The Hidden Cost of Modeling $P(X)$: Membership Inference Attacks in Generative Text Classifiers

Owais Makroo*
Department of Mathematics
Indian Institute of Technology Kharagapur
makroo.owais@kgpian.iitkgp.ac.in

Siva Rajesh Kasa*
Amazon.com Inc.
kasasiva@amazon.com

Sumegh Roychowdhury*
Amazon.com Inc.
sumegr@amazon.com

Karan Gupta*
Amazon.com Inc.
karaniis@amazon.com

Nikhil Pattisapu*
Amazon.com Inc.
npattisa@amazon.com

Santhosh Kasa
Amazon.com Inc.
santkasa@amazon.com

Sumit Negi
Amazon.com Inc.
suminegi@amazon.com

Abstract

Membership Inference Attacks (MIAs) pose a critical privacy threat by enabling adversaries to determine whether a specific sample was included in a model's training dataset. Despite extensive research on MIAs, systematic comparisons between generative and discriminative classifiers remain limited. This work addresses this gap by first providing theoretical motivation for why generative classifiers exhibit heightened susceptibility to MIAs, then validating these insights through comprehensive empirical evaluation. Our study encompasses discriminative, generative, and pseudo-generative text classifiers across varying training data volumes, evaluated on five benchmark datasets. Employing a diverse array of MIA strategies, we consistently demonstrate that fully generative classifiers which explicitly model the joint likelihood $P(X, Y)$ are most vulnerable to membership leakage. Furthermore, we observe that the canonical inference approach commonly used in generative classifiers significantly amplifies this privacy risk. These findings reveal a fundamental utility-privacy trade-off inherent in classifier design, underscoring the critical need for caution when deploying generative classifiers in privacy-sensitive applications. Our results motivate future research directions in developing privacy-preserving generative classifiers that can maintain utility while mitigating membership inference vulnerabilities.

1 Introduction and Related work

Text Classification (TC) is a fundamental task in Natural Language Processing (NLP), serving as the backbone for numerous applications including sentiment analysis, topic detection, intent classification, and document categorization (Yogatama et al., 2017; Castagnos et al., 2022; Roychowdhury et al., 2024; Kasa et al., 2024; Pattisapu et al., 2025). As machine learning models have become increasingly sophisticated and widely deployed, concerns about their privacy implications have grown substantially. One of the most critical privacy vulnerabilities is the **Membership Inference Attack** (MIA), where

*Equal Contribution

Workshop on Regulatable ML at the 39th Conference on Neural Information Processing Systems (NeurIPS 2025).

an adversary attempts to determine whether a specific data point was included in a model's training set (Shokri et al., 2017). MIAs represent a fundamental threat to data privacy by exploiting differential model behaviors on training versus non-training data to infer membership in the training set (Shokri et al., 2017; Carlini et al., 2019). The implications are particularly severe for sensitive personal data, potentially violating privacy expectations and regulatory requirements. Recent surveys have highlighted the growing sophistication of these attacks (Amit et al., 2024; Feng et al., 2025).

Predominant Focus on Discriminative Models. The majority of MIA research has concentrated on discriminative models like BERT (Devlin et al., 2019), which directly model $P(Y|X)$ and learn decision boundaries without explicitly modeling data distributions (Zheng et al., 2023; Kasa et al., 2025). Studies have revealed how factors such as overfitting, model capacity, and training data size influence attack success rates (Amit et al., 2024). Despite this discriminative focus, there has been renewed interest in generative classifiers for text classification (Li et al., 2025; Kasa et al., 2025). Unlike discriminative models, generative classifiers explicitly model the joint distribution $P(X, Y) = P(X|Y)P(Y)$, offering compelling advantages: superior performance in low-data regimes (Kasa et al., 2025; Yogatama et al., 2017), reduced susceptibility to spurious correlations (Li et al., 2025), and principled uncertainty estimates via Bayes' rule (Bouguila, 2011). The renaissance of generative classifiers has been particularly bolstered through scalable model architectures including autoregressive models (Radford et al., 2018), discrete diffusion models (Lou et al., 2024), and masked language models used generatively (Devlin et al., 2019; Wang and Cho, 2019a).

However, the very characteristics that make generative classifiers attractive explicit modeling of data distributions and superior performance with limited data raise important privacy questions Kasa et al. (2025). While MIAs have been extensively studied for discriminative models, a significant gap exists in understanding how different classification paradigms compare in their vulnerability to such attacks. In this work, we present the first large-scale, systematic analysis of the vulnerability of transformer-based classifiers to MIAs across a spectrum of modeling paradigms. Following Kasa et al. (2025), we consider three broad categories: (1) **discriminative models** (e.g., BERT), which model the conditional distribution $P(Y|X)$; (2) **fully generative models** that explicitly model $P(X, Y)$, such as autoregressive or discrete diffusion models; and (3) **pseudo-generative models**, such as MLMs, and pseudo-autoregressive models, where the label is appended at the end of the input sequence.

Contributions. Our work makes three key contributions to understanding privacy vulnerabilities in generative text classification:

1. **First systematic MIA analysis across classification paradigms:** We provide comprehensive theoretical and empirical analysis of MIA vulnerability across discriminative, generative, and pseudo-generative text classifiers. Our results reveal that by virtue of modeling $P(X)$, generative classifiers inherently expose themselves to heightened privacy risks compared to discriminative classifiers that only learn $P(Y|X)$. This vulnerability is particularly pronounced in discrete diffusion models.
2. **Analysis of privacy-utility trade-offs:** We demonstrate complex dynamics between MIA vulnerability and training data volume across different architectures. Our experiments reveal that vulnerability patterns vary with dataset size, and different factorizations of $P(X, Y)$ lead to distinct privacy leakage patterns. Through this analysis, we identify pseudo-generative models as a potential privacy-preserving alternative.
3. **Practical guidance for privacy-aware deployment:** Our findings provide actionable insights through: (a) comprehensive evaluation of attack strategies ranging from simple threshold-based to sophisticated machine learning approaches, (b) quantification of privacy risks through statistical divergence measures, and (c) clear recommendations for mitigating vulnerabilities in real-world deployments.

