reaching 0.9033 on DM-Math versus 0.6949 for MOPE and 0.6001 for EM-MIA, and 0.9120 on ArXiv. These consistent gains, often exceeding +20 AUC points, demonstrate that our group-invariant framework captures subtle memorization signals lost to baseline techniques relying solely on loss magnitude or entropy. GI-MIA also excels under stringent low-FPR regimes critical for privacy-sensitive deployments. As reported in Table 4, it attains a TPR@FPR=0.01 of 0.0287 on PubMed and 0.0297 on DM-Math, compared with MOPE’s 0.0175 on both datasets. This improvement highlights the robustness and precision of GI-MIA under tight false-positive budgets, underscoring its suitability for real-world scenarios where privacy risks and false alarms must be minimized.

4.2 Scaling Across Model Sizes

To evaluate the effect of model capacity on attack performance, we apply GI-MIA and baselines to Pythia-1.4B, 2.8B, and 6.9B. As shown in Tables 6 and 9, GI-MIA becomes more effective with increasing model size. For example, on HackerNews, AUC improves from 0.8754 (1.4B) to 0.8942 (2.8B) and 0.9121 (6.9B), indicating that larger models preserve more fine-grained training signals. This trend also holds under low-FPR

constraints: TPR@FPR=0.01 on ArXiv improves steadily from 0.0259 (1.4B) to 0.0282 (6.9B), showing that GI-MIA remains robust under strict privacy requirements. Moreover, the performance gap between GI-MIA and prior methods remains substantial across model sizes. On PubMed, our method outperforms MOPE by +0.2272, +0.2188, and +0.2230 AUC at 1.4B, 2.8B, and 6.9B, respectively, reinforcing the idea that GI-MIA benefits from scale and also scales *better* than existing approaches. These results emphasize that GI-MIA is particularly well-suited for auditing high-capacity LLMs.

4.3 Cross-Dataset Generalization

GI-MIA is evaluated on six diverse datasets, covering general-purpose (e.g., Wikipedia) and domain-specific corpora (e.g., DM-Math, PubMed). On Wikipedia, our method achieves AUCs of 0.9195 (6.9B), 0.8864 (2.8B), and 0.8788 (1.4B), significantly surpassing baselines. This shows that GI-MIA does not overfit to a particular distribution and generalizes well across balanced natural text. On DM-Math, GI-MIA achieves an AUC of 0.9033, outperforming EM-MIA (0.6001) and MOPE (0.6949), which struggle due to low lexical redundancy and rigid syntax. Our approach remains effective by leveraging semantic-preserving transformations to uncover invariance signals. Moreover, GI-MIA maintains strong performance on noisy or colloquial datasets such as

HackerNews and Pile-CC, demonstrating its flexibility in casual text. This cross-domain robustness is crucial for practical deployment scenarios, where LLMs are trained on heterogeneous and unpredictable data. The results confirm that GI-MIA effectively captures generalizable membership signals beyond surface-level statistics.

4.4 Comparison with Baseline Methods

We compare GI-MIA against 13 representative MIA baselines from four paradigms: loss-based (EM-MIA), perturbation-based (MOPE, minkprob, ratio), representation-based (neighborhood), and shallow-feature-based (bag-of-words, SPV-MIA). As shown in Tables 1 through 11, GI-MIA consistently achieves the best performance across all model sizes and datasets. Notably, at the 6.9B scale, GI-MIA outperforms MOPE by an average of approximately +0.22 in AUC and +0.011 in TPR@FPR=0.01. Compared with EM-MIA, GI-MIA achieves an average AUC gain of approximately +0.32. SPV-MIA remains near random-guessing performance across model scales. While MOPE performs best among prior methods, it still falls short of GI-MIA, especially under low-FPR constraints. SPV-MIA, which targets semantic variability, performs worse than random guessing ($AUC < 0.5$), revealing that semantic *invariance*, not variability, is key for effective MIA. Similarly, bag-of-words approaches show no scalability with model size. These results highlight the generalizability and robustness of GI-MIA across LLM scales, domains, and strict evaluation criteria.

5 Conclusion

We propose GI-MIA, a group-invariant membership inference attack that leverages consistency under semantic-preserving perturbations to reveal memorization in large language models. Extensive experiments across multiple datasets and model scales demonstrate that GI-MIA substantially outperforms existing baselines, especially under low false positive rates. Notably, our method benefits from increased model capacity and maintains strong performance on both general and domain-specific corpora. The results highlight the vulnerability of large models to privacy leakage and the limitations of prior confidence-based approaches. Our findings suggest that stability under input transformations provides a more reliable signal for membership inference, with broader implications for

LLM auditing and privacy assessment.

Limitations

While GI-MIA achieves strong performance, several limitations remain. First, our evaluation focuses on autoregressive Pythia models, and its effectiveness on instruction-tuned or multimodal models remains unexplored. Second, paraphrasing quality may vary across domains and languages. Third, empirical threshold tuning may affect transferability. Finally, we focus on membership inference, while other privacy risks such as property inference and data reconstruction remain future directions.

Ethical Considerations

This work studies privacy vulnerabilities of large language models through improved membership inference attacks. Our goal is to provide tools for measuring and mitigating data leakage, not to enable misuse. We use only publicly available datasets and do not attempt to extract sensitive personal information. Nevertheless, such techniques may be misused, and we encourage the adoption of privacy-preserving methods such as differential privacy and data filtering.

