Calibrated Membership Inference Auditing of NLP Text Classifiers

Muhammad Abusaqer
Minot State University
Minot, ND, USA
muhammad.abusaqer@minotstateu.edu

***Abstract*—Membership inference evaluations can overstate practical privacy risk when attack thresholds are selected on test records, member and nonmember pools differ in class composition, duplicate texts cross partitions, or performance is summarized only by ROC-AUC. This paper presents a corrected and reproducible audit of membership leakage in SST-2 text classifiers. The protocol uses deterministic record-disjoint and normalized-text-disjoint partitions, class-matched attack pools, a separate attack-calibration set, and unchanged threshold transfer to a held-out attack-test set. Three attacks are evaluated: global loss, maximum confidence, and class-conditional standardized loss. For TF-IDF with Logistic Regression, the strongest held-out ROC-AUC rises from 0.5106 to 0.5309 as regularization weakens, while TPR at 1% FPR remains between 0.0074 and 0.0134. DistilBERT reaches substantially higher classification utility, but its primary held-out attack AUCs remain modest, from 0.5119 to 0.5250; across these attacks, TPR at 1% FPR never exceeds 0.0092. A direct protocol-sensitivity comparison over 21 model-attack combinations shows that selecting the threshold on the attack-test labels increases reported attack advantage by 0.0008–0.0120 while leaving threshold-free ROC-AUC unchanged. Thus, same-test threshold tuning introduces a consistent but modest optimistic bias in threshold-dependent metrics, and low-FPR reporting remains essential for interpreting practical membership risk. We additionally distill the corrected protocol into a reusable checklist for NLP membership-inference audits covering data construction, attack calibration, low-FPR reporting, uncertainty, and provenance.**

***Index Terms*—membership inference attacks, NLP privacy, privacy auditing, attack calibration, low-FPR evaluation, reproducibility, text classification**

I. Introduction

Machine-learning models can reveal information about their training data even when they expose only predictions. Membership inference attacks (MIAs) ask whether a particular record was used to train a target model [1], [2]. The attack is relevant to natural language processing (NLP) because text corpora can contain personal statements, behavioral traces, or other sensitive content, while classifier outputs can expose membership-related signals [3], [4].

A numerical MIA result is meaningful only when the evaluation protocol prevents information from flowing from the attack-test set into attack construction. A common threshold attack requires a decision threshold, and selecting that threshold on the same records used for final reporting can inflate apparent performance. Class imbalance between members and nonmembers can also confound the attack if one sentiment class is inherently easier or receives systematically different loss. Duplicate or near-duplicate texts across partitions introduce another source of leakage. Finally, aggregate ROC-AUC can conceal weak behavior at false-positive rates that matter in practice, where even a small FPR can produce many incorrect accusations of membership.

Building on the baseline study in [5], the present work introduces a substantially revised evaluation protocol, newly generated experimental artifacts, and expanded privacy and low-FPR analyses. The earlier study compared TF-IDF with Logistic Regression and DistilBERT on SST-2 using a loss-threshold attack and two simple mitigation settings. The new protocol instead uses deterministic duplicate-aware partitions; separates attack calibration from attack testing; matches member and nonmember pools by class and count; evaluates three attack scores; adds confidence intervals, low-FPR metrics, and top-risk precision; examines four Logistic Regression regularization settings and three DistilBERT epoch settings on the same corrected partitions; and links every stage to record identifiers, normalized-text hashes, configurations, and SHA-256 manifests. Because the data pools, threshold policy, and reporting criteria differ, the earlier numerical values are not treated as directly comparable results. Instead, a dedicated protocol-sensitivity experiment isolates the threshold-selection issue by applying the old test-selected rule and the corrected calibration-transferred rule to the same corrected held-out scores.

The contributions of this paper are:

- a leakage-resistant split and evaluation protocol for auditing SST-2 classifiers, with separate attack calibration and testing and matched attack pools;
- complete held-out privacy analyses for four TF-IDF with Logistic Regression settings and three frozen DistilBERT epoch settings using global loss, maximum confidence, and class-conditional loss, including bootstrap intervals and low-FPR measures;
- a controlled old-versus-corrected threshold-selection comparison on the same corrected data and frozen model scores, isolating the effect of selecting an operating threshold on the test labels; and
- an artifact-oriented reproducibility design that records row identity, text hashes, configurations, device use, and bundle checksums.