2 Related Works and Background

Generative vs. Discriminative Classifiers: Historical Foundations and Evolution. Efron (1975) established foundational theoretical groundwork demonstrating logistic regression's higher efficiency compared to normal discriminant analysis under certain distributional assumptions. Ng and Jordan (2001) provided the seminal analysis showing that while discriminative classifiers achieve lower asymptotic error rates, generative classifiers converge more rapidly with smaller training sets—a

fundamental sample efficiency versus asymptotic performance trade-off. Liang and Jordan (2008) extended these foundations with comprehensive asymptotic analyses revealing that relative performance depends critically on model assumption correctness and data availability. Neural networks brought new perspectives through Raina et al. (2003)'s hybrid approaches combining both paradigms' strengths, while Li et al. (2019) demonstrated generative classifiers' superior robustness to adversarial attacks in neural network settings. Zheng et al. (2023) provided comprehensive theoretical and empirical revisiting with large-scale experiments across multiple domains, developing novel frameworks quantifying bias-variance trade-offs and revealing that generative models achieve better calibration and uncertainty quantification despite higher asymptotic error rates.

The application of generative and discriminative approaches to text classification has evolved significantly with the transformer era witnessing a remarkable resurgence of generative approaches. Early work with recurrent neural networks by Yogatama et al. (2017) demonstrated that generative RNN classifiers, while exhibiting higher asymptotic error rates than discriminative counterparts, showed superior robustness to distribution shifts and faster convergence, echoing classical patterns identified by Ng and Jordan (2001). Modern generative classifiers leverage sophisticated architectures including autoregressive language models (Radford et al., 2018), discrete diffusion models (Lou et al., 2024), and masked language models used generatively (Wang and Cho, 2019a), providing new empirical evidence for advantages in text classification, particularly in low-resource settings where generative classifiers consistently demonstrate superior sample efficiency (Kasa et al., 2025). Jaini et al. (2024) demonstrated that generative classifiers exhibit superior robustness properties, including reduced susceptibility to adversarial perturbations and improved calibration of uncertainty estimates, while Li et al. (2025) showed that generative classifiers naturally avoid shortcut learning by explicitly modeling the full input distribution $P(X, Y) = P(Y) \times P(X|Y)$ rather than merely learning discriminative features $P(Y|X)$. The explicit modeling of class-conditional distributions $P(X|Y)$ enables generative classifiers to provide richer interpretability through likelihood-based analysis and natural incorporation of prior knowledge via Bayes' rule: $P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$ (Bouguila, 2011), allowing for more sophisticated uncertainty quantification and better handling of out-of-distribution inputs. We also acknowledge a separate class of hybrid generative-discriminative models, where some subset of parameters are trained generatively and others discriminatively (Raina et al., 2003; McCallum et al., 2006; Hayashi, 2025). However, we exclude them from our study, as their architectural differences hinder fair comparison with fully generative or discriminative models, placing them outside the scope of this work.

Membership Inference Attacks. Membership inference attacks exploit the differential behavior exhibited by machine learning models on training versus non-training data to infer whether a specific sample was part of the training set (Shokri et al., 2017). Recent advances have introduced sophisticated approaches: Watson et al. (2023) developed scalable attacks using quantile regression that significantly improve efficiency and accuracy, while Shi et al. (2022) provided theoretical foundations through first-principles analysis of membership inference vulnerabilities, and Duan et al. (2022) demonstrated novel output distillation methods extracting membership information from intermediate model representations. The privacy implications have become increasingly nuanced with Tan et al. (2023) discovering a counterintuitive "blessing of dimensionality" phenomenon where increased model parameters, when coupled with proper regularization, can simultaneously improve both privacy and performance, challenging traditional assumptions about overparameterization as privacy liability. Mireshghallah et al. (2023a,b) provided comprehensive empirical analyses of privacy-utility dynamics, developing practical privacy auditing techniques enabling efficient assessment of membership leakage with minimal computational overhead, while Choi et al. (2023) established fundamental connections between memorization and membership inference success, demonstrating that attacks are most effective against samples that models are likely to memorize regardless of distributional properties. These advances collectively underscore the critical importance of incorporating privacy considerations when evaluating different modeling paradigms—a gap our work addresses by systematically comparing MIA vulnerabilities across discriminative, generative, and pseudo-generative text classifiers. The scope of this work is limited to blackbox attacks with the assumption that logits are available from the model. Further, we assume that we can get the ground truth through human labelling. We assume that logits are vended out and cost of inference is negligible—specifically in the generative classifiers setting where the k-pass is the canonical setting. The knowledge distillation based approaches and trajectory based approaches are beyond the scope of this work.

3 Approach

We evaluate privacy vulnerabilities in text classification by training multiple classifiers across datasets and subjecting them to diverse membership inference attacks (MIAs). This enables a systematic comparison of the privacy–utility tradeoffs.

3.1 Classifier Paradigms

Following Kasa et al. (2025), we study 3 main classifier families:

Discriminative: Standard BERT-style encoders modeling $P(Y|X)$ using linear head on top of `[CLS]` token to directly map text X to label Y. There's no explicit memorization signal in this modeling approach.

Fully Generative: Models the joint distribution $P(X, Y)$. We consider the following sub-approaches:

(i) *Label-Prefix Autoregressive* models generate text x conditioned on a label prefix (e.g., `Positive: The film was a masterpiece.`). Classification is performed via logits using likelihood estimation, $\arg\max_{l \in K} \log P(x, y_l)$, in a K-pass fashion (K = number of labels). Such models may be more vulnerable to MIAs since logits expose information about $P(X)$. Alternatively, applying a softmax yields probabilities: $\mathrm{softmax}\big(\log P(x, y_l)\big) = P(x, y_l)/P(x) = P(y_l|x)$, where the shared denominator $P(x)$ cancels across classes. This dilution of $P(X)$ is expected to reduce susceptibility, which we further discuss in Section 5.