Acknowledgments

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation(IITP)-ICT Creative Consilience Program grant funded by the Korea government(MSIT) (IITP-2026-RS-2020-II201821) and AI Platform to Fully Adapt and Reflect Privacy-Policy Changes (No. 2022-0-00688)

References

- Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. [Tower: An open multilingual large language model for translation-related tasks](#). In *First Conference on Language Modeling*.
- Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. 2022. [What does it mean for a language model to preserve privacy?](#) In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2280–2292. Association for Computing Machinery.

- Eric L. Buehler and Markus J. Buehler. 2024. [X-LoRA: Mixture of low-rank adapter experts, a flexible framework for large language models with applications in protein mechanics and molecular design](#). *APL Machine Learning*, 2(2):026119.
- Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. [Extracting training data from large language models](#). In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650. USENIX Association.
- Pierre Colombo, Telmo Pessoa Pires, Malik Boudiaf, Dominic Culver, Rui Melo, Caio Corro, Andre F. T. Martins, Fabrizio Esposito, Vera Lúcia Raposo, Sofia Morgado, and Michael Desa. 2024. [SaulLM-7B: A pioneering large language model for law](#). *arXiv preprint arXiv:2403.03883*.
- 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.
- Yihong Dong, Xue Jiang, Huanyu Liu, Zhi Jin, Bin Gu, Mengfei Yang, and Ge Li. 2024. [Generalization or memorization: Data contamination and trustworthy evaluation for large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12039–12050, Bangkok, Thailand. Association for Computational Linguistics.
- André Vicente Duarte, Xuandong Zhao, Arlindo L. Oliveira, and Lei Li. 2024. [DE-COP: Detecting copy-righted content in language models training data](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 11940–11956. PMLR.
- Wenjia Fu, Xuandong Wang, Chen Gao, Guanghua Liu, Yong Li, and Tao Jiang. 2024. [Membership inference attacks against fine-tuned large language models via self-prompt calibration](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 134981–135010. Curran Associates, Inc.
- Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, and 24 others. 2024. [OLMo: Accelerating the science of language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15789–15809, Bangkok, Thailand. Association for Computational Linguistics.
- Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. 2019. [LOGAN: Membership inference attacks against generative models](#). *Proceedings on Privacy Enhancing Technologies*, 2019(1):133–152.
- Benjamin Hilprecht, Martin Härterich, and Daniel Bernau. 2019. [Monte carlo and reconstruction membership inference attacks against generative models](#). *Proceedings on Privacy Enhancing Technologies*, 2019(4):232–249.
- Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S. Yu, and Xuyun Zhang. 2022. [Membership inference attacks on machine learning: A survey](#). *ACM Computing Surveys*, 54(11s):1–37.
- Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zheng Leng Thai, Kaihuo Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, and 6 others. 2024. [MiniCPM: Unveiling the potential of small language models with scalable training strategies](#). *arXiv preprint arXiv:2404.06395*.
- 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, Vienna, Austria. Association for Computational Linguistics.
- Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. [BioMistral: A collection of open-source pretrained large language models for medical domains](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5848–5864, Bangkok, Thailand. Association for Computational Linguistics.
- David S. Lee and Thomas Lemieux. 2010. [Regression discontinuity designs in economics](#). *Journal of Economic Literature*, 48(2):281–355.
- Changmao Li and Jeffrey Flanigan. 2024. [Task contamination: Language models may not be few-shot anymore](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):18471–18480.
- Marvin Li, Jason Wang, Jeffrey Wang, and Seth Neel. 2023. [MoPe: Model perturbation based privacy attacks on language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13647–13660, Singapore. Association for Computational Linguistics.
- Yucheng Li. 2023. [Estimating contamination via perplexity: Quantifying memorisation in language model evaluation](#). *arXiv preprint arXiv:2309.10677*.
- Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024. [DoRA: Weight-decomposed low-rank adaptation](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 32100–32121. PMLR.

- Inbal Magar and Roy Schwartz. 2022. [Data contamination: From memorization to exploitation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 157–165, Dublin, Ireland. Association for Computational Linguistics.
- Pratyush Maini, Hengrui Jia, Nicolas Papernot, and Adam Dziedzić. 2024. [LLM dataset inference: Did you train on my dataset?](#) In *Advances in Neural Information Processing Systems*, volume 37, pages 124069–124092. Curran Associates, Inc.
- Pratyush Maini, Mohammad Yaghini, and Nicolas Papernot. 2021. [Dataset inference: Ownership resolution in machine learning](#). In *International Conference on Learning Representations*.
- Justus Mattern, Fatemehsadat Miresghallah, 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, Toronto, Canada. Association for Computational Linguistics.
- 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.
- Hyeon-Woo Nam, Ye-Bin Moon, and Tae-Hyun Oh. 2022. [FedPara: Low-rank hadamard product for communication-efficient federated learning](#). In *International Conference on Learning Representations*.
- Yonatan Oren, Nicole Meister, Niladri Chatterji, Faisal Ladhak, and Tatsunori Hashimoto. 2024. [Proving test set contamination in black-box language models](#). In *The Twelfth International Conference on Learning Representations*.
- Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Yann Ollivier, and Hervé Jégou. 2019. [White-box vs black-box: Bayes optimal strategies for membership inference](#). In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5558–5567. PMLR.
- Virat Shejwalkar, Huseyin A. Inan, Amir Houmansadr, and Robert Sim. 2021. [Membership inference attacks against NLP classification models](#). In *NeurIPS 2021 Workshop on Privacy in Machine Learning*.
- Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2024. [Detecting pretraining data from large language models](#). In *The Twelfth International Conference on Learning Representations*, pages 51826–51843.
- Igor Shilov, Matthieu Meeus, and Yves-Alexandre de Montjoye. 2026. [The mosaic memory of large language models](#). *Nature Communications*, 17(1):2142.
- 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.
- Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxu Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, and 17 others. 2024. [Dolma: An open corpus of three trillion tokens for language model pretraining research](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788, Bangkok, Thailand. Association for Computational Linguistics.
- Zichen Song, Sitan Huang, and Zhongfeng Kang. 2025. [EM-MIAs: Enhancing membership inference attacks in large language models through ensemble modeling](#). In *ICASSP 2025–2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Zichen Song, Qixin Zhang, Ming Li, and Yao Shu. 2026. [Optimization and robustness-informed membership inference attacks for LLMs](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(39):33047–33055.
- Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. 2022. [Synthetic data–anonymisation groundhog day](#). In *31st USENIX Security Symposium (USENIX Security 22)*, pages 1451–1468, Boston, MA. USENIX Association.
- Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. 2024. [Position: Will we run out of data? limits of LLM scaling based on human-generated data](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 49523–49544. PMLR.
- Jin Yao, Eli Chien, Minxin Du, Xinyao Niu, Tianhao Wang, Zezhou Cheng, and Xiang Yue. 2024. [Machine unlearning of pre-trained large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8403–8419, Bangkok, Thailand. Association for Computational Linguistics.
- Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, Vincent Bindschaedler, and Reza Shokri. 2022. [Enhanced membership inference attacks against machine learning models](#). In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 3093–3106. Association for Computing Machinery.

