PRIVACY LEAKAGE OR DISTRIBUTION SHIFT? POST-HOC PRIVACY AUDITING OF FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Membership inference attacks (MIA) provide theoretically grounded privacy auditing for machine learning models, but applying it post-hoc to foundation models (LLMs/VLMs) is challenging. Indeed, MIAs quantify leakage by distinguishing data used in training (members) from non-members from the same distribution. In practice, auditors have to collect non-members from different sources, resulting in a distribution shift that biases privacy leakage measurements. In this work, we extend recently proposed debiased post-hoc privacy auditing techniques to make them practical for LLMs. First, we reduce the distribution shift by rewriting non-members in the style of member data and filtering all data for greater empirical overlap. Second, we strengthen MIAs on LLMs by adding paraphrase-related features that compare model behaviour on suspected members with model behaviour on semantically similar paraphrases. A thorough experimental evaluation on text and vision-language MIAs under distribution shifts demonstrates that our approach decreases distribution shift while retaining MIA signal, yielding meaningful privacy leakage estimates on large models.

1 INTRODUCTION

Privacy leakage on training data is a well-established concern in machine learning (Carlini et al., 2019; 2021). Such risk is increasingly important for large (vision) language models (LLMs/VLMs) as they are deployed across a growing range of applications, including settings involving personal or sensitive information (Carlini et al., 2022b; Nasr et al., 2023a; Pinto et al., 2024). Determining whether a trained LLM exposes information about individual training examples is therefore an important component of assessing its privacy risk. Membership inference attacks (MIAs) (Shokri et al., 2017), which test whether a particular example was included in model training (member) or not (non-member), provide a standard and theoretically grounded approach for measuring privacy leakage. More precisely, the success of a MIA can be converted through hypothesis testing into statistically valid empirical lower bounds on privacy leakage, as measured by differential privacy Dwork et al. (2006b); Wasserman & Zhou (2010); Dong et al. (2022), which has lead to MIAs forming the basis of privacy auditing (Nasr et al., 2023b; Steinke et al., 2023).

Nonetheless, auditing LLMs through MIA is considerably more difficult than auditing models in controlled learning settings (Duan et al., 2024; Meeus et al., 2024b). Indeed, modern foundation models are trained on massive and heterogeneous datasets with limited repeated exposure to individual examples, making per-instance membership signals intrinsically weak. Even more importantly, external auditors often have limited visibility and control over the training procedure and the provenance of the training data. As a result, they do not have access to held-out samples from the original training distribution to use as non-member data in the training and evaluation of a MIA. Practical post-hoc audits therefore rely on naturally collected member and non-member samples, often obtained from different sources or time periods outside of a target foundation model's training set Shi et al. (2023); Duan et al. (2024); Meeus et al. (2024a; 2025). The resulting distribution shift between member and non-member data substantially confounds MIA performance, inflating privacy leakage measurements (Das et al., 2025; Duan et al., 2024; Meeus et al., 2025). As a result, foundation model privacy auditing largely falls outside the scope of current post-hoc methods.

Recent works addressing this challenge are largely theoretical, focusing on techniques to correct for the distribution shift of non-member data. For instance, PANORAMIA (Kazmi et al., 2024)

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

calibrates a MIA audit against a data-only baseline that captures the distribution shift arising from imperfect synthetic non-members. Causal approaches use propensity-score estimation and reweighting to remove confounding from distribution shift (Even et al., 2026; Cebere et al., 2026). In the remainder of this paper, we will use the term *discriminator* for Panoramia's baseline and the causal approaches' propensity model, inspired by the GAN's terminology Goodfellow et al. (2014) in which a discriminator differentiates samples from two distributions. These approaches provide the statistical foundations for auditing under distribution shift, but their practical effectiveness relies on strong assumptions: namely (1) sufficient overlap between the member and non-member data, (2) accurate discriminators to correct for distribution shifts and (3) strong MIAs for measurement power. When naturally available non-members are strongly shifted, these requirements can substantially limit the power of the resulting audit. These techniques have indeed never been successfully applied to recent foundation models (Kazmi et al., 2024; Even et al., 2026; Cebere et al., 2026) and, when used out of the box, cannot measure privacy leakage (Section 3.2).

This work introduces three practical mechanisms aligned with the assumptions of Kazmi et al. (2024); Even et al. (2026); Cebere et al. (2026), that make those techniques more effective. First, we reduce the distribution shift of non-members by rewriting non-member text data using member examples as target-style references. Second, we build the audit sets by filtering members and rewritten non-members in regions of small empirical overlap further reducing the distribution shift. Third, we improve MIA by calibrating loss-based scores against the distribution of scores obtained from semantic paraphrases, providing a local reference for each example without requiring a separate reference model. To the best of our knowledge, these improvements enable us to show for the first time and across datasets with different sources of distribution shift, how to quantify nontrivial privacy leakage from several pre-trained foundation models using naturally collected members and non-members. Thus, these contributions close the gap between theory and practice by making the auditing techniques of Kazmi et al. (2024); Even et al. (2026); Cebere et al. (2026) usable.

2 Post-hoc Privacy Auditing under Distribution Shift

Figures 1 and 2 show the structure of a privacy auditing pipeline and MIA, respectively, as well as our extensions to improve auditing power. We start this section by reviewing the foundations of privacy-auditing with MIAs, and of correcting for distribution shifts (§2.1, gray boxes on Figure 1-2). This theory highlights three requirements for an effective privacy audit that are adressed by our contributions. First, member and non-member distributions need to be as close to each other as possible. We rewrite non-member text with the members' style, reducing observable distribution shifts while preserving the underlying content (§2.2). We also increase distribution overlap by filtering both original members and rewritten non-members with an independent discriminator. The resulting candidate set enables stronger privacy measurements (§2.3). Second, controlling for distribution shifts in privacy audits require strong discrimiator models to measure differences between members and non-members. We train complete models to measure residual distinguishability between the two groups (§2.4). Third, the power of the privacy audit depends on the strength of the MIA, and we construct new features to improve MIAs (§2.5). Together, these innovations make it possible to measure privacy leakage on pre-trained LLMs despite distribution shifts (§3).

2.1 Privacy Auditing under Distribution Shift

Membership inference and privacy auditing. Let $f = M(D_{\text{train}})$ be a target model and $S \in \{0, 1\}$ indicate whether a sample X is a *member* of D_{train}. A *MIA* predicts S using X and f, whereas a *discriminator* predicts S from X alone. When there is no distribution shift between members and non-members ($X \perp\!\!\!\perp S$) MIA success is theoretically related to differential privacy (DP) (Dwork et al., 2006b), as an ϵ-DP model satisfies $\text{TPR} \leq e^{\epsilon}\text{FPR}$ (Nasr et al., 2021; Steinke et al., 2023). Privacy audits turn this relation into a hypothesis test, in which sufficiently strong MIA performance rejects a candidate ϵ-DP guarantee, providing an empirical lower-bound on ϵ (we will denote lower-bounds on quantities as ϵ_{lb}).

Privacy auditing under distribution shift. Let $\mathcal{D}$ and $\mathcal{G}$ denote the member and non-member distributions, in which $p_{\mathcal{D}}, p_{\mathcal{G}}$ represent the probability density functions respectively. When $\mathcal{D} \neq \mathcal{G}$, membership can be predicted from the data alone, and MIA success can mistakenly attribute this

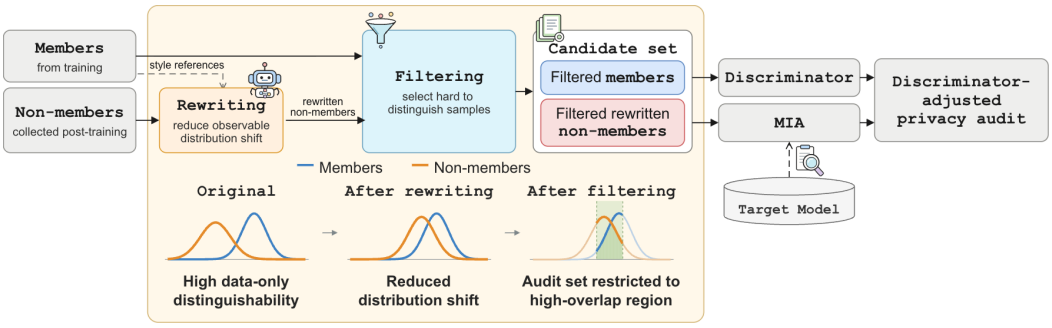


Figure 1: **Overview of the post-hoc privacy auditing pipeline.** Coloured components highlight our contributions to the privacy auditing pipeline (in grey). We reduce distribution shift between members and non-members by rewriting non-members toward the style of members and filtering both to regions of greater empirical overlap. The resulting filtered set is used to train a discriminator and MIA, and evaluate them to compute discriminator-adjusted privacy leakage estimates.

distribution shift to target-model data leakage. Recent post-hoc privacy auditing frameworks provide the theoretical building blocks for removing this distribution shift confounding.

PANORAMIA (Kazmi et al., 2024) characterizes distribution shift through the one-sided c-closeness definition: $\mathcal{G}$ is c-close to $\mathcal{D}$ if $\forall x,\ p_{\mathcal{D}}(x) \leq e^{c} p_{\mathcal{G}}(x)$. Assume that the discriminator makes r_b member predictions, of which t_b are correct, and the MIA makes r_a predictions, of which t_a are correct. Under c-closeness, and when the target model is ϵ-DP, PANORAMIA shows the distribution over correct membership predictions T_b and T_a are statistically dominated by $\mathrm{Binomial}(r_b, e^{c}/(1+e^{c}))$ and $\mathrm{Binomial}(r_b, e^{c+\epsilon}/(1+e^{c+\epsilon}))$ respectively. PANORAMIA uses tail-bounds on these distributions to obtain lower bounds c_{lb} and $(c+\epsilon)_{\mathrm{lb}}$, and reports the privacy leakage estimate as $\widetilde{\epsilon}_{\mathrm{panoramia}} = \max\{0, (c+\epsilon)_{\mathrm{lb}} - c_{\mathrm{lb}}\}$. This gap measures membership evidence in addition to what is revealed by the discriminator (the estimate itself is not a lower bound on ϵ as it subtracts two lower bounds). Importantly, a smaller distribution shift makes c smaller, reducing the confounding of the privacy measurement. A strong MIA increases t_a, making $(c+\epsilon)_{\mathrm{lb}}$ larger and increasing the power of the privacy leakage measurement. When a strong discriminator makes c_{lb} tighter and closer to the true c, in which case $\widetilde{\epsilon}_{\mathrm{panoramia}}$ more closely reflects the model's true privacy leakage.

Causal auditing (Cebere et al., 2026) instead represents distribution shift pointwise through the propensity score $\pi(x) = P(S = 1 \mid X = x)$, in which large values indicate that membership is highly predictable from the data. The causal audit then downweights such examples through randomized abstention. For an ϵ-DP model, with oracle propensities, the distribution of the number of retained correct MIA predictions t is stochastically dominated by $\mathrm{Binomial}(r, e^{\epsilon}/(1+e^{\epsilon}))$, and the resulting hypothesis test yields a finite-sample privacy lower bound. In practice, $\pi(x)$ is estimated from a discriminator and the causal framework bootstraps the propensity-estimation procedure and reports a lower p'-th quantile to account for uncertainty in estimating propensity scores. Similarly to PANORAMIA, closer member and non-member distribution shifts decrease $\pi(x)$, increasing the measurement power, as does a stronger MIA through a larger t. A strong discriminator is required for soundness, through valid estimates of $\pi(x)$. Appendix B provides extended descriptions.

2.2 Reducing Distribution Shifts through Text Rewriting

We reduce distribution shift by using an auxiliary language model to rewrite non-member text in the style of members, using member text as a style references. The task aims to preserve non-members' semantic content while reducing observable stylistic, formatting, or temporal cues that distinguish it from the members. This process is independent of the target model under audit, and only depends on member and non-member data. Sample rewriting prompts are shown in Appendix F. Rewriting avoids difficulties with generating non-member texts from scratch as Kazmi et al. (2024) do. Indeed as training a high-quality generator to approximate the member distribution can require substantial data and computation, and can easily introduce new generator-specific shifts that are easy to detect. We consider both text-only data for LLMs and multi-modal data for VLMs. In the latter case, the rewriting task is additionally conditioned on the paired images.