(ii) *Discrete Diffusion Models* are trained on (X, Y) pairs with a denoising objective. Following Lou et al. (2024), noise is gradually added to corrupt the input sequence to pure `[MASK]` tokens in the forward process, and the original input is reconstructed in the reverse process. At inference, x is given with the label masked and the model predicts y from `[MASK]`, conditional on x. In practice, there are two ways to obtain the predicted variable - a) doing a K-pass argmax on the logits similar to Autoregressive models and b) by simulating a trajectory from the reverse process, conditional on x. In Kasa et al. (2025), the predicted y is obtained by the latter approach. To obtain logits, we use the *Diffusion Weighted Denoising Score Entropy (DWDSE)*, which provides an upper bound on the log-likelihood: $-\log p_0^\theta(x) \leq \mathcal{L}_{DWDSE}(x)$ under the ELBO (Theorem 3.6 in Lou et al. (2024)).

Pseudo-Generative: We include this approach in our study, motivated by prior work on generative classifiers in Li et al. (2019) and Kasa et al. (2025). This category occupies a middle ground between discriminative and fully generative approaches. We consider two main sub-approaches:

(i) *Masked Language Models (MLMs)* are trained on a generative-like objective of reconstructing masked tokens rather than full causal modeling. However, they do not capture the true joint distribution $P(X, Y)$, but instead model the pseudo-likelihood (Wang and Cho, 2019b).

(ii) *Pseudo-Autoregressive Models* represent a recent development where traditional generative classifiers that model $P(X|Y)$ by prepending the label token are modified to append the label at the end of the input sequence. While this approach does not strictly model $P(X|Y)$, recent work (Li et al., 2025) demonstrates that label-appending can yield superior in-distribution performance compared to label-prepending. Notably, these approaches involve minimal architectural modifications to standard transformer models—typically requiring only changes to label placement or loss function computation—while preserving the core model design. This design principle allows for fair comparisons using widely available implementations that are accessible to practitioners, making these models particularly relevant for real-world deployment scenarios.

3.2 Membership Inference Attacks

We examine two main classes of MIAs:

Threshold-Based. Simple metrics derived from model outputs that might potentially contain hidden information useful for MIA: (i) *Max Probability:* $\max(P(y|x))$. (ii) *Entropy:* $H(P(y|x)) = -\sum_i p_i \log p_i$; lower for members and (iii) *Log-Loss:* Cross-entropy on the true label (requires label access).

Model-Based. An explicit attack model is trained in the following fashion: (i) Collect training data by querying the target classifier with member and non-member samples, (ii) represent each sample

using the target model's output probability vector concatenated with its one-hot encoded ground-truth label, (iii) train a Gradient Boosting Classifier with binary 0/1 labels indicating membership class (0 for non-member, 1 for member).

4 Experimental Methodology

This section details the concrete experimental setup used to test our hypotheses. We specify the datasets, training & evaluation procedures and computational details.

4.1 Datasets and Models

Datasets: Our evaluation is conducted on five public text classification benchmarks to ensure robustness across diverse domains and task complexities. The datasets are: **SST-5** Socher et al. (2013), **Hate Speech** Davidson et al. (2017), **Emotion** Saravia et al. (2018), **AG News** Zhang et al. (2015), and **IMDb** Maas et al. (2011). These datasets cover a range of tasks from binary sentiment to fine-grained multi-class topic and emotion classification, with varying text lengths and class balances. Further details are in the Appendix Section A.1.

Models: We implement and compare five classifier paradigms: (1) a discriminative **BERT Classifier**, (2) an **Autoregressive** model, (3) a **Masked Language Model (MLM)**, (4) a **Discrete Diffusion** model, and (5) a **Pseudo-Autoregressive** model. All models are trained from scratch to avoid confounding effects from pre-training, following Li et al. (2025) and Kasa et al. (2025). The specific architectural properties, attention mechanisms, and training objectives for each model are taken from Kasa et al. (2025) and are detailed in Section 3.

4.2 Training and Evaluation Protocol

Training. All models are trained using the AdamW optimizer with a learning rate of 3×10^{-5} and a weight decay of 0.01. We use a batch size of 32 and a maximum sequence length of 256 tokens, truncating longer inputs. A linear warmup is applied for the first 10% of updates, followed by linear decay. To prevent overfitting on the primary task, we employ early stopping with a patience of 20 epochs based on validation accuracy. Each experiment is repeated across three different random seeds to ensure the stability of our findings.

Attack and Evaluation. Each saved model checkpoint is subjected to the three categories of MIA 1. **Label-Agnostic Threshold Attacks**, 2. **Label-Aware Threshold Attacks**, and 3. **Model-Based Attacks**—as specified in Section 3.2. The primary metric for attack success (privacy leakage) is the Area Under the ROC Curve (**AUROC**). A score of 1.0 indicates a perfect attack, while 0.5 signifies performance equivalent to random guessing.

4.3 Computational Details

Experiments were conducted on a cluster of NVIDIA RTX 8000 and A100 GPUs. Training times varied based on model size and paradigm, ranging from approximately one hour for small discriminative models to over twelve hours for full-scale diffusion models on the largest datasets. We employed mixed-precision training to optimize computational efficiency. All model training and evaluation runs were performed in isolated environments to prevent any data contamination.