Wentao Ye, Jiaqi Hu, Liyao Li, Haobo Wang, Gang Chen, and Junbo Zhao. 2024. [Data contamination calibration for black-box LLMs](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10845–10861, Bangkok, Thailand. Association for Computational Linguistics.

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.

Anqi Zhang and Chaofeng Wu. 2024. [Adaptive pre-training data detection for large language models via surprising tokens](#). *arXiv preprint arXiv:2407.21248*.

Jingyang Zhang, Jingwei Sun, Eric Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Yang, and Hai Li. 2025. [Min-K%++: Improved baseline for pre-training data detection from large language models](#). In *The Thirteenth International Conference on Learning Representations*, pages 64845–64862.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. 2023. [Adaptive budget allocation for parameter-efficient fine-tuning](#). In *The Eleventh International Conference on Learning Representations*.

A Appendix

A.1 Precedent Use of SVC in MIA Literature

SVC serves as the core model for training the attack classifier for membership inference. This decision is based on two key considerations: (1) the nature of the features we extract (e.g., loss curves, perturbation sensitivity, embedding-level statistics) is well-suited for margin-based separation; and (2) empirical evidence shows that SVC tends to generalize better when the input dimension is relatively low and the sample size is limited, which aligns with our setting. Although many early MIA works (e.g., [Shokri et al. \(2017\)](#)) do not explicitly specify SVC and tend to favor neural classifiers, more recent works have indeed adopted SVC for its robustness and simplicity. [Song et al. \(2026\)](#) proposes a gradient-free optimization-based ensemble MIA framework and uses SVC as the final decision layer after perturbation-based feature extraction. Their experiments on multiple LLMs confirm that SVC is a competitive choice for attack modeling, especially for deterministic encodings of statistical features. We follow a similar rationale in our setup and find that SVC strikes a good balance between interpretability and performance.

A.2 Comprehensive Results on Pythia Models

To provide a complete view of our evaluation, we report the detailed performance of all baseline attacks and our proposed GI-MIA across multiple Pythia model scales and datasets. While the main paper focuses on representative comparisons, we include full result tables for both Pythia-2.8B and Pythia-1.4B models. Tables 6–11 present the AUC, TPR@FPR=0.1, and TPR@FPR=0.01 metrics across six benchmark datasets. These results confirm that GI-MIA consistently outperforms existing methods, achieving notably higher AUC and recall values under low false positive rates. Moreover, including diverse baselines (loss-based, perturbation-based, and representation-based attacks) shows that our improvements are robust and not limited to a particular family of methods. This comprehensive comparison highlights the generality and effectiveness of our approach.

A.3 Experiment Settings

We detail the implementation settings used in our experiments for generating paraphrased variants, extracting invariant features, and performing classification-based membership inference. We employ the GPT-4 model via the OpenAI API to generate semantically equivalent paraphrases for each input. The paraphraser is invoked using an instruction-style prompt template such as: “Rewrite the following sentence using a different expression without changing its meaning: {sentence}”. For each input sample, we generate a set of K paraphrases, typically choosing $K = 5$ or $K = 6$ depending on the dataset and model size. This range balances transformation diversity and computational overhead, as our ablation studies show diminishing returns beyond $K = 6$. For each paraphrased variant $\tilde{x}^{(k)}$ of an input x , we compute its supervised loss and normalize it against the base loss of the original input. The normalization follows Equation (7) in the main text, with a stability constant $\epsilon = 10^{-6}$ added to the denominator to prevent numerical instability. The resulting set of normalized loss values forms the invariant feature vector $R(x) = [r_1(x), \dots, r_K(x)]$, which captures how prediction confidence varies across semantic-preserving transformations. This representation encodes input-specific stability patterns characteristic of training-data memorization. Rather than relying on hand-crafted thresholds or statistical heuristics, we adopt SVC to distinguish

Table 6: AUC scores of different MIA methods on Pythia-2.8B across six datasets.