Accepted version of a full paper presented at the 38th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2026), Boca Raton, FL, Nov. 2 to 4, 2026. Proceedings to appear in IEEE Xplore. © 2026 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

II. Related Work

Shokri et al. introduced a practical black-box MIA based on shadow models and an attack classifier [1]. Yeom et al. connected membership leakage to overfitting and showed that per-example loss can support a direct threshold attack [2]. Sablayrolles et al. developed loss-oriented membership strategies under a Bayes-optimal formulation [6], and Li et al. evaluated membership leakage, generalization, and defenses for classification models [7]. Salem et al. reduced assumptions and computational requirements for black-box attacks [8], while later evaluations showed that attack success depends strongly on the data distribution, target model, available outputs, and evaluation choices [9], [10].

NLP has received less membership-inference attention than image classification, even though language datasets can contain distinctive or sensitive records. Shejwalkar et al. evaluated MIAs against NLP classifiers and found that attack behavior depends on architecture and task [3]. Mireshghallah et al. showed that reference-based likelihood ratios can reveal more leakage than target-model loss alone for masked language models, including at low false-positive rates [4]. Recent LLM studies further emphasize evaluation design: Puerto et al. show that ill-designed setups can permit shortcut-based apparent success and that aggregation across larger text units can recover membership signal [11], while Chen et al. show that localized window-based comparisons with reference models can outperform global loss averages at low false-positive rates [12]. Recent benchmarking also highlights the need to detect member/nonmember distribution shift [13]. Surveys organize threshold, shadow-model, label-only, and generative-model attacks together with defenses such as regularization and differential privacy [14], [15]. The current study does not propose a new attack algorithm or a formal privacy guarantee. Its contribution is a controlled, artifact-traceable integration of established evaluation safeguards for simple score-based attacks: calibration is isolated, pools are matched, duplicate leakage is controlled, and low-FPR behavior is reported.

The target task is SST-2, a binary sentence-level sentiment benchmark derived from the Stanford Sentiment Treebank and distributed through GLUE [16], [17]. The model families span a sparse linear classifier and DistilBERT, a compact transformer distilled from BERT [18]. This pairing permits a controlled comparison without adding many architectures that would dilute the protocol analysis; comparative NLP benchmarking has likewise shown the value of evaluating traditional and transformer-based text classifiers side by side [19].

III. Corrected Audit Protocol

A. Threat Model

The adversary receives the target classifier's probability vector for a labeled record (x, y) and computes a scalar score. Higher scores indicate stronger evidence of membership. The adversary does not access model parameters, gradients, hidden states, or the training procedure. This is a score-based black-box audit rather than a shadow-model attack.

TABLE I
Corrected SST-2 partitions.

Partition or pool	Records	Use
Target training	52,349	Fit target models
Model validation	5,000	Utility during sweep
Final utility test	872	One-time utility report
Calibration members	5,000	Fit attack threshold
Calibration nonmembers	5,000	Fit threshold/reference
Test members	5,000	Held-out privacy test
Test nonmembers	5,000	Held-out privacy test

The calibration adversary may use a balanced, labeled set of known members and known nonmembers. It selects any required threshold only from this calibration set. The held-out attack-test set is then scored once using the frozen score definition, class-conditional reference statistics, and threshold. Test labels are used only for final evaluation.

B. Dataset Construction

SST-2 supplies a labeled training split and an official validation split [16], [17]. The official training data are partitioned into 52,349 target-training records, 5,000 model-validation records, 5,000 attack-calibration nonmembers, and 5,000 attack-test nonmembers. The official validation split, containing 872 records, is reserved as an untouched final utility test. Two disjoint 5,000-record member pools are sampled from target training for attack calibration and attack testing.

Each attack stage contains 5,000 members and 5,000 nonmembers. Within each membership group, the class counts are identical: 2,211 negative and 2,789 positive records. Normalized text is grouped before assignment so that duplicate normalized-text groups stay in a single partition. Five normalized-text groups with conflicting labels are forced into target training and excluded from member selection. All primary partitions and attack stages are both record-disjoint and normalized-text-disjoint. Thirty-nine automated checks validate counts, uniqueness, disjointness, class matching, and hash consistency. Table I summarizes the design.

C. Target Models

The linear pipeline is implemented with scikit-learn [20] and uses a TF-IDF vocabulary fitted only on target training. It lowercases text, removes English stop words, uses unigrams and bigrams, requires document frequency of at least two, and caps the vocabulary at 20,000 features. Logistic Regression uses the `liblinear` solver, at most 1,000 iterations, seed 42, and $C \in \{0.1, 0.3, 1, 10\}$. The same fitted vectorizer is reused for every split and every C.