5 Analysis and Results

In this section, we present a comprehensive analysis of privacy vulnerabilities in text classification models. We begin by establishing theoretical bounds on MIAs, providing a framework for understanding potential privacy risks. Our empirical evaluation then validates these theoretical insights across different model architectures, comparing discriminative models, fully generative models, and pseudo-generative masked language models (MLMs). We analyze how different model output representations (logits versus probabilities) and various attack strategies affect vulnerability. Finally, we examine how different approaches to modeling the joint distribution $P(X, Y)$ influence privacy leakage, introducing pseudo-autoregressive models as a privacy-utility balanced alternative to fully

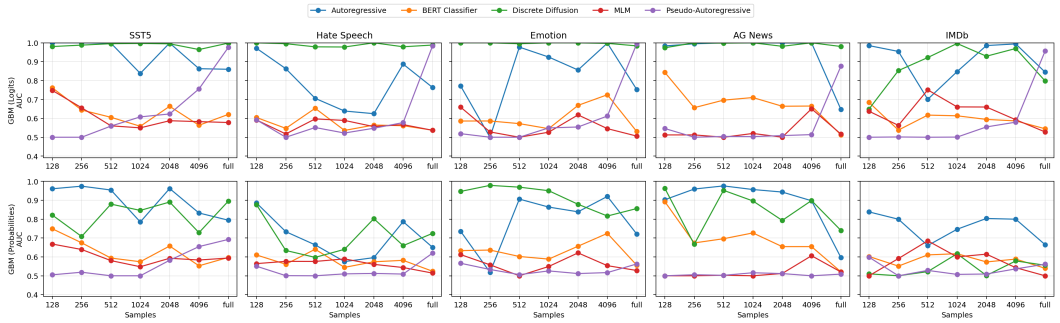

Figure 1: **[Best viewed in color]** Membership Inference Attack success rate (AUROC) compared across model architectures with varying training dataset sizes. We evaluate fully generative classifiers (`Autoregressive`, `Discrete Diffusion`), a discriminative classifier (`BERT Classifier`), and pseudo-generative models (`MLM`, `Pseudo-Autoregressive`). The **top row** displays attack performance using model **logits**, while the **bottom row** shows results using output **probabilities**. Higher AUROC values indicate increased privacy vulnerability.

generative models. Through this analysis, we demonstrate the deep connection between architectural choices and privacy vulnerabilities.

5.1 Theoretical Analysis and Bounds

To systematically analyze privacy vulnerabilities, we first establish a theoretical framework for measuring membership inference risk. We define the Membership Inference Advantage (MIA), which quantifies an adversary's ability to distinguish between training and test samples.

Definition 1 (Membership Inference Advantage). Let A be an adversary with a decision function $A(O(x))$ that outputs 1 if it guesses an input x is a member of the training set D_{train} and 0 otherwise, based on the model's output $O(x)$. The advantage is defined as the absolute difference between the adversary's true positive rate and false positive rate:

$$\varepsilon_{\text{MIA}} = |\mathbb{P}(A(O(x)) = 1 | x \in D_{\text{train}}) - \mathbb{P}(A(O(x)) = 1 | x \in D_{\text{test}})| \tag{1}$$

Building on this definition, we derive an upper bound on the membership inference advantage that decomposes the privacy risk into two components: one related to the marginal distribution $P(X)$ and another to the conditional distribution $P(Y|X)$.

Theorem 1. Let A be any adversary. The membership inference advantage is bounded by:

$$\varepsilon_{\text{MIA}} \leq \sqrt{\frac{D_{KL}(P_{\text{train}}(X)||P_{\text{test}}(X))}{2}} + \sqrt{\frac{\mathbb{E}x[D_{KL}(P_{\text{train}}(Y|X)||P_{\text{test}}(Y|X))]}{2}} \tag{2}$$

The proof for this theorem is presented in Appendix B.

This bound reveals that privacy leakage can occur through two channels: differences in the learned marginal distribution of inputs ($P(X)$) and differences in the learned conditional distribution of labels ($P(Y|X)$)Mahloujifar et al. (2022).

5.2 Empirical Analysis of Privacy Vulnerabilities

Armed with this theoretical framework, we now empirically investigate how these bounds manifest across different model architectures and attack scenarios. Our theoretical bound suggests that privacy leakage can occur through both the marginal distribution $P(X)$ and the conditional distribution $P(Y|X)$. This insight leads us to examine three key aspects: (1) the inherent vulnerability differences between generative models (which model both distributions) and discriminative models (which focus only on $P(Y|X)$), (2) the impact of different output representations (logits vs. probabilities) on information leakage, and (3) the effectiveness of various attack strategies in exploiting these vulnerabilities.

5.2.1 Generative Classifiers are Systematically More Susceptible

Our experimental results consistently demonstrate that fully generative models exhibit significantly higher vulnerability to MIA compared to their discriminative and pseudo-generative counterparts. As shown in Figure 1, this vulnerability gap is particularly pronounced in logit-based attacks, where fully generative models (`Autoregressive` and `Discrete Diffusion`) consistently yield the highest attack AUC across all datasets. While the discriminative `BERT Classifier` shows some vulnerability, its susceptibility remains notably lower.

These findings strongly support our hypothesis that explicitly modeling the joint distribution $P(X, Y)$ forces the model into a memorization-heavy regime. Unlike discriminative models that focus solely on learning the decision boundary for $P(Y|X)$, generative models must capture both the conditional distribution $P(Y|X)$ and the marginal data distribution $P(X)$. This additional modeling requirement significantly increases the likelihood of memorizing specific training samples, thereby amplifying privacy leakage.

Interestingly, our analysis of the relationship between model vulnerability and training data size reveals mixed trends, aligning with findings in Amit et al. (2024). We observe that MIA susceptibility fluctuates—sometimes increasing, sometimes decreasing—with the number of training examples. Table 3 provides additional evidence, comparing different attack types on probabilities (described in Section 3) averaged across all model architectures for varying training sample sizes. This variability can be partially explained by our use of early stopping with patience parameters. When the model's validation loss fails to improve for a specified number of epochs, training stops to prevent potential overfitting. This early stopping mechanism leads to varying levels of model convergence and, consequently, different degrees of memorization across dataset sizes.

5.2.2 Logits as a High-Bandwidth Privacy Leakage Channel

Our experiments demonstrate that MIA conducted using pre-softmax **logits** consistently achieve higher success rates compared to those using post-softmax **probabilities**. As illustrated in Figure 1, comparing the top row (logit-based attacks) with the bottom row (probability-based attacks) reveals a significant and consistent decrease in attack AUC across all model architectures and datasets when only probabilities are accessible. Additional results in Table 4 compare various probability-based attacks with GBM (Logits) across all five datasets. This observation aligns with previous findings in the literature Shokri et al. (2017) and can be attributed to the information-rich nature of logits. Unlike normalized probabilities, logits preserve the raw, unnormalized confidence scores between classes. The softmax transformation, while necessary for obtaining interpretable probabilities, compresses this information through normalization, effectively reducing the attack surface.