Category	Sample text
Original member	A dripping cake, also known as a dripper, is a traditional bread from Great Britain. The main ingredients are dripping, flour, brown sugar, spices, currants and raisins. The ingredients are mixed thoroughly and baked in an oven. Variations of dripping cake can be found in Wales, and in parts of England including Gloucestershire and Yorkshire. In Gloucestershire the dripping is allowed to form a toffee-like layer at the base of the cake. It can be confused with the Lardy cake, which is very similar apart from the principal ingredient being lard instead of beef dripping.
Original non-member	Louise-Félicité de Bréhan de Plélo, duchesse d'Aiguillon (30 November 1726 – 15 September 1796) was a French noblewoman, courtier and letter writer. *Biography.* Louise-Félicité was the daughter of Louis de Bréhan, Comte de Plélo and Louise-Françoise Phélypeaux de La Vrillière. She was orphaned at the age of 11 and subsequently raised as the ward of her uncle, Jean Frédéric Phélypeaux, Count of Maurepas, at Pentemont Abbey.
Rewritten non-member	Louise-Félicité de Bréhan de Plélo, duchesse d'Aiguillon (30 November 1726 – 15 September 1796), was a French noblewoman, courtier, and epistolary figure, noted for her documented correspondence and role within the late Ancien Régime court hierarchy. *Early life.* Louise-Félicité de Bréhan was born on 30 November 1726 in Paris, France, the daughter of Louis de Bréhan, Comte de Plélo, and Louise-Françoise Phélypeaux de La Vrillière. Orphaned at the age of eleven, she was formally placed under the guardianship of her maternal uncle, Jean Frédéric Phélypeaux, Count of Maurepas.

Image	Original caption	Rewritten caption
	<PERSON> decorates the Christmas tree vector character. royalty free illustration	Santa Claus decorates the Christmas tree vector character royalty free illustration
	The Endura Back Pack uses padded shoulder straps for added comfort	Endura Back Pack with Padded Shoulder Straps

Table 1: Examples of text rewriting for Temporal-wiki (original member, original non-member, and its rewritten version) and LAION (original and rewritten image captions).

Text rewriting. For text-only datasets, we randomly sample multiple members as reference provided to the auxiliary LLM when rewriting each non-member. The LLM is instructed to edit the non-member text while preserving its semantic content, and to reduce evidence making non-members distinguishable from the member references. The instructions additionally incorporate dataset-specific knowledge about known sources of shifts (Das et al., 2025). For instance, the instructions target formatting and URL-related artifacts in the Gutenberg benchmark and temporal markers and period-specific terminology in the other benchmarks. Table 1 shows an example.

Image caption rewriting. We extend the same rewriting strategy to image-caption datasets to audit VLMs.In these datasets, we keep non-member images, and rewrite captions which have been shown to exhibit large distribution shifs Das et al. (2025). To best preserve caption semantics, we rewrite using an auxiliary VLM that also gets the image as input. Table 1 illustrates how providing the image to the rewriting model gives additional context and leads to a correct replacement for the cc12m specific <PERSON> token with the Santa Claus on the image. This allows the rewrite to reduce text-specific cues while preserving semantic consistency with the image.

Mitigating rewriting artifacts. Rewriting with auxiliary LLMs can introduce LLM-specific artifacts, such as lexical choices or phrasing, creating a new source of distribution shift. To reduce dependence on the signature of a single LLM, we independently produce rewrites using multiple different models, and combine their outputs into the candidate non-member pool. This introduces heterogeneity in the rewriting process while retaining the same content-preservation and distribution-matching objective.

2.3 Filtering to Focus on Regions with Larger Distributional Overlap

Rewriting reduces the shift between members and non-members, but residual differences can remain, particularly for samples in the tails of the two distributions. These samples are often easily identified as members or non-members, even when a substantial overlapping region exists between the two populations. To counter this, we apply a filtering step to build a candidate set for the audit that concentrates in regions of greater empirical overlap. Specifically, we train a classifier to estimate the membership probability $P(\text{member} \mid x)$ for members and all non-member samples in the pool. The lowest-scoring $k\%$ of members is retained, which the classifier considers least member-like, and the highest-scoring $k\%$ of non-members, which it considers most member-like. The rule targets the lower tail of the member-score distribution and the upper tail of the non-member-score distribution, which correspond to the empirical overlap region when members generally receive higher scores

than non-members. We verify that this pattern holds in our datasets in Figure 4 in Appendix E. The filtering classifier is trained separately using only the input data and membership labels, without access to the target model or the downstream discriminator, and with no parameters shared between them.

2.4 Training the Discriminator

For **text datasets**, the member and non-member samples are encoded using a pretrained BERT model (Devlin et al., 2019), with all encoder parameters kept trainable for fine-tuning. For **image-caption datasets**, the discriminator must account for both the image and caption modalities: images are encoded using ResNet-50 (He et al., 2016), and captions using BERT. To capture cross-modal inconsistency, including possible semantic mismatches to its paired image when rewriting non-member captions, we also add embeddings from a pre-trained CLIP model (Radford et al., 2021) (not trained on the audit member or non-member sets), and their cosine similarity. We concatenate all representations (BERT, ResNet, CLIP), and apply an MLP classifier. All encoders and the final classification head are trained jointly to predict membership (see Fig. 3 in Appendix).

2.5 Membership Inference Attack

A MIA uses information from the target model to predict membership in the training set. High TPR at low FPR is particularly informative, as it indicates that the attack can confidently identify a subset of the training examples (Carlini et al., 2022a) and lower-bounds the ϵ parameter from DP (§2.1). For both text-only and image-caption settings, the MIA receives the same data inputs as the discriminator, together with features derived from the target model. This allows us to measure the membership signal leaking from the model beyond what can be inferred from the data alone. For **text datasets**, we collect loss-based features common in prior work, including token-level negative log-likelihood statistics (see §D), Zlib-calibrated likelihood (Carlini et al., 2021), and the normalized average log-probability of the $k\%$ least likely tokens (Shi et al., 2023; Zhang et al., 2025).

Inspired by LiRA (Carlini et al., 2022a), we also seek a per-input reference for the loss-based statistics. However, for feasibility this reference cannot involve training shadow models. We once again leverage paraphrasing, by generating K semantic paraphrases of each sample and computing all loss-based statistics s on them. For each s, this results in an empirical distribution in the semantic neighborhood of the original text, from which we compute:

$$z(s) = \frac{s_{\mathrm{o}} - \mu_{\mathrm{p}}}{\sigma_{\mathrm{p}}/\sqrt{K}}, \quad z_{\mathrm{robust}}(s) = \frac{s_{\mathrm{o}} - \tilde{\mu}_{\mathrm{p}}}{\mathrm{MAD}(s_{\mathrm{p}})/\sqrt{K}},$$

where s_{o} is the statistic for the original text, μ_{p} and σ_{p} are the mean and standard deviation over its paraphrases, and $\tilde{\mu}_{\mathrm{p}}$ denotes the median. The robust version reduces sensitivity to anomalous paraphrases with extreme scores. We pass all values features through an MLP before concatenating it to the discriminator's features and applying the membership classification head. While prior work has compared loss-based scores their mean over synthetic neighbours (Mattern et al., 2023), we capture how scores vary over their semantic neighbourhood for more precise membership inference.

For **image caption datasets**, we similarly retain the architecture of the corresponding data-only baseline and only change the source of the cross-modal model features, for which we use the embeddings and cosine similarity features from the audited target model instead of a CLIP model.

2.6 Auditing

Both frameworks we leverage (Kazmi et al., 2024; Even et al., 2026) return a lower-bound on privacy leakage ϵ only when the true closeness parameter c or propensity function $\pi(x)$ are upper-bounded (or exact). In practice, these quantities must be estimated, and both approaches differ in how they use the discriminator for this approximation, and account for uncertainty in that estimate.

Global versus pointwise correction. PANORAMIA takes a global view of distribution shift, using a single closeness parameter c shared across audit samples, whereas causal auditing takes a pointwise view and characterizes the same source of confounding through the propensity score $\pi(x)$, which can vary across samples. To link both approaches, let $c(x) \triangleq \log \frac{p_{\mathcal{D}}(x)}{p_{\mathcal{G}}(x)}$ be the point-wise

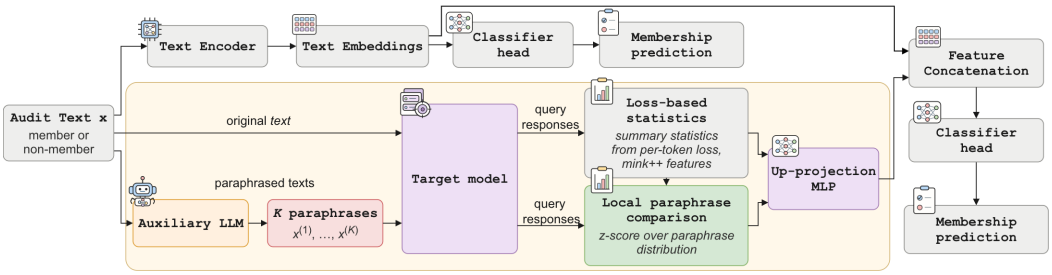


Figure 2: **Overview of MIA.** The upper branch consists of a text encoder and classifier head and serves as the discriminator when used alone. For the MIA, we augment standard loss-based statistics with K auxiliary-LLM paraphrases, by querying the target model on both the original and paraphrased texts to construct local paraphrase-based distributions. The target-model features are then up-projected and concatenated with the text embedding for membership classification.

closeness parameter (Definition C.1). We show (proofs in §C) that, conditioned on the audit set, c-closeness in PANORAMIA is the worst case over heterogeneous pointwise values:

Proposition 2.1 (c-closeness is a pointwise worst case)**.** *Let $\mathbf{x} = (x_1, \ldots, x_m)$ be a realized audit set consisting of m samples. Then $c_{\max}(\mathbf{x}) \triangleq \max_{x \in \mathbf{x}} c(x)$, is the smallest scalar c satisfying the closeness inequality simultaneously for all realized audit samples. That is, $c_{\max}(\mathbf{x}) = \inf \{c \in \mathbb{R} : e^{-c} p_{\mathcal{D}}(x) \leq p_{\mathcal{G}}(x),\ \forall x \in \mathbf{x}\}$.*

This point-wise closeness is directly related to the propensity scores of causal auditing, and upper-bounds of either are equivalent.

Proposition 2.2 (Propensity upper-bound implies point-wise and global closeness upper-bound)**.** *Let $\pi^{ub}(x) \geq \pi(x) \triangleq P(S = 1 \mid X = x)$ denote the oracle propensity function and its upper-bound, and let $c(x) \triangleq \log \frac{p_{\mathcal{D}}(x)}{p_{\mathcal{G}}(x)}$ denote the same for point-wise closeness. If $p_{\mathcal{D}}(x), p_{\mathcal{G}}(x) > 0$:*

$$c(x) = \log \frac{\pi(x)}{1 - \pi(x)} \leq \log \frac{\pi^{ub}(x)}{1 - \pi^{ub}(x)} \triangleq c^{ub}(x).$$

Given a discriminator model that upper-bounds its target $\pi(x)$, the causal audit provides a finite-sample valid privacy lower bound, by Theorem 5.1 and Corollary 6.1 of Cebere et al. (2026). Proposition 2.2 shows that in that case, using $c^{ub} \triangleq \max_{x \in \mathbf{x}} c^{ub}(x)$, where $c^{ub}(x) \triangleq \log \frac{\pi^{ub}(x)}{1 - \pi^{ub}(x)}$, in PANORAMIA also provides a valid privacy lower bound. Note however that default PANORAMIA uses the propensity model to lower-bound c, which in finite samples with be smaller than c even with an orable $\pi(x)$, which would only become tight asymptotically.