The transformer pipeline uses the Hugging Face Transformers implementation [21]. DistilBERT is loaded from `distilbert-base-uncased` at a pinned model revision with maximum sequence length 128, fixed padding and truncation, and global batch size 16. AdamW is configured with a learning rate of 2×10^{-5} and a weight-decay coefficient of 0.01. The remaining settings are a linear schedule with no warmup, gradient clipping at 1.0, FP16 automatic mixed precision, and seed 42. Separate final-epoch models are trained for one, two,

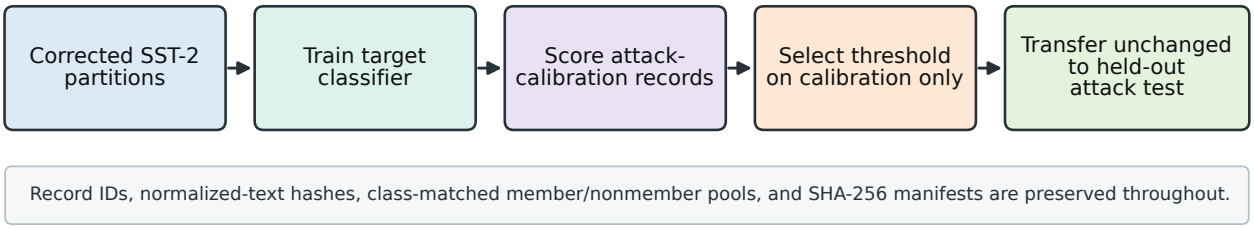


Fig. 1. Audit workflow. Thresholds and class-conditional reference statistics are selected using calibration records only and transferred unchanged to the held-out attack test.

and three epochs. Training uses one native-Windows process with `torch.nn.DataParallel` over two NVIDIA RTX A2000 12GB GPUs. Device markers in training and evaluation outputs verify that both GPUs processed records. The three epoch configurations are frozen using model-validation results before final utility or privacy analysis.

D. Attack Scores

Let $p_k(x)$ be the predicted probability for class k, and let

$$\ell(x, y) = -\log p_y(x) \tag{1}$$

be the true-label cross-entropy loss. Three primary scores are used.

Global loss uses the loss-oriented membership signal studied in prior classifier audits [2], [6], [7]: members are assigned higher scores when their true-label loss is lower,

$$s_{\text{loss}}(x, y) = -\ell(x, y). \tag{2}$$

Maximum confidence ignores the true label and uses

$$s_{\text{conf}}(x) = \max_k p_k(x). \tag{3}$$

Class-conditional loss standardizes loss against calibration nonmembers of the same true class. If μ_y and σ_y are the mean and standard deviation of calibration-nonmember loss for class y, the score is

$$s_{\text{cc}}(x, y) = \frac{\mu_y - \ell(x, y)}{\max(\sigma_y, 10^{-8})}. \tag{4}$$

The calibration reference is frozen and reused unchanged on attack-test records. The score convention is uniform: larger values are more member-like.

E. Threshold Selection and Metrics

For each model and score, the decision threshold maximizes attack advantage $\mathrm{TPR}-\mathrm{FPR}$ on the 10,000 calibration records. Ties are resolved conservatively by lower FPR and then by the highest finite threshold. That threshold is applied unchanged to the corresponding 10,000 attack-test records. Threshold-free test ROC-AUC is also computed on attack-test scores.

Privacy reporting includes test ROC-AUC, a stratified bootstrap 95% confidence interval, accuracy and advantage at the calibration-selected threshold, TPR at 1% FPR, TPR at 0.1% FPR, and precision among the highest-scoring 1% and 5% of records. With 5,000 test nonmembers, 1% FPR permits at most 50 false positives and 0.1% FPR permits at most five. Utility reporting includes accuracy, macro F1, cross-entropy loss, and train-validation gaps. The Logistic Regression sweep uses 1,000 stratified bootstrap iterations; DistilBERT uses 2,000.

To isolate the threshold-selection issue identified in the prior protocol, a separate sensitivity analysis reuses the finalized held-out score vectors without retraining or rescoring. The corrected operating point is the maximum-advantage threshold selected on attack calibration and transferred unchanged to test. The old procedure instead computes the ROC curve on the attack-test labels, selects the first threshold attaining the maximum attack advantage (TPR−FPR), and evaluates that threshold on the same records. Because both procedures use identical score vectors, ROC-AUC is invariant; only threshold-dependent metrics can change.

F. Reproducibility and Provenance

Every record carries a stable row identifier and normalized-text SHA-256 hash. Split files, configurations, record-level prediction and attack-score outputs, validation reports, model artifacts, and scripts are inventoried with SHA-256 checksums. The corrected split bundle hash is `6029ce64...8163796a9`; the finalized Logistic Regression bundle hash is `0222bbc9...e1e6fca9`. All 31 final Logistic Regression validation checks pass. The completed held-out DistilBERT privacy bundle is `f6d8447a...307a408d`, and the threshold-selection sensitivity bundle is `b666f48a...904125e5`. The latter validation confirms 21 primary comparisons, seven global-loss comparisons, unchanged protected-input hashes, and no model or score recomputation. These hashes identify the exact artifacts supporting each numerical statement rather than serving as security guarantees.