This finding has important practical implications: exposing raw logits through APIs, even for legitimate purposes such as temperature scaling or calibration, significantly increases privacy vulnerability. This is particularly concerning as many popular machine learning APIs and frameworks commonly expose logits by default OpenAI (2023). Therefore, practitioners should carefully consider implementing additional privacy-preserving mechanisms when logit access is required, or limit API outputs to probability distributions only.

5.2.3 Attack Strategies and Their Effectiveness

Our analysis demonstrates that membership inference success depends heavily on two factors: the sophistication of the attack strategy and the auxiliary information available to the adversary. Table 1 presents results for both threshold-based and machine learning-based attacks, focusing specifically on probability-based attacks as many of these methods are not applicable to logits.

Label-Agnostic Threshold Attacks represent the simplest approach, operating without knowledge of true labels and relying solely on the model's output probability distribution. We implement two such attacks: (1) **Max Probability**, which examines the highest confidence score in the output vector ($\max(P(y|x))$), and (2) **Entropy** of the output distribution ($H(P(y|x)) = -\sum_i p_i \log p_i$). These methods establish our baseline performance metrics. **Label-Aware Threshold Attacks** enhance the inference capability by incorporating ground-truth label information. This includes **Log-Loss**, which measures the cross-entropy with respect to the true label. Our results indicate that access to ground-truth labels consistently enhances attack performance, particularly against generative models.

Attack	BERT Classifier	Autoregressive	MLM	Discrete Diffusion	Pseudo-Autoregressive
Max Probability	0.56 ± 0.05	0.67 ± 0.13	0.55 ± 0.06	0.52 ± 0.13	0.51 ± 0.02
Entropy	0.56 ± 0.05	0.63 ± 0.12	0.55 ± 0.06	0.46 ± 0.13	0.51 ± 0.02
Log-Loss	0.60 ± 0.06	0.76 ± 0.13	0.55 ± 0.08	0.66 ± 0.13	0.53 ± 0.05
GBM	$\mathbf{0.62 \pm 0.08}$	$\mathbf{0.81 \pm 0.13}$	$\mathbf{0.56 \pm 0.07}$	$\mathbf{0.76 \pm 0.16}$	$\mathbf{0.53 \pm 0.06}$

Table 1: Membership inference attack performance (AUROC) across different model architectures, averaged over all datasets. Higher values indicate greater privacy vulnerability, with the highest values in each column shown in **bold**.

Machine Learning-Based Attacks, implemented here using **GBM**, represent our most sophisticated approach. This method trains a GBDT using probability vectors, ground truth and membership labels to learn complex decision boundaries between training and test samples. The effectiveness of this approach is particularly evident with diffusion models, where it achieves the highest AUC scores.

These results reveal a clear hierarchy: while simple threshold-based methods can breach privacy to some extent, the addition of ground-truth labels and advanced machine learning techniques significantly enhances attack success. This underscores the need for robust privacy protection strategies that account for varying levels of adversarial capabilities.

5.3 The Impact of Factorization: Decomposing Leakage in $P(X, Y)$

Building on our theoretical bounds and empirical findings, we now dive deeper into how different factorizations of the joint distribution $P(X, Y)$ affect privacy leakage. Our theoretical analysis showed that vulnerability stems from differences in both $P(X)$ and $P(Y|X)$ between training and test distributions. Here, we explore how architectural choices in modeling $P(X, Y)$ can shift the balance between these two sources of leakage.

We compare two approaches to modeling the joint distribution:

- **Autoregressive Models (Label-Prefix):** This model is trained to generate the text x conditioned on a label prefix y, thereby factorizing the joint distribution as $P(X, Y) = P(Y)P(X|Y)$. Its primary focus is on learning the class-conditional data distribution.
- **Pseudo-Autoregressive Models (Label-Suffix):** We introduce an autoregressive model trained to generate the full sequence (x, y), with the label appended at the end. This architecture implicitly factorizes the joint distribution as $P(X, Y) = P(X)P(Y|X)$. While still generative, its final step of predicting y after generating all of x mirrors a discriminative task.

This change in factorization from modeling $P(X|Y)$ to $P(X)$ first fundamentally alters the model's memorization patterns. Based on our theoretical bound, we hypothesize that this architectural shift redistributes privacy risk between the two components: the marginal distribution $P(X)$ and the conditional distribution $P(Y|X)$. To validate this, we measure the statistical divergence between loss distributions on training and test samples.

Table 2: Statistical divergence between training and test loss distributions. The Pseudo-Autoregressive model uses label-suffix architecture, while the Autoregressive model uses label-prefix. Higher values indicate greater leakage of the marginal distribution $P(X)$.

	Pseudo-Autoregressive		Autoregressive	
Dataset	JSD	KS Stat.	JSD	KS Stat.
SST-5	**0.8185**	**0.8314**	0.6204	0.6062
HateSpeech	**0.8355**	**0.8490**	0.4419	0.4107
Emotion	**0.8872**	**0.9240**	0.4780	0.4320
AGNews	**0.6230**	**0.6135**	0.2400	0.2143
IMDb	**0.8379**	**0.8681**	0.5232	0.5380

The results in Table 2 strongly support our hypothesis. Across all five datasets, the pseudo-autoregressive model shows substantially higher divergence in both Jensen-Shannon Divergence

(JSD) and Kolmogorov-Smirnov (KS) statistics. This indicates that the label-suffix architecture leads to greater leakage of information about the marginal distribution $P(X)$ compared to the label-prefix approach.

This difference stems from the models' underlying objectives. The label-suffix model must first construct a comprehensive representation of the input x before predicting y, necessitating high-fidelity modeling of $P(X)$. This requirement leads to increased memorization of training samples. In contrast, the label-prefix model focuses on learning conditional distributions $P(X|Y)$ for each class, potentially requiring less memorization of the complete data distribution.