Statistical estimation on uncertainty. In practice, neither π nor c (or their upper-bounds) are known, and they need to be estimated from data using the discriminator. The two frameworks differ on how they manage uncertainty in this estimate. PANORAMIA uses the discriminator's performance to deduce a high probability lower-bound on the global c-closeness parameter. The causal approach on the other hand exploits point-wise propensity scores. This more precise estimate naturally exploits high overlap parts of the member/non-member distributions, but the impact of errors on the final estimate is harder to control, and Even et al. (2026) use a bootstrap to estimate propensity-estimation uncertainty. In both cases smaller sample sizes weaken audits. In PANORAMIA, fewer audit examples reduce statistical power to reject candidate values of c. In causal auditing, fewer examples for propensity estimation lead to higher variance bootstraps and depressed estimates.

3 Experimental Evaluation

We evaluate our approach on the benchmarks and target LLM and VLM models from Table 2. Distribution shifts can be temporal, or based on formatting and collection changes. N(m/n) represents the member/non-member sample size in the benchmarks. The compatible target models are publicly released models for which the member data is known to be included in training, and the non-member

data outside the corresponding training data or temporal cutoff (Duan et al., 2024; Das et al., 2025). Appendix D provides implementation details, and Appendix E additional experiments.

Table 2: Datasets with different types of distribution shifts, and target models.

Dataset	N (m/n)	Type of Distribution Shift	Compatible Target Model
Gutenberg	1000/1000	Formatting difference	OpenLLaMa-3B
WikiMIA-32	388/388	Temporal shifts	Pythia-2.8B/6.9B/12B, LLaMA-13B/30B
BookMIA	4935/4935	Temporal shifts	Pythia-2.8B/6.9B/12B, LLaMA-13B/30B
Temporal-Wiki	1000/1000	Temporal shifts	Pythia-2.8B/6.9B/12B, GPTNeo-125M/1.3B/2.7B
Temporal-Arxiv	1000/1000	Temporal shifts	Pythia-2.8B/6.9B/12B, GPTNeo-125M/1.3B/2.7B
LAION	950k/950k	Collection bias	Open-Clip ViT-B32, Open-Clip ViT-L14
CC12M	982k/982k	Collection bias	Open-Clip ResNet50

3.1 Post-hoc Privacy Auditing Results

Table 3 reports membership inference and privacy auditing results on the constructed sets of filtered members and rewritten non-members.

MIA performance. In most settings, the MIA significantly outperforms the discriminator, particularly in the low-FPR regime. For Gutenberg/OpenLLaMA-3B, for example, TPR increases from 0.03 to 0.25 at 1% FPR. Similarly on BookMIA/LLaMA-30B TPR at 1% FPR increases from 0.14 to 0.40. WikiMIA is the exception: the MIA does not always outperform discriminator, though the privacy audit still yields a positive estimate due to better performance at very low FPR (not shown in the table). Overall, our techniques result in lower distribution shift (low discriminator performance), while the target-model features still provide high-confidence membership information.

Panoramia privacy auditing. We perform PANORAMIA's joint hypothesis test with an overall significance level of $p = 0.05$ ($p = 0.025$ for each of the discriminator and MIA test). For most target models, $(c + \epsilon)_{\text{lb}}$ exceeds c_{lb} and results in a positive privacy leakage estimate $\widetilde{\epsilon}_{\text{panoramia}}$. For example, Gutenberg/OpenLLaMA-3B gives $(c + \epsilon)_{\text{lb}} = 1.79$ and $c_{\text{lb}} = 0.08$, resulting in $\widetilde{\epsilon}_{\text{panoramia}} = 1.71$. There are also situations in which $(c + \epsilon)_{\text{lb}}$ does not exceed c_{lb}, including Temporal-Wiki/GPTNeo-1.3B and Temporal-Arxiv/Pythia-2.8B, for which the resulting estimate is zero. These often coincide with smaller target models, which aligns with prior observations that smaller models can exhibit weaker memorization. Overall, the PANORAMIA results provide evidence of membership information associated with target-model access beyond signals captured by the discriminator in most evaluated settings.

Causal privacy auditing. The causal audit's $\widetilde{\epsilon}_{\text{causal}}$ denotes the empirical privacy lower bound obtained from the audit data and estimated propensity scores, while $\widetilde{\epsilon}^{q_{0.025}}_{\text{causal}}$ incorporate propensity model uncertainty, showing the lower 2.5% quantile over 100 bootstrap repetitions. Values of $\widetilde{\epsilon}_{\text{causal}}$ are positive for every reported benchmark/target-model pair, for example, $\widetilde{\epsilon}_{\text{causal}} = 0.64$ for Gutenberg/OpenLLaMA-3B, and the estimates range from 0.0293 to 0.3692 on Temporal-Wiki. The lower quantile of bootstrap estimates are substantially more conservative, even compared to the PANORAMIA high probability values. For example, we observe $\widetilde{\epsilon}^{q_{0.025}}_{\text{causal}}$ is zero for all WikiMIA and Temporal-Arxiv models and for the GPTNeo models on Temporal-Wiki: the variance from propensity estimation prevents establishing a positive bound on privacy leakage, especially in small datasets. For larger image caption datasets, we can still measure a positive empirical lower bound $\widetilde{\epsilon}^{q_{0.025}}_{\text{causal}}$ which value is close to the $\widetilde{\epsilon}_{\text{panoramia}}$ measured by PANORAMIA audit.

Interpreting Privacy Leakage Estimates. Positive PANORAMIA and causal boostrap gaps indicate evidence of privacy leakage from the target model. Taking the Gutenberg/OpenLLaMA result as an example, $\widetilde{\epsilon}_{\text{panoramia}} = 1.71$, indicating that the MIA finds significant excess membership signal compared to the discriminator. On the setting Gutenberg/OpenLLaMA, an estimate of $\widetilde{\epsilon}_{\text{causal}} = 0.64$ before bootstrap means the causal audit detects an empirical privacy lower bound of 0.64, conditional on the estimated propensity scores from the discriminator. The lower 2.5% quantile of $\widetilde{\epsilon}^{q_{0.025}}_{\text{causal}} = 0.26$, which accounts for propensity estimation variance (on the empirical distribution), is only able to measure a smaller 0.26 value for the audited target model's privacy parameter ϵ. Conversely, on WikiMIA/Pythia-12B, $\widetilde{\epsilon}_{\text{causal}} = 0.41$ and $\widetilde{\epsilon}^{q_{0.025}}_{\text{causal}} = 0$, so we have insufficient evidence to retain a positive bound when accounting for propensity-estimation uncertainty. On substantially larger image-caption benchmarks, $\widetilde{\epsilon}^{q_{0.025}}_{\text{causal}} = 1.74$, 2.09 and 2.06 respectively for OpenCLIP models on LAION and cc12m, larger than the values for $\widetilde{\epsilon}_{\text{panoramia}}$. The causal approach is better able

to leverage the statistical power of larger candidate sets. As discussed in §2.6, PANORAMIA and causal auditing correct for discriminator information and quantify uncertainty in different ways. Their numerical estimates should therefore not be interpreted as interchangeable. When both approaches yield positive estimates, they provide complementary evidence of privacy leakage from the target model after accounting distribution shifts between members and non-members.

Table 3: **Membership inference and privacy auditing results.** Target model D reports the results from discriminator, without access the target model (used to estimate c_{lb}). TPR@FPR measures MIA performance on a balanced set of N members and N non-members. $\tilde{\epsilon}_{\text{causal}}$ is the causal privacy-loss estimate, and $\tilde{\epsilon}_{\text{causal}}^{q0.025}$ is its bootstrap 2.5% quantile. c_{lb} and $(c+\epsilon)_{\text{lb}}$ are lower bounds from the hypothesis test each with $p = 0.025$, computed on an audit set of size N' (PANORAMIA includes samples in the test with a coin flip).

Dataset	Target Model	N	TPR@k%FPR			$\tilde{\epsilon}_{\text{causal}}$	$\tilde{\epsilon}_{\text{causal}}^{q0.025}$	N'(m/n)	c_{lb}	$(c+\epsilon)_{\text{lb}}$	$\tilde{\epsilon}_{\text{panoramia}}$
			1%	5%	10%						
Gutenberg	D	250	0.0360	0.1480	0.2040	\	\	136/114	0.0851	\	\
	OpenLLaMA-3B		0.2520	0.4680	0.5840	0.6413	0.2595		\	1.7963	1.7211
WikiMIA	D	155	0.1061	0.2065	0.3355	\	\	82/73	0.0572	\	\
	Pythia-6.9B		0.0839	0.1806	0.3161	0.4431	0		\	0.2539	0.1967
	Pythia-12B		0.0903	0.1806	0.3097	0.4138	0		\	0.2539	0.1967
	LLaMA-13B		0.0839	0.1806	0.3419	0.3988	0		\	0.2696	0.2124
	LLaMA-30B		0.0774	0.1613	0.3290	0.4138	0		\	0.1168	0.0596
BookMIA	D	625	0.1440	0.5280	0.7920	\	\	301/324	1.6718	\	\
	Pythia-6.9B		0.2848	0.6384	0.8144	0.1620	0.1029		\	2.1663	0.4945
	Pythia-12B		0.2880	0.6864	0.8480	0.3196	0		\	2.1433	0.4715
	LLaMA-13B		0.3376	0.6752	0.8864	0.2349	0		\	2.0113	0.3395
	LLaMA-30B		0.4016	0.7456	0.8944	0.2808	0.0810		\	2.3615	0.6897
Temporal-Wiki	D	500	0.0580	0.1760	0.2960	\	\	233/267	0.1166	\	\
	Pythia-2.8B		0.1280	0.2320	0.3300	0.2579	0.0809		\	0.3514	0.2347
	Pythia-6.9B		0.0960	0.2220	0.3540	0.2833	0.0346		\	0.3064	0.1898
	Pythia-12B		0.0980	0.2360	0.3540	0.3692	0.0851		\	0.4445	0.3278
	GPTNeo-125M		0.0840	0.1940	0.2980	0.0699	0		\	0.1419	0.0253
	GPTNeo-1.3B		0.0760	0.1720	0.2740	0.0293	0		\	0.1114	0
	GPTNeo-2.7B		0.0920	0.1900	0.3080	0.1107	0		\	0.1615	0.0448
Temporal-Arxiv	D	250	0.0720	0.2400	0.3480	\	\	136/114	0.4870	\	\
	Pythia-2.8B		0.1040	0.2320	0.4000	0.1470	0		\	0.4598	0
	Pythia-6.9B		0.1160	0.2720	0.4200	0.1771	0		\	0.4999	0.0129
	Pythia-12B		0.1240	0.2120	0.3840	0.1888	0		\	0.5771	0.0901
	GPTNeo-125M		0.1280	0.2520	0.4000	0.1470	0		\	0.5771	0.0901
	GPTNeo-1.3B		0.1520	0.2560	0.3480	0.2310	0		\	0.6713	0.1842
	GPTNeo-2.7B		0.1200	0.2320	0.3760	0.1771	0		\	0.5459	0.0588
Laion/cc12m	D	23.8k	0.1285	0.3153	0.4442	\	\	11.9k/11.9k	2.3879	\	\
	OpenClip-ViT-B-32		0.2122	0.4293	0.5994	1.4776	0.4203		\	3.0705	0.6826
	OpenClip-ViT-L-14		0.2436	0.4914	0.6216	1.6495	0.5984		\	3.2219	0.8510
cc12m/cc3m	D	25k	0.2137	0.4238	0.5452	\	\	12.5k/12.5k	3.2891	\	\
	OpenClip-Resnet50		0.3959	0.6088	0.7219	2.1166	0.4518		\	4.3963	0.7360

3.2 Ablation Studies

Reducing distribution shift improves auditing. Table 4 evaluates the effect of our rewriting and filtering techniques, by comparing measurements based on the original member/non-member sets, rewritten non-members, and filtered sets with retention rate k. On Temporal-Wiki/Pythia-2.8B, the original non-member data exhibit substantial distribution shift, and the discriminator achieves a TPR of 0.41 at 1% FPR, yielding a PANORAMIA $c_{\text{lb}} = 2.7319$ compared with $(c+\epsilon)_{\text{lb}} = 2.7071$, and thus $\widetilde{\epsilon}_{\text{panoramia}} = 0$. Rewriting reduces distribution shift, with TPR at 1% FPR decreasing to 0.27, and c_{lb} to 1.85. Filtering produces an even larger reduction in discriminator performance. At $k = 0.50$, c_{lb} decreases to 0.1166, while preserving enough MIA signal for $(c+\epsilon)_{\text{lb}} = 0.3514$. Put together, it gives the largest $\widetilde{\epsilon}_{\text{panoramia}} = 0.23$.