Code availability: A public repository for the code and reproducibility artifacts associated with this study is available at https://github.com/msaqer/ICTAI2026-Calibrated-Membership-Inference-NLP. The complete camera-ready reproducibility package will be deposited there following final archival review.

TABLE II
LOGISTIC REGRESSION UTILITY BY C.

C	Val. Acc.	Val. F1	Test Acc.	Test F1	Loss gap
0.1	.8198	.8110	.7546	.7471	.0105
0.3	.8542	.8505	.7821	.7791	.0233
1.0	.8708	.8682	**.7901**	**.7879**	.0435
10.0	**.8844**	**.8824**	.7878	.7867	.0978

TABLE III
HELD-OUT LOGISTIC REGRESSION PRIVACY. VALUES ARE FOR THE STRONGEST SCORE AT EACH C, CLASS-CONDITIONAL LOSS.

C	ROC-AUC [95% CI]	Adv.	TPR@1%	TPR@0.1%
0.1	.5106 [.4989,.5214]	.0318	.0112	.0026
0.3	.5155 [.5040,.5265]	.0392	.0098	.0014
1.0	.5212 [.5110,.5318]	.0448	.0092	.0018
10.0	.5309 [.5194,.5428]	.0584	.0074	.0016

IV. RESULTS

A. Logistic Regression Utility

Table II reports utility for the four regularization settings. Increasing C weakens regularization and improves model-validation macro F1 from 0.8110 at $C = 0.1$ to 0.8824 at $C = 10$. The untouched final test shows a different pattern: macro F1 peaks at 0.7879 for $C = 1$ and slightly declines at $C = 10$. At the same time, the train-validation accuracy gap grows from 0.0252 to 0.0528 and the validation-minus-train loss gap grows from 0.0105 to 0.0978. Thus, validation utility alone would favor $C = 10$, but final-test utility and fit gaps do not support a simple monotonic conclusion.

B. Held-Out Logistic Regression Privacy

Class-conditional loss is the strongest of the three scores for every C. Its held-out ROC-AUC increases from 0.5106 at $C = 0.1$ to 0.5309 at $C = 10$, as shown in Table III and Fig. 2. The 95% interval includes chance at $C = 0.1$, but lies above 0.5 for $C \geq 0.3$. Global loss follows a similar but slightly weaker trend. Maximum confidence is close to chance at strong regularization and reaches 0.5154 at $C = 10$.

The calibrated-threshold advantage also increases with C, from 0.0318 to 0.0584. However, operationally strict metrics tell a less favorable story. Across all 12 model-attack combinations, TPR at 1% FPR ranges only from 0.0074 to 0.0134. For the strongest class-conditional attack, TPR at 1% FPR decreases from 0.0112 at $C = 0.1$ to 0.0074 at $C = 10$, despite the higher aggregate AUC. At 0.1% FPR, the class-conditional TPR remains between 0.0014 and 0.0026. The highest-risk 1% precision is 0.54, 0.49, 0.47, and 0.43 as C increases, while top-5% precision remains near chance. Consequently, the AUC trend indicates weak rank separation, not a practically strong high-precision attack.

C. Privacy-Utility Interpretation

The Logistic Regression results show two distinct trends. Weaker regularization improves validation utility and increases aggregate leakage, consistent with larger train-validation gaps. Yet the best untouched-test macro F1 occurs at $C = 1$, not $C = 10$. More importantly, the setting with the largest AUC does not yield better low-FPR TPR or top-risk precision. A privacy-utility conclusion based only on validation F1 and ROC-AUC would therefore overstate both the utility benefit and attack practicality of $C = 10$.

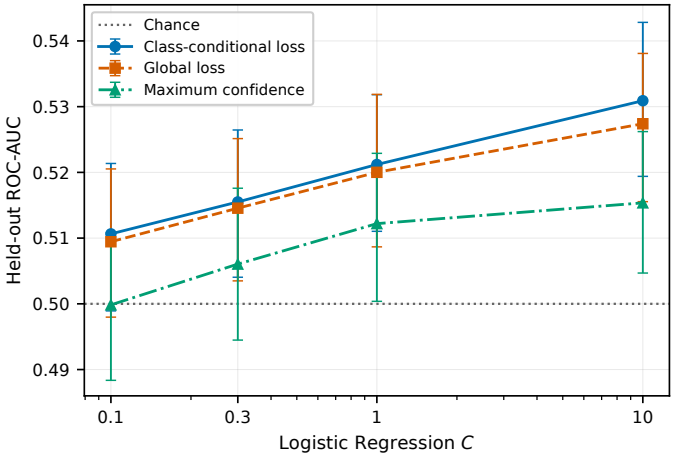


Fig. 2. Held-out attack ROC-AUC for the Logistic Regression sweep. Error bars show 95% bootstrap intervals.