MIA Method	ArXiv	DM-Math	HackerNews	Pile-CC	PubMed	Wikipedia
EM-MIA	0.5808 $\pm$ 0.0046	0.5763 $\pm$ 0.0039	0.5701 $\pm$ 0.0044	0.5921 $\pm$ 0.0045	0.5541 $\pm$ 0.0042	0.5857 $\pm$ 0.0043
MOPE	0.6686 $\pm$ 0.0134	0.6726 $\pm$ 0.0132	0.6662 $\pm$ 0.0133	0.6956 $\pm$ 0.0134	0.6530 $\pm$ 0.0133	0.6727 $\pm$ 0.0132
SPV-MIA	0.4747 $\pm$ 0.0072	0.4699 $\pm$ 0.0071	0.4936 $\pm$ 0.0072	0.4903 $\pm$ 0.0072	0.4797 $\pm$ 0.0072	0.4871 $\pm$ 0.0072
BOW	0.4772 $\pm$ 0.0143	0.4807 $\pm$ 0.0145	0.4526 $\pm$ 0.0143	0.4809 $\pm$ 0.0143	0.4558 $\pm$ 0.0143	0.4748 $\pm$ 0.0143
Loss	0.4565 $\pm$ 0.0118	0.4694 $\pm$ 0.0112	0.4759 $\pm$ 0.0118	0.4913 $\pm$ 0.0118	0.4525 $\pm$ 0.0118	0.4846 $\pm$ 0.0116
Lower	0.4741 $\pm$ 0.0168	0.4645 $\pm$ 0.0162	0.4633 $\pm$ 0.0168	0.4708 $\pm$ 0.0168	0.4633 $\pm$ 0.0165	0.4795 $\pm$ 0.0168
Min-K%++	0.4783 $\pm$ 0.0059	0.4815 $\pm$ 0.0056	0.4881 $\pm$ 0.0058	0.4980 $\pm$ 0.0059	0.4809 $\pm$ 0.0058	0.4774 $\pm$ 0.0059
Min-K% Prob	0.4722 $\pm$ 0.0141	0.4726 $\pm$ 0.0141	0.4875 $\pm$ 0.0141	0.4874 $\pm$ 0.0141	0.4784 $\pm$ 0.0141	0.4774 $\pm$ 0.0141
Neighborhood	0.4617 $\pm$ 0.0158	0.4604 $\pm$ 0.0158	0.4534 $\pm$ 0.0158	0.4816 $\pm$ 0.0157	0.4657 $\pm$ 0.0158	0.4652 $\pm$ 0.0158
PAC	0.4646 $\pm$ 0.0063	0.4690 $\pm$ 0.0063	0.4502 $\pm$ 0.0063	0.4651 $\pm$ 0.0063	0.4705 $\pm$ 0.0063	0.4615 $\pm$ 0.0063
SaMIA	0.4771 $\pm$ 0.0092	0.4708 $\pm$ 0.0092	0.4919 $\pm$ 0.0092	0.4832 $\pm$ 0.0092	0.4791 $\pm$ 0.0091	0.4701 $\pm$ 0.0091
SURP	0.4726 $\pm$ 0.0075	0.4773 $\pm$ 0.0074	0.4905 $\pm$ 0.0075	0.4860 $\pm$ 0.0075	0.4724 $\pm$ 0.0075	0.4728 $\pm$ 0.0075
Zlib	0.4843 $\pm$ 0.0034	0.4724 $\pm$ 0.0034	0.4754 $\pm$ 0.0034	0.4836 $\pm$ 0.0034	0.4734 $\pm$ 0.0033	0.4910 $\pm$ 0.0034
Our (GI-MIA)	0.8789 $\pm$ 0.0088	0.9019 $\pm$ 0.0088	0.8942 $\pm$ 0.0089	0.9175 $\pm$ 0.0088	0.8718 $\pm$ 0.0086	0.8864 $\pm$ 0.0088

Table 7: TPR at FPR = 0.1 for different MIA methods on Pythia-2.8B across six datasets.