These findings reveal a fundamental privacy trade-off aligned with our theoretical framework. Rather than eliminating privacy leakage, different factorizations of $P(X, Y)$ redistribute the vulnerability between the marginal and conditional components. While the label-suffix model may be more resistant to attacks targeting $P(Y|X)$: refer Figure 1 and Table 1, it becomes more vulnerable to those exploiting differences in $P(X)$ between training and test distributions.

6 Conclusion & Future Work

Our investigation reveals fundamental privacy vulnerabilities inherent in different text classification paradigms. Through theoretical analysis and extensive empirical validation, we demonstrate that generative models are systematically more vulnerable to membership inference attacks due to their explicit modeling of $P(X)$. Our theoretical framework, decomposing privacy risk into components from $P(X)$ and $P(Y|X)$, explains the observed hierarchy of vulnerability: generative models are most susceptible, followed by pseudo-generative approaches, with discriminative models showing the highest resistance. We show that this vulnerability is particularly pronounced when accessing logits rather than probabilities, and that sophisticated machine learning-based attacks can effectively exploit these vulnerabilities. Importantly, our analysis of different factorization strategies reveals that architectural choices in modeling $P(X, Y)$ fundamentally affect privacy leakage patterns.

Future research should focus on developing defense mechanisms specifically tailored to generative architectures, investigating privacy-preserving training methods that can maintain utility while reducing memorization, and exploring the impact of model scale and pre-training on privacy vulnerabilities. These directions, combined with our current findings, will help practitioners make informed decisions about model selection and deployment in privacy-sensitive applications.

References

Amit, G., Goldsteen, A., and Farkash, A. (2024). Sok: Reducing the vulnerability of fine-tuned language models to membership inference attacks.

Bouguila, N. (2011). Bayesian hybrid generative discriminative learning based on finite liouville mixture models. *Pattern Recognition*, 44(6):1183–1200.

Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., and Song, D. (2019). The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX security symposium (USENIX security 19)*, pages 267–284.

Castagnos, F., Mihelich, M., and Dognin, C. (2022). A simple log-based loss function for ordinal text classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4604–4609.

Choi, J., Chandrasekaran, V., Tople, S., and Jha, S. (2023). Exploring connections between memorization and membership inference. In *The Eleventh International Conference on Learning Representations*.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17, pages 512–515.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T.,

editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Duan, A. K., Kasiviswanathan, S. P., Kumar, R., and Mantrach, A. (2022). Flashing lights in my cnn: Membership inference by output distillation. *arXiv preprint arXiv:2208.08270*.

Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association*, 70:892–898.

Feng, Q., Kasa, S. R., KASA, S. K., Yun, H., Teo, C. H., and Bodapati, S. B. (2025). Exposing privacy gaps: Membership inference attack on preference data for llm alignment. In Li, Y., Mandt, S., Agrawal, S., and Khan, E., editors, *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 5221–5229. PMLR.

Hayashi, H. (2025). A hybrid of generative and discriminative models based on the gaussian-coupled softmax layer. *IEEE Transactions on Neural Networks and Learning Systems*, 36(2):2894–2904.

Jaini, P., Clark, K., and Geirhos, R. (2024). Intriguing properties of generative classifiers. In *The Twelfth International Conference on Learning Representations*.

Kasa, S. R., Goel, A., Gupta, K., Roychowdhury, S., Priyatam, P., Bhanushali, A., and Srinivasa Murthy, P. (2024). Exploring ordinality in text classification: A comparative study of explicit and implicit techniques. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5390–5404, Bangkok, Thailand. Association for Computational Linguistics.

Kasa, S. R., Gupta, K., Roychowdhury, S., Kumar, A., Biruduraju, Y., Kasa, S. K., Pattisapu, N. P., Bhattacharya, A., Agarwal, S., et al. (2025). Generative or discriminative? revisiting text classification in the era of transformers. *arXiv preprint arXiv:2506.12181*.

Li, A. C., Kumar, A., and Pathak, D. (2025). Generative classifiers avoid shortcut solutions. In *The Thirteenth International Conference on Learning Representations*.

Li, Y., Bradshaw, J., and Sharma, Y. (2019). Are generative classifiers more robust to adversarial attacks? In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3804–3814. PMLR.

Liang, P. and Jordan, M. I. (2008). An asymptotic analysis of generative, discriminative, and pseudolikelihood estimators. In *International Conference on Machine Learning*.

Lou, A., Meng, C., and Ermon, S. (2024). Discrete diffusion modeling by estimating the ratios of the data distribution. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011). Learning word vectors for sentiment analysis. In Lin, D., Matsumoto, Y., and Mihalcea, R., editors, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Mahloujifar, S., Sablayrolles, A., Cormode, G., and Jha, S. (2022). Optimal membership inference bounds for adaptive composition of sampled gaussian mechanisms.

McCallum, A., Pal, C., Druck, G., and Wang, X. (2006). Multi-conditional learning: Generative/discriminative training for clustering and classification. In *AAAI*, volume 1, page 6.

Mireshghallah, F., Goyal, K., Uniyal, A., Berg-Kirkpatrick, T., and Shokri, R. (2023a). An empirical analysis of the privacy-utility tradeoff in membership inference attacks. *arXiv preprint arXiv:2302.12580*.

Mireshghallah, F., Goyal, K., Uniyal, A., Berg-Kirkpatrick, T., and Shokri, R. (2023b). Privacy auditing with one (1) training run. *arXiv preprint arXiv:2310.08015*.

Ng, A. Y. and Jordan, M. I. (2001). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in Neural Information Processing Systems*.

OpenAI (2023). Openai api reference. `https://platform.openai.com/docs/api-reference`. Accessed: August 2023.

Pang, B. and Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In Knight, K., Ng, H. T., and Oflazer, K., editors, *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 115–124, Ann Arbor, Michigan. Association for Computational Linguistics.