The class-conditional score consistently improves on global loss because sentiment classes have different loss distributions. Standardizing against class-matched calibration nonmembers reduces the opportunity for class identity to act as a surrogate membership feature. Because attack pools are already class matched, the remaining difference reflects within-class loss separation rather than aggregate class imbalance.

D. DistilBERT Utility and Held-Out Privacy

The three DistilBERT configurations were frozen using model validation before protected data were accessed. Three epochs achieved the highest model-validation accuracy (0.9460) and macro F1 (0.9453), while the independent one-epoch run achieved the lowest validation cross-entropy loss (0.1918). All configurations then received a single final-utility evaluation on the 872-record official validation split.

Table IV shows that two epochs yield the highest final-test accuracy (0.9106) and macro F1 (0.9105). One epoch is close in classification utility and has the lowest final-test cross-entropy loss (0.2718). Three epochs slightly reduce final-test accuracy and macro F1 while increasing cross-entropy loss to 0.4383. At attack calibration, the nonmember-minus-member mean-loss gap grows from 0.0615 at one epoch to 0.1165 at two and 0.1739 at three.

The frozen thresholds and class-conditional reference statistics were then transferred unchanged to the 10,000-record held-out attack test. Across the three primary attacks, test ROC-AUC ranges from 0.5119–0.5137 after one epoch, 0.5163–0.5218 after two, and 0.5203–0.5250 after three. Table V reports the strongest primary AUC per epoch. The three-epoch global-loss attack is largest at 0.5250 [0.5141, 0.5365]. These aggregate differences remain operationally weak: across primary attacks, TPR at 1% FPR never exceeds 0.0092, TPR at 0.1% FPR never exceeds 0.0022, top-1% precision ranges from 0.43 to 0.49, and top-5% precision ranges from 0.492 to 0.522.

TABLE IV
DISTILBERT FINAL UTILITY AND CALIBRATION-STAGE LOSS SEPARATION.

Ep.	Test Acc.	Test F1	Test CE	Mem. loss	Nonmem. loss	Gap
1	.9071	.9071	**.2718**	.1223	.1838	.0615
2	**.9106**	**.9105**	.3618	.0678	.1843	.1165
3	.9060	.9058	.4383	.0409	.2148	.1739

TABLE V
HELD-OUT DISTILBERT PRIVACY. AUC IS THE STRONGEST PRIMARY ATTACK AT EACH EPOCH; LOW-FPR COLUMNS ARE THE MAXIMUM TPR ACROSS THE THREE PRIMARY ATTACKS.

Ep.	Strongest score	ROC-AUC [95% CI]	TPR@1%	TPR@0.1%
1	Class-cond.	.5137 [.5020,.5252]	.0086	.0022
2	Class-cond.	.5218 [.5108,.5329]	.0092	.0006
3	Global loss	.5250 [.5141,.5365]	.0084	.0008

Figure 3 summarizes the cross-model privacy–utility relationship using the strongest held-out attack score for each configuration. DistilBERT occupies a substantially higher utility regime than Logistic Regression while remaining in a similarly modest AUC range; within each family, additional fitting tends to move the configuration upward in aggregate leakage without a monotonic final-test utility gain.

V. DISCUSSION

A. Recommended Checklist for NLP Membership-Inference Audits

The controls used here can be stated as a reusable audit checklist. Table VI separates data-construction, attack-construction, reporting, and provenance controls so that later studies can identify which source of bias each safeguard addresses.

These controls are not specific to SST-2. For supervised text classifiers that expose probabilistic outputs, they provide a practical template for separating membership signal from partition artifacts and attack-selection leakage. The exact low-FPR resolution depends on the number of held-out nonmembers, and bootstrap intervals over attack records do not quantify variation across independently trained models. These distinctions should therefore be reported explicitly rather than folded into a single attack score.

B. Why Calibration Changes the Claim

An attack threshold is part of the attack, not merely a reporting choice. The direct sensitivity analysis quantifies the effect of choosing it on the held-out labels. Across all 21 primary model-attack combinations, the old same-test procedure increases reported attack advantage by 0.0008–0.0120 and accuracy by 0.0004–0.0060 relative to the calibration-transferred rule. The direction is consistent in all 21 comparisons because the old procedure explicitly maximizes advantage on the test labels; the magnitude is modest rather than dramatic. ROC-AUC is unchanged because both procedures use the same score vectors.