The causal audit shows a similar benefit from reduced shift: $\widetilde{\epsilon}_{\text{causal}} = 0$ for the original benchmark, but it reaches 0.19 when rewriting non-members, consistent with the assumption that propensity correction retains more informative guesses when member and non-member are similar (Cebere et al., 2026). Soft filtering is also helpful, with the larget $\widetilde{\epsilon}_{\text{causal}} = 0.25$ when $k = 0.50$. Filtering too aggressively trades off reduced confounding for lower statistical power: at $k = 0.25$, where only 250 members and non-members are retained, both $\widetilde{\epsilon}_{\text{panoramia}}$ and $\widetilde{\epsilon}_{\text{causal}}$ fall to zero. The same

pattern appears on the LAION benchmark, in which $c\mathrm{lb}$ decreases with more aggressive filtering, and $\widetilde{\epsilon}_{\mathrm{panoramia}}$ increases from 0 to positive estimates. Similarly $\widetilde{\epsilon}_{\mathrm{causal}}$ is largest when filtering at $k = 0.25$. Overall, these ablations show that rewriting and filtering can reduce distribution shift and reveal privacy leakage from the target model, but effective auditing requires balancing member–non-member overlap against the sample size retained for statistical estimation.

Table 4: **Effect of reducing distribution shift by rewriting and filtering.** *Original* denotes unchanged members and non-members. *Rewritten N* rewrites the non-members, and *Filtered (k)* keeps a fraction k of members and non-members. D is the discriminator, which does not access the target model, and MIA is the attack with target model access. Other columns are as in Table 3.

Dataset/Model	Setup		N	TPR@k%FPR			$\widetilde{\epsilon}_{\mathrm{causal}}$	$\widetilde{\epsilon}^{q0.025}_{\mathrm{causal}}$	N'(m/n)	c_{lb}	$(c+\epsilon)_{\mathrm{lb}}$	$\widetilde{\epsilon}_{\mathrm{panoramia}}$
				1%	5%	10%						
Temporal-wiki/ Pythia-2.8B	Original M, N	D	1000	0.4190	0.5950	0.6540	\	\	476/524	2.7319	\	\
		MIA		0.3860	0.5590	0.6550	0	0		\	2.7071	0
	Original M, Rewritten N	D	1000	0.2760	0.5640	0.7520	\	\	476/524	1.8596	\	\
		MIA		0.5367	0.7867	0.9100	0.1919	0		\	1.8970	0.0373
	Filtered M, N (k=0.75)	D	750	0.0560	0.2200	0.4227	\	\	365/385	0.8533	\	\
		MIA		0.0533	0.2587	0.4067	0.0124	0		\	0.9042	0.0510
	Filtered M, N (k=0.50)	D	500	0.0580	0.1760	0.2960	\	\	233/267	0.1166	\	\
		MIA		0.1280	0.2320	0.3300	0.2579	0.0809		\	0.3514	0.2347
	Filtered M, N (k=0.25)	D	250	0.0080	0.0800	0.1080	\	\	136/114	0	\	\
		MIA		0.0120	0.0640	0.1160	0	0		\	0	0
LAION/ ViT-B-32	Original M, N	D	95.1k	0.9766	0.9845	0.9884	\	\	47.6k/47.6k	7.3307	\	\
		MIA		0.9772	0.9853	0.9889	0.1.6964	0.1355		\	7.0437	0
	Original M, Rewritten N	D	95.1k	0.9037	0.9641	0.9777	\	\	47.6k/47.6k	6.8188	\	\
		MIA		0.9099	0.9663	0.9801	1.9658	0.2248		\	6.6028	0
	Filtered M, N (k=0.75)	D	71.4k	0.5441	0.7501	0.8331	\	\	35.7k/35.7k	5.1495	\	\
		MIA		0.6286	0.8108	0.8799	1.6678	0.2202		\	5.6417	0.4922
	Filtered M, N (k=0.50)	D	47.6k	0.3588	0.6072	0.7238	\	\	23.8k/23.8k	3.9700	\	\
		MIA		0.4817	0.7104	0.8075	1.7040	0.2763		\	4.7329	0.7629
	Filtered M, N (k=0.25)	D	23.8k	0.1285	0.3153	0.4442	\	\	11.9k/11.9k	2.3879	\	\
		MIA		0.2122	0.4293	0.5594	1.4776	0.4203		\	3.0705	0.6826

Strength of MIA. Table 5 examines how target-model features affect membership inference performance and privacy auditing results. As shown in row 2-5, threshold-based attacks using a single statistic provide weak membership signal, and both $\widetilde{\epsilon}_{\mathrm{panoramia}} = 0$ and $\widetilde{\epsilon}_{\mathrm{causal}} = 0$. Our candidate members and non-members are more difficult to distinguish by construction, making model-specific membership difficult to recover from a single statistic. Rows 6-8 show that MIA performance improves when combining multiple target-model features for text representation. Using loss-based features alone gives $(c+\epsilon)_{\mathrm{lb}} = 0.1070 < c_{\mathrm{lb}} = 0.1166$, yielding $\widetilde{\epsilon}_{\mathrm{panoramia}} = 0$, although the causal audit gives a small positive estimate of 0.0434. Adding Min-K%++ features (Zhang et al., 2025) increases TPR at 1% FPR from 0.045 to 0.060 and increases $(c+\epsilon)_{\mathrm{lb}}$ to 0.1821, giving $\widetilde{\epsilon}\mathrm{panoramia} = 0.0655$ and $\widetilde{\epsilon}_{\mathrm{causal}} = 0.0990$. Finally, adding paraphrasing-based z-score and robust z-score features produces the strongest attack, with TPR of 0.12 at FPR of 1%, and the highest audit estimates of $\widetilde{\epsilon}_{\mathrm{panoramia}} = 0.2347$ and $\widetilde{\epsilon}_{\mathrm{causal}} = 0.2579$, with even $\widetilde{\boldsymbol{\epsilon}}^{\boldsymbol{q0.025}}_{\mathrm{causal}} > 0$. These ablation results show that enhancing a classifier-based MIA with multiple statistics and a local reference distributions constructed from semantic paraphrases substantially improves membership inference and strengthens the resulting privacy audit.

Table 5: **Ablation of MIA features.** On Temporal-wiki with Pythia-2.8B as the target model. Top: data-only baseline D. Middle: threshold-based attacks using a single statistic. Bottom: classifier-based attacks, with features added cumulatively (+). Columns are as in Table 3.

MIA Features	TPR@k%FPR			$\widetilde{\epsilon}_{\mathrm{causal}}$	$\widetilde{\epsilon}^{q0.025}_{\mathrm{causal}}$	c_{lb}	$(c+\epsilon)_{\mathrm{lb}}$	$\widetilde{\epsilon}_{\mathrm{panoramia}}$
	1%	5%	10%					
Data-only baseline D	0.0580	0.1760	0.2960	\	\	0.1166	\	\
Mean NLL Loss	0.0300	0.1000	0.1400	0	0	\	0	0
Min-K%++ ($k = 10$)	0.0300	0.0920	0.1680	0	0	\	0	0
z-score with Min-K%++ ($k = 60$)	0.0300	0.0820	0.1100	0	0	\	0	0
Loss-based features	0.0450	0.1380	0.2680	0.0434	0	\	0.1070	0
+ Min-K%++	0.0600	0.1560	0.2440	0.0990	0	\	0.1821	0.0655
+ Paraphrase-relative features	0.1280	0.2320	0.3300	0.2579	0.0809	\	0.3514	0.2347

AI USE STATEMENT

In this work, we use on open-source LLM models to generate synthetic data that are used in two steps in our pipeline: this is a core part of our methodology for privacy auditing, described in depth in the paper. We have also used Claude Code to assist with code implementation and graphic art design. GPT and Claude were also used to skim through articles when exploring related work, and to dig into some technical details. Generative AI tools were *not* used to formulate the research questions, develop the main ideas, or auto-run experiments. AI assistance in writing was limited to grammar correction, fluency editing, and latex formatting. It was *not* used to generate long chunks of text, which was all written and edited by the authors. We have reviewed all AI-assisted work, and the LLM-generated code was verified and tested for correctness. We take full responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pp. 267–284, Santa Clara, CA, August 2019. USENIX Association. ISBN 978-1-939133-06-9. URL https://www.usenix.org/conference/usenixsecurity19/presentation/carlini.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650. USENIX Association, 2021.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2022a. doi: 10.1109/SP46214.2022.9833649.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. Quantifying memorization across neural language models. *ArXiv*, abs/2202.07646, 2022b. URL https://api.semanticscholar.org/CorpusID:246863735.

Tudor Cebere, Mathieu Even, Linus Bleistein, and Aurélien Bellet. Privacy auditing with zero (0) training run. *arXiv preprint arXiv:2605.14591*, 2026.

Debeshee Das, Jie Zhang, and Florian Tramèr. Blind baselines beat membership inference attacks for foundation models. In *Deep Learning and Security Workshop*, 2025.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL https://arxiv.org/abs/1810.04805.

Jinshuo Dong, Aaron Roth, and Weijie J Su. Gaussian differential privacy. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2022.

Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. Do membership inference attacks work on large language models? In *Conference on Language Modeling (COLM)*, 2024.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In Shai Halevi and Tal Rabin (eds.), *Theory of Cryptography*, pp. 265–284, Berlin, Heidelberg, 2006a. Springer Berlin Heidelberg. ISBN 978-3-540-32732-5.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, 2006b.

Mathieu Even, Clément Berenfeld, Linus Bleistein, Tudor Cebere, Julie Josse, and Aurélien Bellet. Causal evaluation of membership inference attacks. *arXiv preprint arXiv:2602.02819*, 2026.

Wenjie Fu, Huandong Wang, Chen Gao, Guanghua Liu, Yong Li, and Tao Jiang. Membership inference attacks against fine-tuned large language models via self-prompt calibration. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Mishaal Kazmi, Hadrien Lautraite, Alireza Akbari, Qiaoyue Tang, Mauricio Soroco, Tao Wang, Sébastien Gambs, and Mathias Lécuyer. PANORAMIA: Privacy auditing of machine learning models without retraining. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Zhan Li, Yongtao Wu, Yihang Chen, Francesco Tonin, Elias Abad Rocamora, and Volkan Cevher. Membership inference attacks against large vision-language models. *Advances in Neural Information Processing Systems*, 37:98645–98674, 2024.

Saeed Mahloujifar, Luca Melis, and Kamalika Chaudhuri. Auditing f-differential privacy in one run. *arXiv preprint arXiv:2410.22235*, 2024.

Pratyush Maini, Hengrui Jia, Nicolas Papernot, and Adam Dziedzic. Llm dataset inference: Did you train on my dataset? *Advances in Neural Information Processing Systems*, 37:124069–124092, 2024.

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

Matthieu Meeus, Shubham Jain, Marek Rei, and Yves-Alexandre de Montjoye. Did the neurons read your book? document-level membership inference for large language models. In *33rd USENIX Security Symposium (USENIX Security 24)*, pp. 2369–2385, 2024a.

Matthieu Meeus, Igor Shilov, Manuel Faysse, and Yves-Alexandre De Montjoye. Copyright traps for large language models. *arXiv preprint arXiv:2402.09363*, 2024b.

Matthieu Meeus, Igor Shilov, Shubham Jain, Manuel Faysse, Marek Rei, and Yves-Alexandre de Montjoye. Sok: Membership inference attacks on llms are rushing nowhere (and how to fix it). In *IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 385–401. IEEE, 2025. doi: 10.1109/SaTML64287.2025.00028.

Milad Nasr, Shuang Song, Abhradeep Thakurta, Nicolas Papernot, and Nicholas Carlini. Adversary instantiation: Lower bounds for differentially private machine learning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 866–882. IEEE, 2021. doi: 10.1109/SP40001.2021.00069.

Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. Scalable extraction of training data from (production) language models, 2023a. URL `https://arxiv.org/abs/2311.17035`.

Milad Nasr, Jamie Hayes, Thomas Steinke, Borja Balle, Florian Tramèr, Matthew Jagielski, Nicholas Carlini, and Andreas Terzis. Tight auditing of differentially private machine learning. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 1631–1648, 2023b.

Ashwinee Panda, Xinyu Tang, Christopher Choquette-Choo, Milad Nasr, and Prateek Mittal. Privacy auditing of large language models. In *International Conference on Learning Representations*, volume 2025, pp. 92555–92571, 2025.

Francesco Pinto, Nathalie Rauschmayr, Florian Tramèr, Philip Torr, and Federico Tombari. Extracting training data from document-based vqa models. *arXiv preprint arXiv:2407.08707*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Lorenzo Rossi, Bartłomiej Marek, Franziska Boenisch, and Adam Dziedzic. Natural identifiers for privacy and data audits in large language models. *arXiv preprint arXiv:2606.24408*, 2026.

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

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

Thomas Steinke, Milad Nasr, and Matthew Jagielski. Privacy auditing with one (1) training run. *Advances in Neural Information Processing Systems*, 36:49268–49280, 2023.

Marika Swanberg, Meenatchi Sundaram Muthu Selva Annamalai, Jamie Hayes, Borja Balle, and Adam Smith. A unified framework for adversary-aware differential privacy bounds. *arXiv preprint arXiv:2507.08158*, 2025.

Larry Wasserman and Shuheng Zhou. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 2010.

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

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

Weichao Zhang, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. Pretraining data detection for large language models: A divergence-based calibration method. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5263–5274, 2024.

A Extended Related Work

MIA on LLMs. A growing line of work has focused on extracting stronger membership signals for pretrained LLMs. Shi et al. (2023) propose Min-K%Prob that averages the log-likelihoods of the $k\%$ least-likely tokens, motivated by the observation that unseen texts are more likely to contain low-probability outlier tokens. Zhang et al. (2025) further propose Min-K%++, which normalizse each token probability with statistics of the categorical distribution over the whole vocabulary to better reflect the relative likelihood of the target token compared with other candidate tokens. Similarly, Zhang et al. (2024) addressses limitations of raw token likelihood by calibrating token probabilities with respect to token-frequency information. Beyond statistics computed from a single text sequence, other work derives membership signals by comparing the candidate text with auxiliary texts. ReCaLL (Xie et al., 2024) measures the relative change in a candidate's log-likelihood when it is conditioned on a prefix constructed from non-member samples, compared with its unconditional likelihood. Mattern et al. (2023) generate neighbouring texts through word replacement and compare the candidate text loss against the mean loss over its neighbours. Related work has also considered other LLM membership inference settings, such as fine-tuned LLMs (Fu et al., 2024), document or dataset-level membership inference (Meeus et al., 2024a;b), and MIA on vision-language models (VLLMs) (Li et al., 2024).

Distribution-shift confounding in MIA. Recent work has highlighted that strong MIA performance does not necessarily reflect training-data leakage. In many post-hoc MIA benchmarks, held-out non-member samples from the original LLM training distribution are often unavailable, and members and non-members are collected from different time periods or data sources. As a result, membership labels can be confounded by substantial distributional differences between the data populations irrespective of the target model (Shi et al., 2023). Duan et al. (2024) shows temporal differences between members and non-members can substantially increase MIA performance, whereas MIAs are generally weaker on more controlled member and non-member samples when they are from the same distribution. Das et al. (2025) demonstrates the confounding directly by showing that classifiers trained to predict membership from the samples alone, without access to the target model, can outperform target-model-based MIAs on several commonly used benchmarks. Meeus et al. (2025) systematically identifies similar distribution shifts across existing LLM MIA benchmarks and advocate randomized evaluation when possible, or explicit post-hoc controls when randomized evaluation is unavailable. Maini et al. (2024) similarly observes that sample-level MIAs perform close to random guessing when members and non-members are drawn from the same distribution, and consequently shift the focus toward aggregating weak membership signals for dataset-level inference.

Privacy auditing on LLMs. Privacy auditing builds on MIA and uses the success rates of membership inference to obtain statistically meaningful measurement of privacy leakage, such as a lower bound on the privacy parameter ϵ (Nasr et al., 2023b; Steinke et al., 2023; Mahloujifar et al., 2024). The standard way to audit LLMs is to construct auditing samples before model training. For instance, Panda et al. (2025) designes canaries that are highly susceptible to memorization and insert them into LLM training, allowing strong MIA performance to be converted into empirical lower bounds on privacy leakage. Such controlled audits provide well-defined member and non-member datasets, but require intervention in the training procedure and therefore cannot generally be applied post-hoc to an already trained model. Recent work has also explored post-hoc auditing settings in which suitable non-members can be obtained without retraining samples prior to training. Rossi et al. (2026) proposes the generation of non-members using natural identifiers such as structured random strings that already occur in training corpora. The approach avoids the member–non-member distribution-shift problem by construction, but applies to special classes of data for which such matched non-members can be generated.

Privacy auditing methods accounting for distribution shifts. Recent work by Swanberg et al. (2025) introduces a general framework for bounding an adversary's success rate against a differentially private mechanism under non-uniform prior knowledge. This prior relates to a baseline attack success probability without observing the mechanism and is conceptually close to our distinction between data-only distinguishability of members and non-members and additional information revealed by the target model. However, their framework does not directly provide a post-hoc auditing mechanism and assumes that the relevant adversarial prior is known.

Although these methods have not yet been applied to auditing LLMs under distribution shift, a closely related line of work develops privacy auditing methods that account for distribution shifts between member and non-member populations. PANORAMIA (Kazmi et al., 2024) addresses such shifts by generating synthetic non-members and calibrating the membership signals against a data-only baseline. It trains a target-model-independent baseline MIA to quantify membership signals from data only, and compares it with an MIA that additionally accesses the target model to estimate the combined signal. The difference between these quantities provides a calibrated estimate of privacy leakage while accounting for distributional differences introduced by the generated non-members. A complementary line of work approaches the same problem from a causal perspective, by treating data inclusion in training as a treatment and the resulting MIA performance as observational results, they account for confounding from distribution shift through propensity-score estimation and inverse-probability weighting (Even et al., 2026; Cebere et al., 2026). Similar to PANORAMIA, the method require sufficient overlap between the member and non-member distributions and an accurate estimate of the propensity of membership from the data alone. Even et al. (2026) uses propensity-based reweighting to evaluate MIA performance under a reweighted non-member population that more closely matches the member population. Cebere et al. (2026) further extends it to privacy auditing, in which the estimated propensity score for each sample determines an abstention probability that remove the contribution of MIA guesses that can be explained by large

distributional differences, and the uncertainty in the estimated propensity scores is accounted for through bootstrapping.

B EXTENDED BACKGROUND

Our practical extensions build on two existing frameworks for accounting distribution shift in post-hoc privacy auditing, namely PANORAMIA (Kazmi et al., 2024) and causal auditing (Cebere et al., 2026). We summarize their theoretical framework for testing violations against ϵ-differential privacy (DP) target model.

A randomized training algorithm M is ϵ-differential privacy (DP) (Dwork et al., 2006a) if for any pair of datasets (D, D') differing by one record and every measurable outcome E,

$$\Pr[M(D) \in E] \leq e^{\varepsilon} \Pr[M(D') \in E].$$

Consequently, a membership inference attack (MIA) distinguishing these neighboring datastes must satisfy $\mathrm{TPR} \leq e^{\varepsilon}\mathrm{FPR}$ Nasr et al. (2021). An audit tests such restrictions to reject candidate ϵ privacy guarantees.

Steinke et al. (2023) show that privacy auditing can be performed with a single training run by independently including each auditing example X_i, $i \in [m]$ with probability $1/2$. Following their notation, let S be uniform on $-1, +1^m$, encoding exclusion and inclusion, and let $T = M(S) \in \mathbb{R}^m$ be the auditor's guesses. Positive and negative values predict inclusion and exclusion AND zero denotes abstention. When M satisfy ϵ-DP, Proposition 5.1 (Steinke et al., 2023) states that, for every $v \in \mathbb{R}$ and almost every realized t,

$$\Pr\left[\sum_{i=1}^{m} \max 0, T_i S_i \geq v, \middle| , T = t\right] \leq \Pr_{\check{S} \sim \mathrm{Bernoulli}\left(\frac{e^{\varepsilon}}{1+e^{\varepsilon}}\right)^m}\left[\sum_{i=1}^{m} |t_i| \check{S}_i \geq v\right],$$

i.e. the number of successful guesses is stochastically dominated by a sum of independent Bernoulli variables. Evaluating the right-hand tail at the observed statistic gives a test p-value and performing hypothesis test yields a lower confidence bound on ϵ. This framework requires control over inclusion of audit samples before training. PANORAMIA (Kazmi et al., 2024) and causal auditing (Cebere et al., 2026) address the post-hoc setting, where the model is already trained and the auditor observes known members and non-members.

PANORAMIA (Kazmi et al., 2024) pairs known members with independently generated non-members and independently selects one example from each pair using a fair coin. Let $\mathcal{D}, \mathcal{G}$ denote respectively the member and non-member distribution in which $p_{\mathcal{D}}, p_{\mathcal{G}}$ is their probability density function. The one-sided c-closeness assumption requires

$$p_{\mathcal{G}}(x) \leq e^{c} p_{\mathcal{D}}(x).$$

The condition is one-sided because PANORAMIA only predicts member or abstains, as high confidence in correctly identifying members better align with model memorization signals Carlini et al. (2022a). Let $T_i^b, T_i^a \in 0, 1$ denote discriminator and MIA predictions where 1 predicts membership and 0 abstains, and the discriminator uses only the data, whereas the MIA additionally accesses the target model θ. Given observed predictions t^b, t^a and membership labels s, define $r_j = \sum_{i=1}^{m} t_i^j, v_j = \sum_{i=1}^{m} t_i^j s_i, j \in b, a$ as the numbers of predictions and correct predictions respectively. Under PANORAMIA's privacy game (Definition 2 (Kazmi et al., 2024)), the discriminator tests a candidate closeness parameter c, while the MIA tests a candidate combined parameter $c + \varepsilon$. Propositions 1 and 2 (Kazmi et al., 2024) provide the following tests:

$$\mathbb{P}_{S,X,T^b}\left[\sum_{i=1}^{m} T_i^b S_i \geq v \,\middle|\, T^b = t^b\right] \leq \mathbb{P}_{S' \sim \mathrm{Bernoulli}\left(\frac{e^c}{1+e^c}\right)^m}\left[\sum_{i=1}^{m} t_i^b S_i' \geq v\right]$$

$$\mathbb{P}_{S,X,T^a}\left[\sum_{i=1}^{m} T_i^a S_i \geq v \,\middle|\, T^a = t^a\right] \leq \mathbb{P}_{S' \sim \mathrm{Bernoulli}\left(\frac{e^{c+\varepsilon}}{1+e^{c+\varepsilon}}\right)^m}\left[\sum_{i=1}^{m} t_i^a S_i' \geq v\right].$$

Because the predictions are binary, these are equivalently the upper tails of the Binomial distribution. Let α denote the overall significance level, allocated across the discriminator and MIA tests and any

searched prediction thresholds. Under the framework's assumptions, $c_{\text{lb}} \leq c$ and $(c+\varepsilon)_{\text{lb}} \leq c+\varepsilon$ then hold simultaneously with probability at least $1-\alpha$. PANORAMIA estimates privacy leakage attributable to target model by,

$$\widetilde{\epsilon}_{\text{panoramia}} = \max\{0, (c+\varepsilon)_{\text{lb}} - c_{\text{lb}}\}.$$

The difference of lower bounds is a calibrated privacy leakage estimate instead of a lower confidence bound on ϵ. Its interpretation depends on how tightly c_{lb} captures the actual distribution shift. A stronger baseline can improve this correction, but even an optimal classifier does not eliminate finite-sample uncertainty. An upper bound on c would instead permit a valid lower bound on ϵ.