Pattisapu, N., Kasa, S. R., Roychowdhury, S., Gupta, K., Bhanushali, A., and Murthy, P. S. (2025). Leveraging structural information in tree ensembles for table representation learning. *WWW*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2018). Language models are unsupervised multitask learners. *OpenAI*.

Raina, R., Shen, Y., Mccallum, A., and Ng, A. (2003). Classification with hybrid generative/discriminative models. *Advances in neural information processing systems*, 16.

Roychowdhury, S., Gupta, K., Kasa, S. R., and Srinivasa Murthy, P. (2024). Tackling concept shift in text classification using entailment-style modeling. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5647–5656.

Saravia, E., Liu, H.-C. T., Huang, Y.-H., Wu, J., and Chen, Y.-S. (2018). CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium. Association for Computational Linguistics.

Shi, J., Xu, J., Chen, Y., Wang, D., Li, J., Tian, Z., and Shokri, R. (2022). Membership inference attacks from first principles. *arXiv preprint arXiv:2211.00463*.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. (2017). Membership inference attacks against machine learning models.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Tan, J., LeJeune, D., Mason, B., Javadi, H., and Baraniuk, R. G. (2023). A blessing of dimensionality in membership inference through regularization. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 10968–10993. PMLR.

Wang, A. and Cho, K. (2019a). Bert has a mouth, and it must speak: Bert as a markov random field language model. *arXiv preprint arXiv:1902.04094*.

Wang, A. and Cho, K. (2019b). BERT has a mouth, and it must speak: BERT as a Markov random field language model. In Bosselut, A., Celikyilmaz, A., Ghazvininejad, M., Iyer, S., Khandelwal, U., Rashkin, H., and Wolf, T., editors, *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 30–36, Minneapolis, Minnesota. Association for Computational Linguistics.

Watson, L., Guo, C., Cormode, G., and Sablayrolles, A. (2023). Scalable membership inference attacks via quantile regression. *arXiv preprint arXiv:2307.03694*.

Yogatama, D., Dyer, C., Ling, W., and Blunsom, P. (2017). Generative and discriminative text classification with recurrent neural networks. *arXiv preprint*.

Zhang, X., Zhao, J., and LeCun, Y. (2015). Character-level convolutional networks for text classification. In *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 649–657, Cambridge, MA, USA. MIT Press.

Zheng, C., Wu, G., Bao, F., Cao, Y., Li, C., and Zhu, J. (2023). Revisiting discriminative vs. generative classifiers: Theory and implications. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J., editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 42420–42477. PMLR.

A Dataset Details

To provide further context, we briefly describe each dataset and its characteristics: **AG News** Zhang et al. (2015) contains roughly 120K training and 7.6K test samples, categorized into World, Sports, Business, and Technology. Each entry is a short news article comprising the title and initial sentences. **Emotion** Saravia et al. (2018) is composed of English tweets labeled with six core emotions: anger, fear, joy, love, sadness, and surprise, totaling 20K samples (16K train, 2K validation, 2K test). **Stanford Sentiment Treebank (SST)** Socher et al. (2013) is used in both its SST-2 (binary sentiment: positive/negative) and SST-5 (five sentiment classes: very negative, negative, neutral, positive, very positive) formats, enabling evaluation on both coarse and fine-grained sentiment tasks. **Multiclass Sentiment Analysis**[2] includes 41.6K samples labeled as positive, negative, or neutral, with notable class imbalance that tests a model's robustness to skewed distributions. **Twitter Financial News Sentiment**[3] is a domain-specific dataset of 11,932 finance-related tweets, annotated as Bearish, Bullish, or Neutral, requiring nuanced understanding of financial terminology. **IMDb** Maas et al. (2011) offers 50K equally split positive and negative long-form movie reviews, challenging models to process extended, opinion-rich text. **Rotten Tomatoes** Pang and Lee (2005) comprises 10,662 short movie review sentences (5,331 positive and 5,331 negative), emphasizing concise sentiment expression. Finally, **Hate Speech Offensive** Davidson et al. (2017) contains approximately 25K tweets categorized as hate speech, offensive (non-hate) language, or neutral, posing the challenge of fine-grained discrimination between harmful and non-harmful expressions.

B Theoretical Results

B.1 Proof of the General MIA Bound

Membership Inference Advantage:

Let A be an adversary with a decision function $A(O(x))$ that outputs 1 if it guesses an input x is a member of the training set D_{train} and 0 otherwise, based on the model's output $O(x)$. The advantage is defined as the absolute difference between the adversary's true positive rate and false positive rate:

$$\varepsilon_{\text{MIA}} = |\mathbb{P}(A(O(x)) = 1 | x \in D_{\text{train}}) - \mathbb{P}(A(O(x)) = 1 | x \in D_{\text{test}})| \tag{3}$$

Here, we provide a formal proof for the theorem that bounds the Membership Inference Advantage (ε_{MIA}) by the Kullback-Leibler (KL) divergence between the training and test data distributions. The proof synthesizes three key inequalities from information and probability theory.

Theorem 1. *Let A be any adversary. The membership inference advantage is bounded by:*

$$\varepsilon_{MIA} \leq \sqrt{\frac{D_{KL}(P_{train}(X)||P_{test}(X))}{2}} + \sqrt{\frac{\mathbb{E}_x[D_{KL}(P_{train}(Y|X)||P_{test}(Y|X))]}{2}} \tag{4}$$

Proof. First, we would like to clarify the meaning of the P_{train} and P_{test}.

$P_{train}(.)$ - represents the probability distribution of examples in the training set. Yeag $P_{test}(.)$ - represents the probability distribution of examples in the test set.

The proof proceeds in three steps.