Table VII gives the most direct comparison for global loss, matching the loss-threshold attack used in the prior baseline. Here the old procedure increases advantage by 0.0012–0.0084. This experiment therefore supports a narrower claim than an order-of-magnitude correction: separating calibration from testing removes a small but systematic optimistic bias in threshold-dependent metrics and yields an operating rule that can be transferred reproducibly.

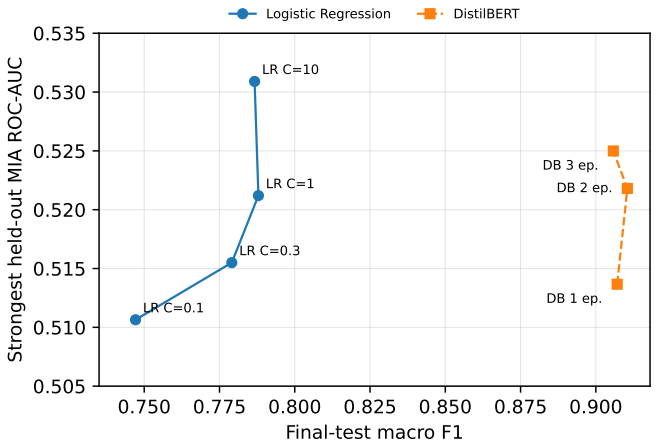


Fig. 3. Final-test macro F1 versus the strongest held-out MIA ROC-AUC for each model configuration. Higher F1 is better utility; ROC-AUC closer to 0.5 indicates weaker observed membership ranking.

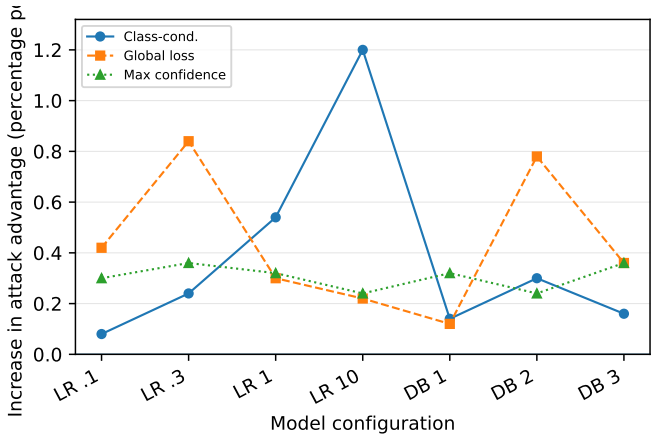


Fig. 4. Increase in reported attack advantage when the operating threshold is selected on the attack-test labels rather than transferred from calibration. The figure includes all 21 primary model–attack combinations.

Figure 4 extends this comparison to all 21 primary model–attack combinations. The positive deltas across every score and configuration visualize the systematic direction of the bias while also showing that its magnitude is generally small.

The use of 10,000 records per attack stage also gives explicit low-FPR resolution: one false positive corresponds to FPR 0.0002.

C. Aggregate AUC Is Not Operational Risk

The strongest Logistic Regression AUC is statistically above chance but only 0.5309, while the strongest DistilBERT primary AUC is 0.5250. Even these modest separations may look more consequential when presented without low-FPR results. At 1% FPR, the best observed Logistic Regression TPR is 1.34%, and the best DistilBERT primary TPR is 0.92%. In a deployment containing many nonmembers, such rules would identify very few true members while producing false accusations at the

TABLE VI
RECOMMENDED CHECKLIST FOR NLP MEMBERSHIP-INFERENCE AUDITS.

Audit component	Recommended practice	Failure controlled
DATA CONSTRUCTION		
Partition identity	Preserve stable record IDs and normalized-text hashes.	Hidden record or duplicate overlap
Member/nonmember pools	Match pool sizes and class composition.	Class or prevalence confounding
Duplicate handling	Keep normalized-text groups within one partition.	Text leakage across stages
ATTACK CONSTRUCTION		
Attack calibration	Use calibration records disjoint from attack test.	Test-set tuning
Threshold policy	Freeze calibration-selected thresholds and transfer them unchanged.	Optimistic threshold metrics
Conditional references	Estimate class/reference statistics on calibration only.	Reference leakage
REPORTING		
Primary reporting	Report ROC-AUC with uncertainty and calibrated-threshold metrics.	Overreliance on one aggregate metric
Strict-FPR reporting	Report TPR at 1% and, when supported, 0.1% FPR.	Hidden operational weakness
Risk concentration	Report precision among the highest-risk 1% and 5%.	AUC hiding poor top-ranked precision
UNCERTAINTY AND PROVENANCE		
Uncertainty/provenance	Separate attack-sample uncertainty from training-seed variability; preserve configurations and hashes.	Unclear reproducibility and uncertainty

TABLE VII
GLOBAL-LOSS THRESHOLD SENSITIVITY ON THE SAME HELD-OUT SCORES. "CORR." SELECTS THE THRESHOLD ON CALIBRATION; "OLD" SELECTS IT ON THE ATTACK-TEST LABELS.

