Membership Inference Attacks on Language Models and Vison-Language Models*

Darwin Zhang **Marco Wang** **Tianhao Cao** **Yusen Huang**

Abstract

Membership inference attacks (MIAs) are the canonical privacy-auditing tool for generative models, yet recent studies find that naive single-statistic attacks barely outperform chance on finetuned large language models (LLMs) and collapse under distributional controls on vision-language models (VLMs). We study the MIA problem across both modalities with a shared methodology: stack multiple individually-weak signals into a supervised meta-classifier. For finetuned LLMs, we target two SmolLM2 models (135M, 360M) finetuned on medical clinical notes and fuse cross-model loss ratios, neighborhood-perturbation gaps, and shallow text statistics into a 16-dimensional feature vector, lifting AUC from a raw-loss baseline of 0.584 to 0.907 and TPR at FPR=0.1 from 0.155 to 0.697. For VLMs, we combine cross-modal CLIP similarity features with a dual-model text/image neighborhood attack, achieving a Kaggle AUC of 0.8119 on a finetuned SmolVLM-256M and outperforming both pure M^{4}I-style and pure neighborhood variants. Analysis shows that the dominant signal is cross-model for LLMs but cross-modal — specifically image-to-text CLIP similarity — for VLMs, and that CLIP alignment and neighborhood perturbation signals are orthogonal: each identifies members the other misses.

1 Introduction

Modern generative models are trained on corpora whose provenance is rarely fully auditable. Large language models (LLMs) ingest web-scale text that can include personal correspondence, clinical notes, and copyrighted material, while vision-language models (VLMs) additionally consume billions of image-caption pairs scraped without explicit consent from content creators. Regulators, data owners, and model providers alike therefore face a sharpening question: given query access to a deployed model, can we decide whether a specific example was part of its training set? This is the classical *membership inference attack* (MIA) problem [12], and in the context of modern generative models it is the central tool of privacy auditing. Beyond its regulatory motivation, MIA is also the experimental handle through which researchers study memorization in high-capacity models.

Prior work has made the limits of MIA on modern models painfully clear. On pretrained LLMs, Duan et al. [3] and Chen et al. [1] show that widely cited attacks barely outperform random guessing once distributional shortcuts between member and non-member corpora are eliminated; the combination of web-scale data and near-one-epoch training keeps overfitting small, and the membership signal is weak and noisy. Reference-based and perturbation-based attacks [11, 8, 4, 14] recover useful signal on *finetuned* LLMs, but each commits to a single statistic and few are evaluated on small, domain-specialized finetuned models representative of real-world clinical deployments. On VLMs the story is sharper still. Hu et al. [6] and Li et al. [7] report high attack success on multimodal captioning and instruction-tuned LVLMs, but Miyamoto et al. [9] demonstrate on a distribution-controlled benchmark that those same attacks collapse to near-chance, suggesting that much of the reported signal was an artifact of dataset construction rather than of genuine memorization. No single attack has been shown to reliably audit a modern instruction-tuned VLM under bias-controlled conditions.

We take this two-sided fragility as our starting point. For a single team, across two linked tasks, we ask: can a principled *feature-fusion* strategy — stacking multiple individually-weak membership signals into a single supervised meta-classifier — outperform the best published single-statistic attacks on both a finetuned small LLM and a modern instruction-tuned VLM?

In the LLM setting (Task 1, COLX 531), we target two SmolLM2 models (135M and 360M parameters) finetuned on 50,000 medical clinical notes from a Kaggle competition provided for the course. We design a 16-dimensional feature vector that combines losses from both finetuned and both base models, reference ratios across model sizes, neighborhood-perturbation gaps from token-drop neighbors, and shallow text statistics such as

*Repository: https://github.ubc.ca/snalyf/TFC

length and lexical diversity. A gradient-boosting classifier trained on these features lifts validation AUC from a raw-loss baseline of 0.584 to 0.907 and TPR at FPR=0.1 from 0.155 to 0.697, with feature-importance analysis indicating that *cross-model* reference signals (the loss gap between the 135M and 360M finetuned models) dominate over neighborhood perturbation effects.