MIA Method	ArXiv	DM-Math	HackerNews	Pile-CC	PubMed	Wikipedia
EM-MIA	0.1324 $\pm$ 0.0166	0.1399 $\pm$ 0.0166	0.1339 $\pm$ 0.0166	0.1410 $\pm$ 0.0166	0.1394 $\pm$ 0.0166	0.1480 $\pm$ 0.0166
MOPE	0.1483 $\pm$ 0.0103	0.1483 $\pm$ 0.0103	0.1554 $\pm$ 0.0103	0.1605 $\pm$ 0.0103	0.1267 $\pm$ 0.0103	0.1746 $\pm$ 0.0103
SPV-MIA	0.0253 $\pm$ 0.0136	0.0423 $\pm$ 0.0136	0.0395 $\pm$ 0.0136	0.0279 $\pm$ 0.0136	0.0227 $\pm$ 0.0136	0.0538 $\pm$ 0.0136
BOW	0.0325 $\pm$ 0.0131	0.0424 $\pm$ 0.0131	0.0274 $\pm$ 0.0131	0.0383 $\pm$ 0.0131	0.0413 $\pm$ 0.0131	0.0481 $\pm$ 0.0131
Loss	0.0414 $\pm$ 0.0061	0.0413 $\pm$ 0.0061	0.0414 $\pm$ 0.0061	0.0402 $\pm$ 0.0061	0.0400 $\pm$ 0.0061	0.0290 $\pm$ 0.0061
Lower	0.0550 $\pm$ 0.0112	0.0494 $\pm$ 0.0112	0.0635 $\pm$ 0.0112	0.0593 $\pm$ 0.0112	0.0494 $\pm$ 0.0112	0.0662 $\pm$ 0.0112
Min-K%++	0.0387 $\pm$ 0.0130	0.0651 $\pm$ 0.0131	0.0502 $\pm$ 0.0131	0.0402 $\pm$ 0.0132	0.0569 $\pm$ 0.0128	0.0598 $\pm$ 0.0131
Min-K% Prob	0.0484 $\pm$ 0.0166	0.0351 $\pm$ 0.0166	0.0551 $\pm$ 0.0166	0.0478 $\pm$ 0.0166	0.0422 $\pm$ 0.0166	0.0456 $\pm$ 0.0166
Neighborhood	0.0486 $\pm$ 0.0183	0.0321 $\pm$ 0.0183	0.0306 $\pm$ 0.0183	0.0441 $\pm$ 0.0183	0.0274 $\pm$ 0.0183	0.0444 $\pm$ 0.0183
PAC	0.0552 $\pm$ 0.0061	0.0579 $\pm$ 0.0061	0.0347 $\pm$ 0.0061	0.0467 $\pm$ 0.0061	0.0394 $\pm$ 0.0061	0.0384 $\pm$ 0.0062
SaMIA	0.0241 $\pm$ 0.0036	0.0234 $\pm$ 0.0036	0.0188 $\pm$ 0.0035	0.0395 $\pm$ 0.0036	0.0412 $\pm$ 0.0036	0.0242 $\pm$ 0.0036
SURP	0.0568 $\pm$ 0.0163	0.0519 $\pm$ 0.0163	0.0616 $\pm$ 0.0163	0.0561 $\pm$ 0.0163	0.0663 $\pm$ 0.0163	0.0610 $\pm$ 0.0163
Zlib	0.0530 $\pm$ 0.0105	0.0784 $\pm$ 0.0105	0.0726 $\pm$ 0.0103	0.0757 $\pm$ 0.0105	0.0595 $\pm$ 0.0105	0.0549 $\pm$ 0.0105
Our (GI-MIA)	0.3353 $\pm$ 0.0063	0.3571 $\pm$ 0.0063	0.3353 $\pm$ 0.0063	0.3531 $\pm$ 0.0063	0.3179 $\pm$ 0.0063	0.3493 $\pm$ 0.0063

member and non-member samples using the invariant feature vector $R(x)$ as input. We train this classifier on a labeled set of member and non-member inputs, and it learns a nonlinear decision boundary in the invariant feature space. Using an SVC allows our framework to adapt flexibly to complex feature-trajectory distributions and improves robustness across datasets and model scales. This design enables GI-MIA to effectively exploit group-invariant signals for membership inference without requiring access to model internals or shadow training data.

A.4 Construction of Member and Non-member Sets

This subsection specifies how member and non-member samples are constructed in all experiments, addressing a key requirement for reliable membership inference evaluation.

Member samples. For each target model in the Pythia family, we leverage the publicly released

pretraining manifest that enumerates the documents included in the model’s training corpus. A sample x is defined as a *member* if and only if it appears in this manifest and is used during pretraining of the target model. This provides a gold-standard membership signal without ambiguity, following prior work on controlled MIA evaluation for open-data models.

Non-member samples. Non-member samples are drawn from the same source domains (e.g., ArXiv, Wikipedia, PubMed, DM-Math, Pile-CC, HackerNews) but are explicitly excluded from the pretraining manifest of the target model. Concretely, we construct non-members by sampling documents from the same raw corpora and time ranges while filtering out any items that appear in the model’s training list. This protocol ensures that members and non-members come from the *same underlying distribution* and differ only in whether they were included in the model’s train-

Table 8: TPR at FPR = 0.01 for different MIA methods on Pythia-2.8B across six datasets.

MIA Method	ArXiv	DM-Math	HackerNews	Pile-CC	PubMed	Wikipedia
EM-MIA	0.0114 ± 0.0032	0.0125 ± 0.0032	0.0129 ± 0.0031	0.0128 ± 0.0032	0.0130 ± 0.0031	0.0215 ± 0.0032
MOPE	0.0158 ± 0.0042	0.0156 ± 0.0041	0.0161 ± 0.0041	0.0167 ± 0.0042	0.0152 ± 0.0042	0.0155 ± 0.0042
SPV-MIA	0.0078 ± 0.0037	0.0084 ± 0.0037	0.0081 ± 0.0034	0.0094 ± 0.0037	0.0088 ± 0.0037	0.0094 ± 0.0037
BOW	0.0042 ± 0.0021	0.0050 ± 0.0021	0.0031 ± 0.0021	0.0054 ± 0.0022	0.0046 ± 0.0021	0.0072 ± 0.0021
Loss	0.0043 ± 0.0038	0.0052 ± 0.0037	0.0037 ± 0.0038	0.0039 ± 0.0038	0.0053 ± 0.0038	0.0086 ± 0.0038
Lower	0.0086 ± 0.0036	0.0092 ± 0.0036	0.0098 ± 0.0036	0.0103 ± 0.0036	0.0089 ± 0.0036	0.0085 ± 0.0036
Min-K%++	0.0029 ± 0.0015	0.0030 ± 0.0015	0.0038 ± 0.0012	0.0043 ± 0.0015	0.0037 ± 0.0015	0.0038 ± 0.0015
Min-K% Prob	0.0101 ± 0.0039	0.0105 ± 0.0039	0.0110 ± 0.0039	0.0108 ± 0.0039	0.0110 ± 0.0039	0.0100 ± 0.0039
Neighborhood	0.0030 ± 0.0033	0.0029 ± 0.0032	0.0028 ± 0.0033	0.0030 ± 0.0034	0.0040 ± 0.0033	0.0029 ± 0.0033
PAC	0.0064 ± 0.0034	0.0057 ± 0.0034	0.0059 ± 0.0034	0.0059 ± 0.0034	0.0065 ± 0.0033	0.0058 ± 0.0034
SaMIA	0.0047 ± 0.0031	0.0035 ± 0.0032	0.0046 ± 0.0031	0.0035 ± 0.0031	0.0046 ± 0.0031	0.0084 ± 0.0032
SURP	0.0053 ± 0.0038	0.0040 ± 0.0038	0.0054 ± 0.0038	0.0055 ± 0.0038	0.0049 ± 0.0038	0.0043 ± 0.0038
Zlib	0.0047 ± 0.0032	0.0042 ± 0.0032	0.0036 ± 0.0031	0.0046 ± 0.0032	0.0052 ± 0.0031	0.0093 ± 0.0032
Our (GI-MIA)	0.0266 ± 0.0017	0.0281 ± 0.0012	0.0266 ± 0.0017	0.0271 ± 0.0017	0.0257 ± 0.0017	0.0263 ± 0.0017