Causal auditing (Cebere et al., 2026) uses per-sample propensity score $\pi_i = P(\text{member}|X_i)$ to captures how predictable membership is from the data alone. Values near $1/2$ indicate greater overlap between members and non-members, whereas values near 0 or 1 indicate stronger data-only membership information. Let c_i denote the MIA score for sample $i \in [m]$, causal auditing rank predictions by their propensity-adjusted confidence $c_i \times b_i$, where

$$b_i = \frac{\min(\pi_i, 1-\pi_i)}{\max(\pi_i, 1-\pi_i)}.$$

The distribution-shift correction then independently samples $B_i \mid X \sim \text{Bernoulli}(b_i)$ and result in $T_i' = B_i T_i$ where MIA prediction T_i is abstained $B_i = 0$. The hypothesis test evaluates the number of correct retained predictions under the total number of guesses r. Examples whose membership is highly predictable from the data alone are assigned a higher probability of abstention, reducing their contribution to the privacy audit even when the MIA scores is high. Assuming i.i.d. data-generation model and known oracle propensity scores, Theorem 5.1 (Cebere et al., 2026) gives a test for model satisfying ϵ-DP,

$$\Pr\left[\sum_{i=1}^{m} B_i \mathbf{1}\{T_i = S_i\} \geq v \middle| (X_i)_{i\in[m]}\right] \leq \Pr_{Z\sim\text{Binom}\left(r, \frac{e^\varepsilon}{1+e^\varepsilon}\right)}[Z \geq v].$$

The retention weights satisfy the theorem's sufficient condition

$$b_i \leq \frac{1+e^{-\epsilon-\epsilon_{\text{DS}}(X_i)}}{1+e^{-\epsilon}}, \quad \text{for every } \epsilon \geq 0,$$

in which $\epsilon_{\text{DS}}(X_i)$ measures the strength of the membership signal available from the data alone. With known propensities, hypothesis testing results in a lower bound on ϵ. In practice, the propensities are unknown and are estimated by a discriminator. Substituting these estimates does not automatically preserve the lower bound guarantee. Causal auditing therefore estimates propensity-estimation uncertainty from finite data through a bootstrap procedure. Let α denote the significance level of the main ϵ-DP hypothesis test. Using the observed audit set and estimated propensity scores, the causal audit reports $\widetilde{\epsilon}_{\text{causal}}$. Each bootstrap run resamples the propensity-fitting, refits the discriminator, and recomputes the propensity correction and privacy estimate, where the target model, audit examples, and MIA scores remain fixed from the observed finite sample audit set. The lower α'-quantile of these bootstrap estimates, denoted by $\widetilde{\epsilon}^{q}_{\text{causal}}$, is reported with an intended confidence level of $1-\alpha-\alpha'$. This bootstrap interpretation is asymptotic and subject to appropriate propensity-estimation and bootstrap regularity conditions (Section 6, (Cebere et al., 2026)).

C PANORAMIA WITH POINTWISE C-CLOSENESS

Definition C.1 (Pointwise closeness)**.** Let $\mathcal{D}$ denote the member distribution and $\mathcal{G}$ the non-member distribution, with probability density functions $p_{\mathcal{D}}$ and $p_{\mathcal{G}}$, respectively. For any $x \in \mathcal{X}$ such that $p_{\mathcal{G}}(x) > 0$, define the *pointwise closeness parameter*

$$c(x) \triangleq \log \frac{p_{\mathcal{D}}(x)}{p_{\mathcal{G}}(x)}.$$

Equivalently, $c(x)$ is the smallest scalar satisfying $e^{-c(x)} p_{\mathcal{D}}(x) \leq p_{\mathcal{G}}(x)$. A function $c^{ub} : \mathcal{X} \to \mathbb{R}$ is a *pointwise closeness dominating function* if $c(x) \leq c^{ub}(x),\ \forall x \in \mathcal{X}$.

We next connect pointwise closeness to the c-closeness definition in PANORAMIA (Definition 3 of Kazmi et al., 2024).

Proposition C.2 (c-closeness is a pointwise worst case)**.** *Let $\mathbf{x} = (x_1, \ldots, x_m)$ be a realized audit set consisting of m samples. Then $c_{\max}(\mathbf{x}) \triangleq \max_{x \in \mathbf{x}} c(x)$, is the smallest scalar c satisfying the closeness inequality simultaneously for all realized audit samples. That is, $c_{\max}(\mathbf{x}) = \inf \{c \in \mathbb{R} : e^{-c} p_{\mathcal{D}}(x) \leq p_{\mathcal{G}}(x),\ \forall x \in \mathbf{x}\}$.*

Proof. By Definition C.1, for every $x_i \in \mathbf{x}$, $c(x_i) = \log \frac{p_{\mathcal{D}}(x_i)}{p_{\mathcal{G}}(x_i)}$. For any scalar c, $e^{-c} p_{\mathcal{D}}(x_i) \leq p_{\mathcal{G}}(x_i)$ is equivalent to $c(x_i) \leq c$. Thus, a single scalar c satisfies the closeness inequality for all realized audit examples if and only if $c \geq c(x_i)$ for every $x_i \in \mathbf{x}$. The smallest such scalar is therefore $c_{\max}(\mathbf{x}) = \max_{x \in \mathbf{x}} c(x)$. □

The pointwise closeness parameter $c(x)$ is directly related to the propensity score used in causal privacy auditing. PANORAMIA constructs its audit set through a privacy game (Definition 2 (Kazmi et al., 2024)), in which $S_i \sim \text{Bernoulli}\left(\frac{1}{2}\right)$ independently for every sample i, and the auditor sees either a member or non-member given the result of the Bernoulli coin flip, $X_i \mid S_i = 1 \sim \mathcal{D}$, $X_i \mid S_i = 0 \sim \mathcal{G}$. We show that PANORAMIA with pointwise closeness and the causal propensity score can be seen as two characterizations on the same data-only likelihood ratio $p_{\mathcal{D}}/p_{\mathcal{G}}$. The c-closeness definition in PANORAMIA can be seen as collapsing the pointwise closeness quantities into a single constant bound shared across all samples in the audit set, whereas causal auditing retains the pointwise values $\pi(x)$ and corrects each membership predictions separately.

Proposition C.3 (Propensity upper-bound implies point-wise and global closeness upper-bound)**.** *Let $\pi^{ub}(x) \geq \pi(x) \triangleq P(S = 1 \mid X = x)$ denote the oracle propensity function and its upper-bound, and let $c(x) \triangleq \log \frac{p_{\mathcal{D}}(x)}{p_{\mathcal{G}}(x)}$ denote the same for point-wise closeness. If $p_{\mathcal{D}}(x), p_{\mathcal{G}}(x) > 0$:*

$$c(x) = \log \frac{\pi(x)}{1 - \pi(x)} \leq \log \frac{\pi^{ub}(x)}{1 - \pi^{ub}(x)} \triangleq c^{ub}(x).$$

Proof. Under the PANORAMIA audit game, $P(S = 1) = P(S = 0) = 1/2$. By Bayes' rule,

$$\begin{aligned} \pi(x) &= P(S = 1 \mid X = x) \\ &= \frac{p_{\mathcal{D}}(x) P(S = 1)}{p_{\mathcal{D}}(x) P(S = 1) + p_{\mathcal{G}}(x) P(S = 0)} \\ &= \frac{p_{\mathcal{D}}(x)}{p_{\mathcal{D}}(x) + p_{\mathcal{G}}(x)}. \end{aligned}$$

Therefore,

$$\frac{\pi(x)}{1 - \pi(x)} = \frac{p_{\mathcal{D}}(x)}{p_{\mathcal{G}}(x)},\ c(x) = \log \frac{p_{\mathcal{D}}(x)}{p_{\mathcal{G}}(x)} = \log \frac{\pi(x)}{1 - \pi(x)}.$$

Since the logit function $\log \frac{u}{1-u}$ is monotonically increasing on $(0, 1)$ and $\pi(x) \leq \pi^{ub}(x)$,

$$c(x) \leq \log \frac{\pi^{ub}(x)}{1 - \pi^{ub}(x)} = c^{ub}(x).$$

Since the inverse of logit function is also monotone, any pointwise upper bound $c^{ub}(x) \geq c(x)$ induces $\pi^{ub}(x) = \frac{e^{c^{ub}(x)}}{1 + e^{c^{ub}(x)}} \geq \pi(x)$. □

Given a valid pointwise upper bound $\pi^{ub}(x)$ on the propensity, Proposition 2.2 gives the corresponding pointwise closeness upper bound. Proposition C.4 then shows that these pointwise bounds can be incorporated into PANORAMIA's hypothesis test, yielding a finite-sample valid privacy lower bound. In contrast, standard PANORAMIA does not use the pointwise values and it estimates a constant lower bound c_{lb} from the discriminator's predictions. Therefore even with an oracle discriminator, c_{lb} need not equal the true closeness parameter at finite sample size. As the number of audit samples increases, the hypothesis test gains power and the statistical gap can vanish asymptotically.

Proposition C.4 (Privacy auditing with known pointwise closeness upper bounds)**.** *Let c^{ub} be a pointwise closeness upper-bound function for $\mathcal{D}$ and $\mathcal{G}$, let S, X be the random variables corresponding to s, x in the PANORAMIA audit game (Definition 2), and let $T^a \triangleq A(S, X, f)$ denote the vector of guesses from MIA against an ϵ-DP model f. Then, for any realized $x = (x_1, \ldots, x_m)$ in the support of X, for all $v \in \mathbb{R}$ and all t^a in the support of T^a,*

$$\Pr_{S,X,T^a}\left[\sum_{i=1}^{m} T_i^a S_i \geq v \,\middle|\, X = x,\, T^a = t^a\right] \leq \Pr_{S'}\left[\sum_{i=1}^{m} t_i^a S_i' \geq v\right],$$

where $S_i' \sim \text{Bernoulli}(p_i(\epsilon))$, $p_i(\epsilon) \triangleq \frac{e^{\epsilon + c_i^{ub}}}{1+e^{\epsilon+c_i^{ub}}}$, $c_i^{ub} = c^{ub}(x_i)$. If $t_i^a \in \{0, 1\}$, then $\sum_i t_i^a S_i'$ follows a Poisson-binomial distribution with parameters $\{p_i(\epsilon)\}_{i=1}^m$.

Proof. The proof follows Proposition 2 of Kazmi et al. (2024) by replacing the c-closeness condition by the pointwise inequality $\frac{P_{\mathcal{G}}[x_i]}{P_{\mathcal{D}}[x_i]} \geq e^{-c_i^{ub}}$. For every $i \in [m]$, $s_{<i}$, and $x_{\leq i}$, the same conditional independence and Bayes rule argument gives,

$$\Pr[S_i = 1 \,|\, T^a = t^a,\, S_{<i} = s_{<i},\, X_{\leq i} = x_{\leq i}] \leq \frac{1}{1 + e^{-\epsilon - c_i^{ub}}} = \frac{e^{\epsilon + c_i^{ub}}}{1 + e^{\epsilon + c_i^{ub}}} \triangleq p_i(\epsilon).$$

Since c_i^{ub} depends on x_i, we applying the law of total probability over the past observations $X_{<i}$ only,

$$\begin{aligned}
&\Pr[S_i = 1 \mid T^a = t^a,\, S_{<i} = s_{<i},\, X_i = x_i] \\
&= \sum_{x_{<i}} \Pr[S_i = 1 \mid T^a = t^a,\, S_{<i} = s_{<i},\, X_{\leq i} = (x_{<i}, x_i)] \\
&\qquad \cdot \Pr[X_{<i} = x_{<i} \mid T^a = t^a,\, S_{<i} = s_{<i},\, X_i = x_i] \\
&\leq p_i(\epsilon).
\end{aligned}$$

Condition on $X = x$ and $T^a = t^a$, the same stochastic dominance argument applies, with the common Bernoulli parameter $\frac{e^{\epsilon+c}}{1+e^{\epsilon+c}}$ replaced by $p_i(\epsilon)$ for each i. Consequently, $\sum_{i=1}^m t_i^a S_i$ is dominated by $\sum_{i=1}^m t_i^a S_i'$, where $S_i' \sim \text{Bernoulli}(p_i(\epsilon))$ and result in the tail-probability inequality. □

D Implementation Details

D.1 Discriminator and MIA for the image-caption data modality

Figure 3 presents an overview of our proposed Discriminator/MIA for the image-caption data modality.