Model	Setting	Corr. Acc.	Old Acc.	Corr. Adv.	Old Adv.	ΔAdv.
DistilBERT	1 ep.	.5166	.5172	.0332	.0344	.0012
DistilBERT	2 ep.	.5214	.5253	.0428	.0506	.0078
DistilBERT	3 ep.	.5266	.5284	.0532	.0568	.0036
LR	$C = .1$	.5125	.5146	.0250	.0292	.0042
LR	$C = .3$	.5153	.5195	.0306	.0390	.0084
LR	$C = 1$	.5240	.5255	.0480	.0510	.0030
LR	$C = 10$	.5340	.5351	.0680	.0702	.0022

allowed rate. This does not prove privacy safety. It shows that these specific simple scores, target models, and SST-2 partitions do not provide a strong operational attack under strict FPR constraints.

D. Regularization and Epoch Count

For Logistic Regression, increasing C enlarges coefficient norms, fit gaps, validation performance, and held-out AUC, but utility on the untouched test plateaus. The result supports a cautious association between weaker regularization and leakage rather than a causal privacy guarantee from regularization. A defense claim would require repeated seeds, more datasets, stronger attacks, and explicit utility constraints.

For DistilBERT, additional epochs reduce member loss much faster than nonmember loss in calibration, while final-test cross-entropy worsens after one epoch. The held-out audit now confirms a modest increase in aggregate membership ranking: the strongest primary AUC rises from 0.5137 at one epoch to 0.5250 at three. However, strict-FPR behavior remains weak and does not improve monotonically with epochs. Two epochs retain the best final-test macro F1 (0.9105), while three epochs have the largest aggregate attack AUC but no corresponding operationally strong low-FPR attack. The result supports an association between longer fine-tuning and stronger aggregate membership signal for this fixed seed, not a causal or general privacy claim.

E. Relation to the Prior Baseline

The prior MICS paper [5] established a pedagogical baseline with two model families, a single loss-threshold attack, smaller attack pools, and basic mitigation settings. The present work changes the scientific object being measured through duplicate-aware partitions, separate calibration and test stages, class matching, multiple scores, low-FPR metrics, bootstrap intervals, broader configuration sweeps, final utility testing, and artifact checksums. Because the data pools and reporting criteria differ, the numerical values from the two papers are not treated as direct replications. Instead, Table VII isolates one specific methodological difference by applying the old and corrected threshold policies to identical corrected scores. This separation also avoids reusing earlier text, figures, or tables as new contributions.

VI. LIMITATIONS AND VALIDITY THREATS

The study uses one binary sentiment dataset and one fixed seed. SST-2 is public and does not itself demonstrate harm on a sensitive corpus. The target models are limited to a sparse linear classifier and DistilBERT, and the evaluated attacks are simple score-based methods using prediction outputs. They do not include shadow models, augmentation, reference-model likelihood ratios, localized window attacks, or label-only strategies. Prior language-model studies show that reference-based, aggregated, or localized evidence can expose leakage missed by simple global scores [4], [11], [12]; accordingly, weak performance of the present scores must not be interpreted as a general privacy guarantee.

Repeated seeds are needed to quantify training variance, especially for DistilBERT. A second dataset with a security-relevant domain would improve external validity. Confidence intervals quantify attack-sample uncertainty for a fixed trained model, not variation across model training. Finally, SHA-256 manifests support provenance and change detection but do not replace public artifact release or independent replication.

VII. Conclusion

This paper presents a methodologically corrected audit of membership inference in SST-2 text classifiers. Logistic Regression shows weak but increasing held-out rank separation as regularization is reduced, with the strongest AUC reaching 0.5309 while TPR remains below 1.34% at 1% FPR. DistilBERT provides much higher classification utility, yet its strongest held-out primary AUC is only 0.5250 and its TPR at 1% FPR never exceeds 0.0092. The old-versus-corrected comparison further shows that selecting a threshold on the attack-test labels raises reported attack advantage by 0.0008–0.0120 across all 21 comparisons while leaving ROC-AUC unchanged. Thus, test-set threshold tuning creates a consistent but modest optimistic bias, and aggregate AUC alone remains insufficient for operational interpretation. Reliable privacy auditing requires disjoint calibration and testing, matched pools, low-FPR reporting, conservative claims, and traceable artifacts; the proposed audit checklist summarizes these controls as a reusable template for future NLP membership-inference evaluations.