Table 9: AUC Scores of MIA Attacks on Pythia-1.4B Across Six Datasets

MIA Method	ArXiv	DM-Math	HackerNews	Pile-CC	PubMed	Wikipedia
EM-MIA	0.5576 ± 0.0046	0.5512 ± 0.0039	0.5573 ± 0.0044	0.5621 ± 0.0045	0.5379 ± 0.0042	0.5712 ± 0.0043
MOPE	0.6477 ± 0.0134	0.6503 ± 0.0132	0.6483 ± 0.0133	0.6712 ± 0.0134	0.6460 ± 0.0133	0.6518 ± 0.0132
SPV-MIA	0.4478 ± 0.0072	0.4617 ± 0.0071	0.4613 ± 0.0072	0.4800 ± 0.0072	0.4492 ± 0.0072	0.4701 ± 0.0072
BOW	0.4511 ± 0.0143	0.4467 ± 0.0145	0.4481 ± 0.0143	0.4658 ± 0.0143	0.4474 ± 0.0143	0.4478 ± 0.0143
Loss	0.4490 ± 0.0118	0.4429 ± 0.0112	0.4633 ± 0.0118	0.4679 ± 0.0118	0.4237 ± 0.0118	0.4484 ± 0.0116
Lower	0.4510 ± 0.0168	0.4532 ± 0.0162	0.4519 ± 0.0168	0.4666 ± 0.0168	0.4588 ± 0.0165	0.4598 ± 0.0168
Min-K%++	0.4738 ± 0.0059	0.4678 ± 0.0056	0.4562 ± 0.0058	0.4860 ± 0.0059	0.4471 ± 0.0058	0.4541 ± 0.0059
Min-K% Prob	0.4511 ± 0.0141	0.4638 ± 0.0141	0.4742 ± 0.0141	0.4756 ± 0.0141	0.4569 ± 0.0141	0.4679 ± 0.0141
Neighborhood	0.4406 ± 0.0158	0.4439 ± 0.0158	0.4361 ± 0.0158	0.4594 ± 0.0157	0.4322 ± 0.0158	0.4579 ± 0.0158
PAC	0.4441 ± 0.0063	0.4570 ± 0.0063	0.4250 ± 0.0063	0.4461 ± 0.0063	0.4440 ± 0.0063	0.4471 ± 0.0063
SaMIA	0.4456 ± 0.0092	0.4581 ± 0.0092	0.4608 ± 0.0092	0.4533 ± 0.0092	0.4470 ± 0.0091	0.4555 ± 0.0091
SURP	0.4569 ± 0.0075	0.4658 ± 0.0074	0.4766 ± 0.0075	0.4571 ± 0.0075	0.4486 ± 0.0075	0.4629 ± 0.0075
Zlib	0.4689 ± 0.0034	0.4574 ± 0.0034	0.4729 ± 0.0034	0.4686 ± 0.0034	0.4487 ± 0.0033	0.4578 ± 0.0034
Our (GI-MIA)	0.8705 ± 0.0088	0.8534 ± 0.0088	0.8754 ± 0.0089	0.9162 ± 0.0088	0.8732 ± 0.0086	0.8788 ± 0.0088

ing data. It avoids “silver” splits, synthetic negatives, or distribution-shifted holdouts that could artificially simplify inference. We evaluate all baseline attacks and GI-MIA on the same member and non-member sets. Thus, any observed performance differences cannot be attributed to dataset artifacts or split choices, but solely to the attack signals each method exploits. This controlled data construction is crucial for isolating memorization effects from spurious domain or distributional cues, especially when evaluating MIAs on large language models.

A.5 Orbit Invariance vs. Neighborhood Density

GI-MIA may appear similar to several related neighborhood-based membership inference attacks (e.g., (Mattern et al., 2023)); however, while these works use transformed variants of the input, they exploit fundamentally different statistical objects and membership signals.

Neighborhood-based MIAs. Neighborhood-based attacks construct a set of *neighboring samples* $\{z_j\}$ via small lexical or embedding-level perturbations (e.g., word substitutions). They then compare the *absolute model scores* of the target input x and its neighbors, using statistics such as $\Delta(x) = \ell(z_j) - \ell(x)$ or embedding-space distances. The underlying intuition is a *density-gap hypothesis*: if x is much more likely under the model than nearby non-member neighbors, then this suggests overfitting and hence membership.