D.2 Target-Model Features for Membership Inference

We extract a set of loss-based features from the target model for each input sequence. Given the per-token negative log-likelihoods (NLLs), we compute the mean and maximum NLL as two global summary statistics. We additionally include a compression-normalized score, which is the mean NLL losses divided by the zlib-compressed length of the text, which adjusts the sequence loss by textual compressibility (Carlini et al., 2021). To characterize the high-loss tail of the token-level distribution, we use four Min-K++ style features at $k \in \{10, 20, 40, 60\}$, which are the mean NLLs over the top-$k\%$ highest-loss tokens, with the token score normalized relative to the full vocabulary logit distribution at each position, and the least-confident $k\%$ of tokens under this normalized score are aggregated (Zhang et al., 2025). Finally, we include five quantiles of the per-token NLL distribution, corresponding to the 10th, 25th, 50th, 75th, and 90th percentiles. Together, these features capture both the overall likelihood of the sequence and the shape and tail behaviour of its token-level loss distribution. For each raw feature, we compare the value on the original text with the corresponding values obtained from its K auxiliary-LLM paraphrases and compute standard z-score and robust z-score, each measuring how anomalous the original example is relative to its own local paraphrase distribution for the same target-model feature.

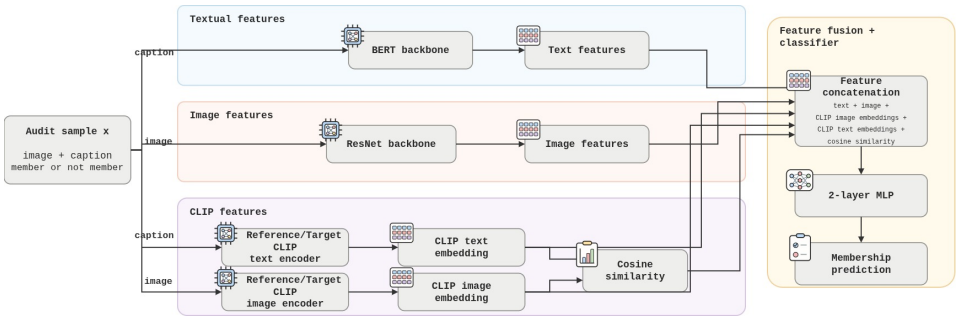


Figure 3: **Overview of the discriminator and MIA architecture for the Image-caption modality.** Audited samples are passed through pre-train ResNet and Bert in order to extract image and textual features respectively. Audit samples are also passed in either a reference CLIP model for the discriminator or the audited target model for the MIA. Cosine similarity is computed between image and text embeddings. Extracted textual and image features, textual and image embeddings as well as cosine similarity are concatenated together and passed through a 2-layer MLP classifier to obtain the membership prediction.

Discriminator and MIA evaluation protocol. For smaller datasets, we use cross-validation to obtain out-of-fold membership scores for every sample in the audit set. Each sample is predicted only by a classifier trained on the remaining folds, and predictions from all held-out folds are concatenated to evaluate MIA performance over the full set. For sufficiently large datasets, we use a standard train–test split and evaluate on the held-out test set. For each dataset, the data-only baseline and target-model-accessing MIA use the same data splits and evaluation protocol. For the causal audit bootstrap we use the same strategy, cross-fitting models for smaller dataset and resampling a fresh training and validation/callibration set for each bootstrap.

Bonferroni correction to compute audit values Once the Discriminator or MIA model is trained, the auditor needs to choose a set of r positive guesses. For PANORAMIA we apply the original method suggested the authors, ranking all guesses and testing for all possible subset of r-guesses and apply Bonferroni correction to account for multiple tests.

For the causal audit, we also follow the implementation suggested by Cenbere and co-authors Cebere et al. (2026), sorting sample according to their prediction but only test for a few values of r. Table 6 presents the grid of values used to conduct the causal audit. We increase the value of the r-guesses as the audit dataset grows. Rateher than using the membership guesses alone to rank the audit sample we employ the proposed adjusted confidence by fitting an initial discriminator to obtain initial propensity score: $b_i^{\text{initial}} \times t_i$.

Table 6: Grid of guess budgets r tried by the causal audit.

Dataset	Grid of r
Gutenberg	{10,50,100,150,. . . ,250}
WikiMIA-32	{10,50,100,150}
BookMIA	{10,50,100,150,. . . ,450}
Temporal-Wiki	{10,50,100,150,. . . ,450}
Temporal-Arxiv	{10,50,100,150,. . . ,250}
CC12M	$\{50, 100, \ldots, 500\}$
LAION (raw captions)	$\{200, 400, \ldots, 2000\}$
LAION (rewritten)	$\{200, 400, \ldots, 2000\}$
LAION (rewritten, keep 75%)	$\{150, 300, \ldots, 1500\}$
LAION (rewritten, keep 50%)	$\{100, 200, \ldots, 1000\}$
LAION (rewritten, keep 25%)	$\{50, 100, \ldots, 500\}$

Propensity score model training for image-caption data modality. The causal audit is quite sensible to variation of the propensity score during the bootstrap step. In order to increase stability of the estimates propensity score we replace individual model training by ensembling 5 models on the same data subset and taking the average of the predicted propensity score. We use this ensemble modelling for both the initial discriminator training and the bootstrapping phase.

E MORE EXPERIMENT RESULTS

Strength of discriminator. The discriminator plays an important role in both PANORAMIA and causal auditing, in which PANORAMIA relies on it to obtain a tight estimate of c_{lb} and the causal audit uses the same data-only signal to estimate the pointwise propensity scores. Therefore, we compare several discriminator architectures to verify that our experiments use a sufficiently strong data-only classifier. Table 7 shows the MIA performance and the resulting c_{lb} under different discriminator. For text benchmarks, we include TF-IDF and Bag-of-Words classifiers following (Das et al., 2025; Meeus et al., 2025), together with pretrained BERT encoders equipped with a classification head. On Temporal-Wiki, the TF-IDF and Bag-of-Words baselines achieve AUCs between 0.63 and 0.65 and yield $c_{\mathrm{lb}} = 0$, whereas a frozen BERT encoder improves the AUC to 0.6969 and gives $c_{\mathrm{lb}} = 0.1394$. Jointly fine-tuning BERT with the classifier head performs best, reaching an AUC of 0.7173 and the largest $c_{\mathrm{lb}} = 0.1628$. We therefore adopt the trainable BERT classifier in all our text benchmark experiments. For image caption benchmark LAION, we combine both text and vision representations to strengthen the baseline, with c_{lb} increasing from 1.623 using BERT alone and 1.521 using ResNet50 alone to 1.902 with BERT+ResNet and 2.388 with BERT+ResNet+CLIP, which we use in the VLM experiments. Overall, these results confirm the strength of our baselines to better capture data-only membership signals and resulting in tighter c_{lb} estimates for PANORAMIA and more informative propensity-score estimates for the causal audit.

Table 7: **Strength of discriminator.** We compare performance of different discriminator on Temporal-Wiki and LAION. Stronger discriminators produce tight c_{lb} and more accurate propensity scores. On Temporal-Wiki, trainable BERT variants outperform simpler lexical baselines, while on LAION, jointly using image, text, and CLIP cross-modal features gives the strongest discriminator and the largest c_{lb}.

Dataset	Baseline Classifier	TPR@FPR1%	TPR@FPR5%	TPR@FPR10%	AUC	c_{lb}
Temporal-wiki	TF-IDF + LR	0.0100	0.0580	0.1760	0.6512	0.0
	TF-IDF + RF	0.0	0.0420	0.1460	0.6349	0.0
	Bag of Words + LR	0.0060	0.0560	0.1640	0.6427	0.0
	Bag of Words + RF	0.0	0.0560	0.1620	0.6454	0.0
	Frozen Bert + clf head	0.0320	0.1260	0.2500	0.6969	0.1094
	Trainable Bert + clf head	0.0580	0.1760	0.2960	0.7173	0.1166
	Trainable Bert + large clf head	0.0940	0.1720	0.2440	0.6561	0.1032
LAION	Bag of Words + RF	0.036	0.123	0.205	0.630	0.906
	BERT only	0.071	0.194	0.293	0.677	1.623
	ResNet50 only	0.060	0.177	0.276	0.666	1.521
	BERT + ResNet	0.091	0.248	0.361	0.728	1.902
	BERT + ResNet + CLIP (ours)	0.129	0.315	0.444	0.777	2.388

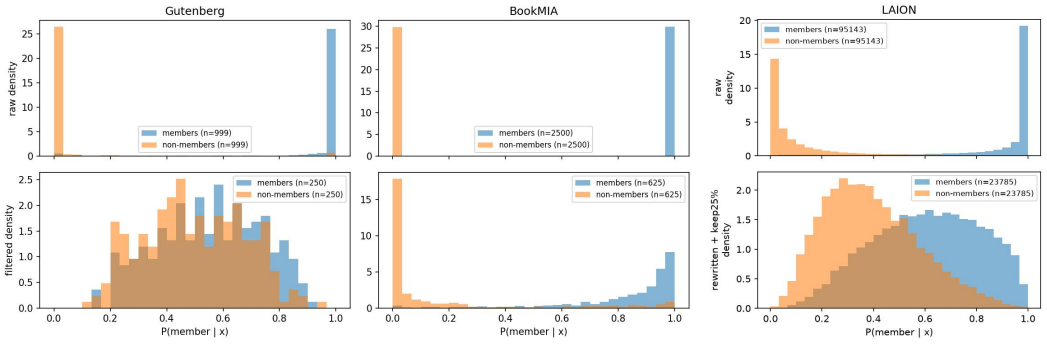


Figure 4: **Discriminator score distributions before and after rewriting and filtering.** We train a data-only discriminator to predict membership on the original benchmark data and on the constructed audit sets. On the original data, member and non-member scores are strongly separated with members generally receiving higher scores. After rewriting and filtering, the score distributions overlap substantially more, indicating reduced data-only distribution shift.

Discriminator score distributions. Figure 4 compares the discriminator scores on the original benchmarks and after rewriting and filtering. On the original data, members and non-members are

strongly separated, with their score distributions concentrated toward opposite ends of the range. This verifies the empirical pattern underlying our filtering rule. When members generally receive higher discriminator scores, the lower tail of the member distribution and the upper tail of the non-member distribution correspond to the region of greatest overlap. After rewriting and filtering, the two distributions overlap substantially more, indicating that readily exploitable data-only differences are reduced. The extent of this reduction varies across datasets. For more challenging benchmarks such as BookMIA, substantial residual separation remains, consistent with the smaller adjusted privacy-leakage estimates we obtained, which the discriminator can still explain a large fraction of the observed membership signal.

Histogram of causal audit estimates. Figure 5 shows the distribution of $\widetilde{\epsilon}_{\text{causal}}$ across 100 bootstrap replicates, obtained by resampling the propensity-estimation data and refitting the propensity model. The black and red vertical lines indicate the estimate on the observed data (the candidate set we obtain after rewriting and filtering) and the lower 2.5% bootstrap quantile $\widetilde{\epsilon}^{q0.025}_{\text{causal}}$ respectively. We observe positive estimates in many bootstrap replicates even when $\widetilde{\epsilon}^{q0.025}_{\text{causal}} = 0$. For smaller benchmarks such as WikiMIA, the limited sample size leads to a more discrete distribution of audit estimates and with a large mass at zero.