References

[1] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *Proc. IEEE Symp. Security Privacy*, 2017, pp. 3–18.

[2] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha, "Privacy risk in machine learning: Analyzing the connection to overfitting," in *Proc. IEEE 31st Comput. Security Foundations Symp.*, 2018, pp. 268–282.

[3] V. Shejwalkar, H. A. Inan, A. Houmansadr, and R. Sim, "Membership inference attacks against NLP classification models," in *NeurIPS Workshop Privacy in Machine Learning*, 2021.

[4] F. Mireshghallah, K. Goyal, A. Uniyal, T. Berg-Kirkpatrick, and R. Shokri, "Quantifying privacy risks of masked language models using membership inference attacks," in *Proc. EMNLP*, 2022, pp. 8332–8347.

[5] W. Novak and M. Abusaqer, "Empirical evaluation of membership inference attacks on NLP text classifiers: A baseline study on SST-2," 2026, authors' version; presented at the 58th Midwest Instruction and Computing Symposium (MICS 2026), Eau Claire, WI, USA, Mar. 27–28, 2026. [Online]. Available: https://arxiv.org/abs/2609.10935

[6] A. Sablayrolles, M. Douze, C. Schmid, Y. Ollivier, and H. Jégou, "White-box vs black-box: Bayes optimal strategies for membership inference," in *Proc. 36th Int. Conf. Machine Learning*, 2019, pp. 5558–5567.

[7] J. Li, N. Li, and B. Ribeiro, "Membership inference attacks and defenses in classification models," in *Proc. 11th ACM Conf. Data and Application Security and Privacy*, 2021, pp. 5–16.

[8] A. Salem, Y. Zhang, M. Humbert, M. Fritz, and M. Backes, "ML-Leaks: Model and data independent membership inference attacks and defenses on machine learning models," in *Proc. Network Distrib. Syst. Security Symp.*, 2019.

[9] L. Song and P. Mittal, "Systematic evaluation of privacy risks of machine learning models," in *Proc. 30th USENIX Security Symp.*, 2021, pp. 2615–2632.

[10] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramèr, "Membership inference attacks from first principles," in *Proc. IEEE Symp. Security Privacy*, 2022, pp. 1897–1914.

[11] H. Puerto, M. Gubri, S. Yun, and S. J. Oh, "Scaling up membership inference: When and how attacks succeed on large language models," in *Findings of the Association for Computational Linguistics: NAACL 2025*. Albuquerque, NM, USA: Association for Computational Linguistics, Apr. 2025, pp. 4165–4182.

[12] Y. Chen, Y. Du, K. Zhang, A. Kundu, C. Fleming, B. Ribeiro, and N. Li, "Window-based membership inference attacks against fine-tuned large language models," in *35th USENIX Security Symposium (USENIX Security 26)*. Baltimore, MD, USA: USENIX Association, Aug. 2026, pp. 1187–1206.

[13] J. G. Wang, J. Wang, M. Li, and S. Neel, "CheckMIABench: Firm foundations for membership inference attacks on language models," in *Proc. 64th Annu. Meeting Assoc. Comput. Linguistics, Volume 2: Short Papers*, 2026, pp. 364–370.

[14] J. Niu *et al.*, "A survey on membership inference attacks and defenses in machine learning," *J. Information and Intelligence*, vol. 2, no. 5, pp. 404–454, 2024.

[15] H. Hu, Z. Salcic, L. Sun, G. Dobbie, P. S. Yu, and X. Zhang, "Membership inference attacks on machine learning: A survey," *ACM Computing Surveys*, vol. 54, no. 11s, pp. 1–37, 2022, art. no. 235.

[16] R. Socher *et al.*, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proc. EMNLP*, 2013, pp. 1631–1642.

[17] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *Proc. ICLR*, 2019.

[18] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," *arXiv:1910.01108*, 2019.

[19] M. Abusaqer and J. M. Saquer, "A comparative analysis of transformer and traditional ML models for cyberbullying detection on twitter (now X)," in *Proc. IEEE 49th Annu. Computers, Software, and Applications Conf. (COMPSAC)*, Toronto, ON, Canada, 2025, pp. 1607–1612.

[20] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[21] T. Wolf *et al.*, "Transformers:state-of-the-art natural language processing," in *Proc. EMNLP System Demonstrations*, 2020, pp. 38–45.