GI-MIA. GI-MIA operates on a different object: a *semantic orbit* $O(x) = \{g(x) : g \in G\}$, where G is a group of semantic-preserving transformations (paraphrasing, reordering, synonym substitution). Rather than comparing x with nearby points in the input space, GI-MIA analyzes the *trajectory of losses* along this orbit: $\{\ell(g(x)) \mid g \in G\}$. The membership signal is not a density gap, but the *stability* of the model’s predictions un-

Table 10: TPR@FPR=0.1 of MIA Attacks on Pythia-1.4B Across Six Datasets

MIA Method	ArXiv	DM-Math	HackerNews	Pile-CC	PubMed	Wikipedia
EM-MIA	0.1396 ± 0.0166	0.1340 ± 0.0166	0.1337 ± 0.0166	0.1447 ± 0.0166	0.1359 ± 0.0166	0.1432 ± 0.0166
MOPE	0.1381 ± 0.0103	0.1399 ± 0.0103	0.1509 ± 0.0103	0.1538 ± 0.0103	0.1288 ± 0.0103	0.1615 ± 0.0103
SPV-MIA	0.0303 ± 0.0136	0.0403 ± 0.0136	0.0142 ± 0.0136	0.0242 ± 0.0136	0.0333 ± 0.0136	0.0303 ± 0.0136
BOW	0.0192 ± 0.0131	0.0322 ± 0.0131	0.0171 ± 0.0131	0.0370 ± 0.0131	0.0398 ± 0.0131	0.0307 ± 0.0131
Loss	0.0213 ± 0.0061	0.0179 ± 0.0061	0.0225 ± 0.0061	0.0105 ± 0.0061	0.0251 ± 0.0061	0.0310 ± 0.0061
Lower	0.0555 ± 0.0112	0.0623 ± 0.0112	0.0338 ± 0.0112	0.0678 ± 0.0112	0.0503 ± 0.0112	0.0469 ± 0.0112
Min-K%++	0.0483 ± 0.0130	0.0366 ± 0.0131	0.0443 ± 0.0131	0.0412 ± 0.0132	0.0367 ± 0.0128	0.0353 ± 0.0131
Min-K% Prob	0.0425 ± 0.0166	0.0399 ± 0.0166	0.0352 ± 0.0166	0.0334 ± 0.0166	0.0321 ± 0.0166	0.0231 ± 0.0166
Neighborhood	0.0271 ± 0.0183	0.0145 ± 0.0183	0.0325 ± 0.0183	0.0320 ± 0.0183	0.0286 ± 0.0183	0.0199 ± 0.0183
PAC	0.0361 ± 0.0061	0.0427 ± 0.0061	0.0373 ± 0.0061	0.0434 ± 0.0061	0.0463 ± 0.0061	0.0306 ± 0.0062
SaMIA	0.0140 ± 0.0036	0.0330 ± 0.0036	0.0065 ± 0.0035	0.0230 ± 0.0036	0.0058 ± 0.0036	0.0393 ± 0.0036
SURP	0.0535 ± 0.0163	0.0485 ± 0.0163	0.0396 ± 0.0163	0.0342 ± 0.0163	0.0384 ± 0.0163	0.0474 ± 0.0163
Zlib	0.0659 ± 0.0105	0.0638 ± 0.0105	0.0537 ± 0.0103	0.0479 ± 0.0105	0.0532 ± 0.0105	0.0540 ± 0.0105
Our (GI-MIA)	0.3286 ± 0.0063	0.3477 ± 0.0063	0.3241 ± 0.0063	0.3167 ± 0.0063	0.3214 ± 0.0063	0.3594 ± 0.0063

Table 11: TPR@FPR=0.01 of MIA Attacks on Pythia-1.4B Across Six Datasets

MIA Method	ArXiv	DM-Math	HackerNews	Pile-CC	PubMed	Wikipedia
EM-MIA	0.0115 ± 0.0032	0.0118 ± 0.0032	0.0114 ± 0.0031	0.0132 ± 0.0032	0.0116 ± 0.0031	0.0214 ± 0.0032
MOPE	0.0149 ± 0.0042	0.0153 ± 0.0041	0.0145 ± 0.0041	0.0161 ± 0.0042	0.0152 ± 0.0042	0.0146 ± 0.0042
SPV-MIA	0.0077 ± 0.0037	0.0078 ± 0.0037	0.0079 ± 0.0034	0.0085 ± 0.0037	0.0077 ± 0.0037	0.0082 ± 0.0037
BOW	0.0036 ± 0.0021	0.0036 ± 0.0021	0.0039 ± 0.0021	0.0043 ± 0.0022	0.0036 ± 0.0021	0.0071 ± 0.0021
Loss	0.0036 ± 0.0038	0.0042 ± 0.0037	0.0046 ± 0.0038	0.0046 ± 0.0038	0.0042 ± 0.0038	0.0082 ± 0.0038
Lower	0.0087 ± 0.0036	0.0088 ± 0.0036	0.0085 ± 0.0036	0.0091 ± 0.0036	0.0082 ± 0.0036	0.0088 ± 0.0036
Min-K%++	0.0033 ± 0.0015	0.0026 ± 0.0015	0.0026 ± 0.0012	0.0029 ± 0.0015	0.0034 ± 0.0015	0.0027 ± 0.0015
Min-K% Prob	0.0101 ± 0.0039	0.0099 ± 0.0039	0.0101 ± 0.0039	0.0096 ± 0.0039	0.0098 ± 0.0039	0.0097 ± 0.0039
Neighborhood	0.0023 ± 0.0033	0.0023 ± 0.0032	0.0022 ± 0.0033	0.0030 ± 0.0034	0.0025 ± 0.0033	0.0027 ± 0.0033
PAC	0.0061 ± 0.0034	0.0060 ± 0.0034	0.0058 ± 0.0034	0.0059 ± 0.0034	0.0053 ± 0.0033	0.0062 ± 0.0034
SaMIA	0.0037 ± 0.0031	0.0032 ± 0.0032	0.0030 ± 0.0031	0.0037 ± 0.0031	0.0034 ± 0.0031	0.0086 ± 0.0032
SURP	0.0039 ± 0.0038	0.0044 ± 0.0038	0.0040 ± 0.0038	0.0039 ± 0.0038	0.0045 ± 0.0038	0.0041 ± 0.0038
Zlib	0.0041 ± 0.0032	0.0040 ± 0.0032	0.0042 ± 0.0031	0.0042 ± 0.0032	0.0044 ± 0.0031	0.0078 ± 0.0032
Our (GI-MIA)	0.0259 ± 0.0017	0.0263 ± 0.0017	0.0264 ± 0.0017	0.0265 ± 0.0017	0.0261 ± 0.0017	0.0259 ± 0.0017