Step 1: Bounding Advantage with Total Variation Distance. First, we bound the attacker's advantage by the total variation distance between the distributions of the model's outputs. Let $P_{\text{out\_train}}$ and $P_{\text{out\_test}}$ be the distributions of the model's outputs on members and non-members,

[2] `https://huggingface.co/datasets/Sp1786/multiclass-sentiment-analysis-dataset`

[3] `https://huggingface.co/datasets/zeroshot/twitter-financial-news-sentiment`

respectively. The maximum advantage of any statistical test to distinguish between two distributions is bounded by half of their total variation distance (TVD) - this follows from the definition of TVD:

$$\varepsilon_{\text{MIA}} \leq \frac{1}{2}\|P_{\text{out\_train}} - P_{\text{out\_test}}\|_1 \tag{5}$$

Step 2: Applying the Data Processing Inequality. A machine learning model is a function that processes input data to produce an output. The Data Processing Inequality states that no such processing can increase the statistical distance between distributions. Therefore, for a generative classifier, the distance between the output distributions cannot be greater than the distance between the original input data distributions, $P_{\text{train}}(X, Y)$ and $P_{\text{test}}(X, Y)$:

$$\|P_{\text{out\_train}} - P_{\text{out\_test}}\|_1 \leq \|P_{\text{train}}(X, Y) - P_{\text{test}}(X, Y)\|_1 \tag{6}$$

Combining these first two steps yields:

$$\varepsilon_{\text{MIA}} \leq \frac{1}{2}\|P_{\text{train}}(X, Y) - P_{\text{test}}(X, Y)\|_1 \tag{7}$$

Step 3: Connecting Total Variation to KL Divergence (Pinsker's Inequality). Finally, we relate the TVD to the KL divergence using Pinsker's Inequality:

$$\frac{1}{2}\|P - Q\|_1 \leq \sqrt{\frac{1}{2}D_{KL}(P||Q)} \tag{8}$$

Applying this to our bound from Step 2, we get:

$$\varepsilon_{\text{MIA}} \leq \sqrt{\frac{1}{2}D_{KL}(P_{\text{train}}(X, Y)||P_{\text{test}}(X, Y))} \tag{9}$$

Using the chain rule for KL-divergence, we can decompose the divergence of between joint distributions as:

$$D_{KL}(P(X, Y)||Q(X, Y)) = D_{KL}(P(X)||Q(X)) + \mathbb{E}_{x\sim P(X)}[D_{KL}(P(Y|X)||Q(Y|X))] \tag{10}$$

By applying this chain rule to the term inside our bound, we can separate the two primary sources of information leakage for a generative model that learns the joint distribution $P(X, Y)$:

$$\varepsilon_{\text{MIA\_gen}} \leq \sqrt{\frac{1}{2}\left(D_{KL}(P_{\text{train}}(X)||P_{\text{test}}(X)) + \mathbb{E}_x[\dots]\right)} \tag{11}$$

$$\leq \sqrt{\frac{D_{KL}(P_{\text{train}}(X)||P_{\text{test}}(X))}{2}} + \sqrt{\frac{\mathbb{E}_x[D_{KL}(P_{\text{train}}(Y|X)||P_{\text{test}}(Y|X))]}{2}} \tag{12}$$

where the second line uses the inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$.

This completes the proof. The attacker's advantage is fundamentally limited by how much the training data distribution diverges from the test data distribution.

This decomposed bound reveals two distinct vulnerability terms:

1. **Input Memorization Term**: $\sqrt{\frac{D_{KL}(P_{\text{train}}(X)||P_{\text{test}}(X))}{2}}$
   This term quantifies the leakage from the model memorizing the distribution of the training *inputs* themselves. This vulnerability exists because a generative model's objective function explicitly requires it to learn $P(X)$.

2. **Conditional Memorization Term**: $\sqrt{\frac{\mathbb{E}_x[D_{KL}(P_{\text{train}}(Y|X)||P_{\text{test}}(Y|X))]}{2}}$
   This term quantifies the leakage from the model overfitting the mapping from inputs to labels. This vulnerability exists for both generative and discriminative models.

□

C Extra Results

Attack	128	256	512	1024	2048	4096	Full Data
Entropy	0.51 ± 0.12	0.49 ± 0.10	0.50 ± 0.11	0.50 ± 0.11	0.51 ± 0.11	0.50 ± 0.11	0.50 ± 0.10
GBM	$\mathbf{0.61 \pm 0.16}$	0.58 ± 0.13	0.60 ± 0.15	0.60 ± 0.13	$\mathbf{0.62 \pm 0.13}$	0.60 ± 0.12	0.55 ± 0.07
Log Loss	0.61 ± 0.14	$\mathbf{0.60 \pm 0.12}$	$\mathbf{0.61 \pm 0.12}$	$\mathbf{0.61 \pm 0.12}$	0.61 ± 0.12	$\mathbf{0.61 \pm 0.11}$	$\mathbf{0.60 \pm 0.11}$
Max Probability	0.54 ± 0.12	0.52 ± 0.11	0.53 ± 0.11	0.53 ± 0.10	0.53 ± 0.10	0.53 ± 0.10	0.52 ± 0.09

Table 3: Membership inference attack performance (mean ± standard deviation AUROC) across varying training sample sizes. Higher values indicate greater privacy vulnerability, with the highest values in each column shown in **bold**.

Dataset	Entropy	GBM	Ground Truth Predictions	Log Loss	Max Probability
AG News	0.54 ± 0.11	$\mathbf{0.62 \pm 0.16}$	0.62 ± 0.15	0.62 ± 0.15	0.58 ± 0.13
Emotion	0.47 ± 0.14	0.61 ± 0.13	0.64 ± 0.11	$\mathbf{0.65 \pm 0.11}$	0.52 ± 0.12
HateSpeech	0.52 ± 0.06	$\mathbf{0.56 \pm 0.09}$	0.56 ± 0.06	0.56 ± 0.06	0.53 ± 0.06
IMDb	0.51 ± 0.08	$\mathbf{0.56 \pm 0.12}$	0.54 ± 0.07	0.54 ± 0.07	0.51 ± 0.08
SST5	0.48 ± 0.10	0.61 ± 0.13	0.66 ± 0.14	$\mathbf{0.67 \pm 0.14}$	0.51 ± 0.10

Table 4: Membership inference attack performance (mean ± standard deviation AUROC) across different datasets. Higher values indicate greater privacy vulnerability, with the highest values in each row shown in **bold**.