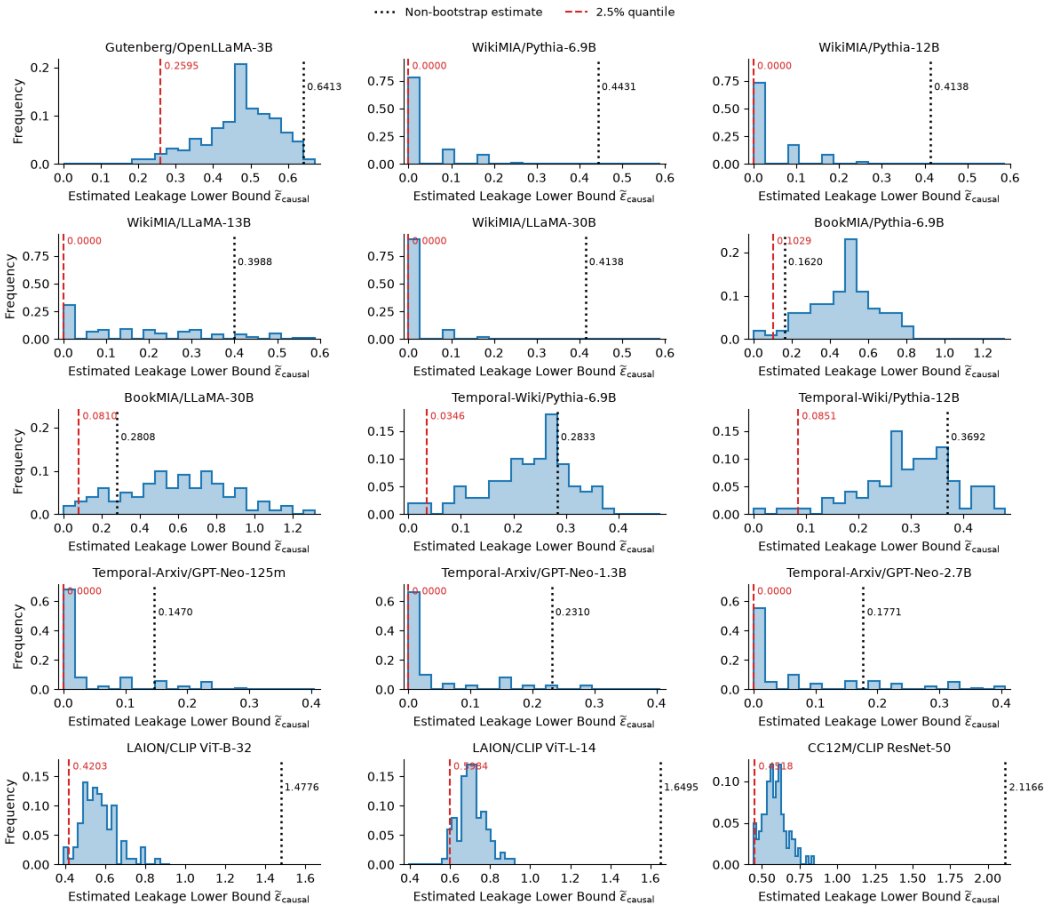


Figure 5: **Distribution of $\widetilde{\epsilon}_{\text{causal}}$ estimates.** Computed across 100 bootstrap experiments when resampling from the observed audit set. The black dotted line denotes the non-bootstrap estimate, while the red dashed line denotes the lower 2.5% quantile of the bootstrap distribution.

F PROMPT SAMPLE

We show two example prompts used for Gutenberg, WikiMIA/Temporal-Wiki and LAION respectively, where an auxiliary LLM rewrites non-member text to match the style of member text. For text rewriting and paraphrasing, we use `Qwen2.5-7B-Instruct`, `Qwen3-30B-A3B-Instruct-2507-FP8` and `llama-4-Scout`.

Rewriting prompt for the Gutenberg dataset

```
You are a careful text editor. You rewrite short passages so that
    they match the SURFACE STYLE of a reference passage (typography,
    punctuation, header/footer boilerplate, era-specific orthography,
    archaic spellings, scanned-OCR artifacts, chapter/section labels)
    while fully preserving the SEMANTIC CONTENT of the passage you
    are rewriting. Do not add, remove, or paraphrase the underlying
    facts, characters, plot, or arguments. Match formatting quirks
of the reference: line break patterns, use of asterisks,
    header/footer wording if present, dashes, quote styles. Output
    only the rewritten passage, with no preamble, commentary, or
    markdown fences.
%\end{lstlisting}
```

Rewriting prompt for WikiMIA/Temporal-Wiki datasets

```
You are a careful Wikipedia editor rewriting a snippet so that it
    could plausibly have been written before 2017. The reference
    passage is an authentic pre-2017 Wikipedia snippet. The passage
    to rewrite is a more recent Wikipedia snippet. Your job is to
    rewrite it so that, looking at vocabulary and content alone, a
    reader cannot tell whether it was written before 2017 or after.

Specific instructions:

  - Remove or genericize references to people, organizations,
    products, technologies, events, and terms that did not exist or
    were not notable before 2017. Replace specific recent entities
    with plausible pre-2017 equivalents OR with generic descriptions
    (e.g. 'a new smartphone model' instead of a 2024 product name).

  - Avoid post-2017 buzzwords, slang, and jargon. Prefer vocabulary
    that was standard on Wikipedia before 2017.

  - Match the reference passage's tone, sentence rhythm, and use of
    citations/templates if any.

  - Preserve the overall topic and factual claims at the abstract
    level. Do not invent specific dates or specific people.

  - If the passage describes an event whose occurrence after 2017 is
    essential to its meaning, rewrite it as a hypothetical or
    describe a comparable pre-2017 event in the same domain.

Output only the rewritten passage with no preamble, commentary,or
    markdown fences.
```

Rewriting prompt for LAION dataset

```
You are a caption style-transfer tool. LAION captions are raw,
    unfiltered alt-text scraped from the open web  overwhelmingly
```

```
    e-commerce product listings, SEO/Pinterest keyword strings,
    stock-photo pages, news/editorial photos, and blog titles. They
    read like product titles and page titles, NOT like clean
    descriptive sentences. Your job is to rewrite CC12M captions so
    they look like authentic LAION web scrapes.

You are shown the image the caption belongs to. Use it only to ground
    what is actually in the picture (and to resolve any person
    placeholder); do not start describing details the source caption
    never mentioned.

Here are 30 real LAION captions  study their raw, product-listing
    quality:

1. The Legacy Study Bible  Black Bonded Leather
2. Simple-Long-Ankara-jacket-outfit
3. Keys to the Elementary Classroom
4. Kate Bosworth - SS-GB S01 E03 1080p nightwear scene
5. ET2 Minx 3-Light Rapidjack Pendant And Canopy
6. Star Wars Force and Destiny Ghosts of Dathomir
7. Peach Mandarin Neck Embroidered Kurta
8. Parliament Square
9. Items for sewing or DIY Stock Photos
10. 2014-12-17 FB Ralph Green Signs with Northwestern State
11. lined paper word excel pdf documents  kindergarten hand writing
    lined paper pdf
12. Boy eating a green Peas Royalty Free Stock Photo
13. Microsoft Certified Power Platform Fundamentals Exams
14. A forestry worker sawing a tree trunk.  Stock Photo
15. La Roche-Posay Toleriane Purifying Foaming Cream (4.22 fl oz/ 125
    ml)
16. Nightmares Fear Factory
17. "Rubbermaid FG354000BEIG 23-gal Rectangle Slim Trash Can, 20""L x
    11""W x 30""H, Beige"
18. JDX Abstract Cushions Cover
19. Jennette McCurdy, Nathan Kress & Jerry Trainor Talk 'iCarly' at
    the Creative Arts Emmy Awards 2011
20. Plus One | [Brighton Walsh]
21. Allen Fontenot & The Country Cajuns
22. Landscape with volcano. Moor. National Natural Park snow. Andean
    Royalty Free Stock Images
23. Wall Tile Addition in Pink
24. Mario Chalmers Calls Out Lying Girls on Twitter Who Claimed They
    Drank with Him
25. 2019 McLaren 720S Supercar 1:32 Scale Model Car Diecast Toy
    Vehicle Red Kids
26. Bravo - Homestyle Complete Turkey - Freeze Dried Dog Food - 2 lb
27. Image of John Barrowman
28. dirty street with garbage all... | Shutterstock .eps vector
    #1090782695
29. 2015 calendar. Vector 2015 calendar on a grey background Royalty
    Free Stock Images
30. Obama, with Sasha, center, and Malia, pays for a purchase at the
    Politics & Prose bookstore in northwest Washington in November
    2013.

CRITICAL RULES  read carefully before transforming:

DO NOT write clean descriptive sentences ("A man riding a horse on a
    beach."). LAION captions are titles/fragments, not descriptions.
DO NOT add editorial adjectives the original doesn't imply
    ("breathtaking", "majestic", "stunning", "vibrant",
    "picturesque", "showcasing").
```

```
DO NOT end captions with a period. Almost no LAION caption ends with
    a period. Match the examples.
DO NOT expand short captions. If the original is 5 words, the output
    can be 5 words too.
DO NOT over-strip: keep roughly the same amount of information. A
    rich source caption should become a LAION-style title of about
    6-12 words, NOT a bare 3-4 word phrase. Retain secondary keywords
    (materials, colors, place, brand, event, model numbers) rather
    than deleting them.
DO NOT make the caption more grammatically fluent than the LAION
    examples above.

LAION surface patterns to replicate when appropriate (the model
    decides when they fit):
- Title Case is dominant. Most LAION captions are written in Title
    Case like a product or page title (e.g. "Round Cubic Zirconia
    Stud Earrings", "Bronze Table Top Patio Heater").
- Product/e-commerce titles with brand names, model/SKU numbers,
    colors, sizes, and materials (e.g. "Tom Ford FT5458 001", "GO
    GEAR In-Ear Headphones Turbos - Black", "4.5 x 37 Frasier Fir
    Christmas Tree w/ 402T Lights").
- Product-type suffixes: "Royalty Free Stock Photo", "Fine Art
    Print", "Wall Mural", "Shower Curtain", "Poster", "Canvas Print",
    "for Sale". Add when the subject is a sellable/printable image.
- SEO / Pinterest keyword strings and page titles (e.g. "Short Bob
    Hippie Perm Hair Trend Archives Kpop Korean").
- News/editorial photo lines with a trailing credit (e.g. "... Globe
    photo by James Miller").
- Occasionally a bare brand/product name, a headline, or a "how-to"
    listicle title (e.g. "18 Best Ways to Retire").

KEEP proper nouns and key terms  a LAION caption still names its
    subject:
- Keep real names of people, brands, places, products, and model/SKU
    numbers exactly as in the original.
- The source caption may contain a `<PERSON>` placeholder. That is
    CC12M's anonymisation marker standing in for a person's name.
    Real LAION captions never contain it, so it must NEVER appear in
    your output. Each request tells you exactly how to resolve it,
    and anything you write in its place must be consistent with the
    person actually visible in the image.
- The marker is not always reliable: it sometimes replaced a word
    that was never a person's name at all. Trust the image over the
    placeholder, and drop the reference rather than forcing a person
    into a caption that has none.
- Where several `<PERSON>` placeholders stand for the same person,
    resolve them consistently, and do not repeat a name where the
    caption reads better with a single mention.
- Do NOT strip the caption down so far that the main subject/keywords
    are lost. "Premiere of AFP" or a title ending mid-phrase is WRONG
     keep the who/what.

BAD transformed captions (too polished / descriptive, or
    over-stripped  avoid this):
- "A pair of round stud earrings with cubic zirconia stones."
- "A bronze patio heater standing on a table outdoors."
- "A man is playing an electric guitar on stage."
- "Premiere of AFP"  (main subject dropped)

GOOD transformed captions (match real LAION web-scrape quality):
- "Round Cubic Zirconia Stud Earrings"
- "Bronze Table Top Patio Heater"
- "A Young Bride Holding a Bouquet - Fine Art Print"
```

```
- "Christy Turlington Returns to the Catwalk - Poster"

FINAL CRITICAL rules:
- Do not deviate from the original caption meaning  the caption
    should still describe the same image.
- Do not copy part of the provided examples in the transformed
    caption.
- Keep the output in English.
- Output exactly ONE caption. No numbering, no quotes, no
    explanation, no preamble.

<|vision_start|><|image_pad|><|vision_end|>Source CC12M caption: A
    visual comparison of an original photo of a sunset vs. the
    version Plex serves up.
Transformed LAION-style caption:
```