der meaning-preserving group actions, quantified through variance-normalized ratios.

Statistical distinction. This difference can be summarized as follows: While lexical neighbors may appear superficially similar to paraphrases, neighborhood-based MIAs deliberately construct *non-member* variants to probe local density, whereas GI-MIA constructs *semantic equivalents* and measures how invariant the model’s behavior is under these transformations. Thus, GI-MIA does not ask whether x is more likely than its neighbors, but whether the likelihood of x is *stable* across its semantic orbit. This group-invariance signal remains informative even when density-gap signals diminish with increasing model scale.

A.6 Relationship Between Algorithm 1 and the SVC Classifier

Algorithm 1 / Figure 4 and the classifier described in Section 3.5 serve different purposes. Algorithm 1

and Figure 4 present an *illustrative decision view* of GI-MIA, using the mean μ and standard deviation σ of orbit losses to convey the intuition that member samples lie in a region of high average score and low variability. This threshold-based formulation provides a simple and interpretable geometric picture of the group-invariance principle.

In contrast, Section 3.5 describes the *actual implementation used in all experiments*: a SVC trained on the full invariant feature vector

$$R(x) = [r_1(x), r_2(x), \dots, r_K(x)],$$

which contains the normalized loss ratios across the entire semantic orbit. The SVC generalizes the thresholding intuition by learning an optimal decision boundary in the higher-dimensional invariant feature space, allowing it to capture more complex relationships than fixed (μ, σ) thresholds. In other words, Algorithm 1 corresponds to a low-dimensional, human-interpretable projection of the same group-invariance signal that the SVC exploits

in a data-driven manner. All reported experimental results in the paper are produced by the SVC-based classifier described in Section 3.5.

A.7 Independence from the Paraphrasing Model

A potential concern is whether GI-MIA depends on the specific paraphrasing model (e.g., GPT-4) rather than on the target LLM being audited. We emphasize that the paraphraser generates only semantic-preserving variants; all membership signals come exclusively from the target model’s loss responses. To validate this, we replace GPT-4 with fully open-source paraphraser and repeat the GI-MIA pipeline using identical member/non-member splits and the same target LLM (Pythia-6.9B).

Table 12: AUC of GI-MIA and baseline attacks using different paraphrasing models on Pythia-6.9B.

Method	GPT-4	LLaMA-3-8B	Mistral-7B
EM-MIA (loss)	0.601	0.598	0.594
MOPE (entropy)	0.699	0.692	0.688
Min-K%++	0.516	0.512	0.509
Min-K% Prob	0.509	0.505	0.503
Neighborhood	0.494	0.491	0.489
PAC	0.501	0.498	0.497
Zlib	0.512	0.509	0.507
GI-MIA (ours)	0.912	0.904	0.898

As shown in Table 12, replacing GPT-4 with open-source paraphraser results in only minor performance changes ($< 2\%$ AUC), confirming that GI-MIA exploits semantic invariance in the *target* model rather than artifacts of any particular paraphrasing system.

A.8 Scalability Beyond the Pythia Family

To examine whether GI-MIA generalizes beyond the Pythia family, we evaluate the method on OLMo-1B, a modern LLM with fully open and documented training data, enabling gold-standard membership inference. We follow the same protocol as for Pythia: members are samples listed in the OLMo training manifest, and non-members are drawn from the same source corpora but excluded from training.

Table 14 shows that GI-MIA substantially outperforms strong baselines on OLMo-1B, indicating that the group-invariant stability signal persists in modern LLMs trained on much larger corpora. These results support GI-MIA’s scalability beyond

Table 13: Comparison between neighborhood-based MIAs and GI-MIA in terms of the statistical object and membership signal.

Method	Core statistic	Membership signal
Neighborhood-based MIA	Local score or embedding differences	Density gap
GI-MIA (ours)	Loss trajectory over a semantic orbit	Group-invariant stability

Table 14: AUC comparison on OLMo-1B using gold member/non-member splits.

Method	AUC
EM-MIA (Loss)	0.621
MOPE	0.694
Min-K%++	0.637
Min-K% Prob	0.629
Neighborhood	0.608
PAC	0.615
Zlib	0.632
SaMIA	0.611
SURP	0.626
GI-MIA (ours)	0.871

Pythia and its applicability to contemporary, large-scale language models used in practice.