In the VLM setting (Task 2, COLX 585), we target a finetuned SmolVLM-256M-Instruct on 6,000 image-text pairs. We adapt the M^{4}I framework of Hu et al. [6] by replacing its custom multimodal feature extractor with off-the-shelf CLIP (ViT-B/32) and its SVM meta-classifier with a modern gradient-boosted / logistic classifier. Our best attack combines cross-modal CLIP similarity features (image-to-text, image-to-generation, and gap signals) with a dual-model text/image neighborhood attack adapted from Task 1 (token-drop text neighbors and noise/patch-mask image neighbors scored against both the finetuned and base VLM), reaching a **Kaggle AUC of 0.8119**. A lean CLIP-only variant without neighborhood features remains within a small margin, indicating that the usable membership signal in modern instruction-tuned VLMs is primarily *data-centric* (how naturally aligned the image-text pair is in CLIP space) rather than *model-centric* (what the target VLM produces), while the neighborhood signals provide orthogonal, complementary evidence.

Contributions. This paper makes four contributions:

1. A dual-model feature-fusion MIA for finetuned LLMs that achieves AUC 0.907 and TPR@FPR=0.1 of 0.697 on medical clinical notes, substantially improving over single-statistic baselines.

2. A CLIP+Neighborhood MIA for finetuned VLMs that fuses cross-modal CLIP similarity features with a dual-model text/image neighborhood attack, achieving Kaggle AUC 0.8119 on SmolVLM-256M and outperforming pure M^{4}I-style or pure neighborhood variants.

3. Empirical evidence that VLM membership is primarily data-centric: CLIP alignment features alone already match most of the full attack's performance, while neighborhood signals provide orthogonal gains — consistent with the distributional-bias critique of Miyamoto et al. [9].

4. A cross-task observation that the effective *reference signal* differs by modality: cross-model (two finetuned sizes) for LLMs, and cross-modal (image vs. text in a shared embedding space) for VLMs.

The remainder of the paper is organized as follows. Section 2 surveys related work. Section **??** describes the two datasets. Section 4 presents our methods for both tasks. Section 5 details the experimental setup. Section **??** reports results and analysis.

2 Related Work

We organize prior work into three threads: membership inference on language models (§2.1), membership inference on vision-language models (§2.2), and the theoretical foundations of privacy auditing (§2.3). Our two tasks build directly on the first two threads while being informed by the third.

2.1 Membership Inference on Language Models

Membership inference attacks (MIAs) were introduced by Shokri et al. [12], who formalized the problem as a binary classification over a target model's outputs and used shadow models to learn members-vs-non-members decision boundaries on classical machine learning tasks. Extending the framework to large language models (LLMs) has proven difficult: at the pretraining scale, the combination of web-scale corpora and near-one-epoch training sharply limits overfitting, and Duan et al. [3] showed across Pythia, GPT-Neo, and OLMo model families that popular MIAs barely outperform random guessing once member and non-member corpora are drawn from comparable distributions. Chen et al. [1] reinforce this finding with a large statistical study showing that most published methods do not significantly beat simple loss-based baselines once multiple-threshold and multiple-domain variance is accounted for.

Within this difficult regime, three lines of attack have shown real gains. *Reference-based* methods calibrate the target model's loss against a second model to remove per-sample difficulty bias; Min-K% Prob [11] isolates the bottom-k% of token log-probabilities as a more discriminative signal than mean loss, while SPV-MIA [4] synthesizes its own reference distribution by self-prompting the target, lifting AUC from roughly 0.7 to 0.92 on finetuned GPT-2, GPT-J, Falcon-7B, and LLaMA-7B. *Neigh-*

borhood attacks [8] bypass the reference-dataset assumption altogether by comparing the target's loss on the original sample against losses on locally perturbed neighbors, outperforming LiRA by up to 100% in realistic settings. *Conditional-likelihood* methods such as RECALL [14] measure the drop in log-likelihood when the sample is prefixed with non-member context, achieving state-of-the-art performance on WikiMIA.

Our Task 1 method sits in this landscape as a *feature-fusion* generalization of the above black-box attacks. Rather than committing to a single statistic, we combine cross-model reference ratios across two finetuned model sizes, neighborhood-perturbation gaps, and shallow text statistics into a 16-dimensional feature vector and learn a supervised decision boundary with gradient boosting. Unlike G-Drift we remain strictly black-box; unlike SPV-MIA we require no self-prompted reference; and unlike pure neighborhood attacks we show empirically that *cross-model* signals dominate.

2.2 Membership Inference on Vision-Language Models

Multimodal MIA was opened by Hu et al. [6], who proposed two complementary strategies for image-captioning models built on ResNet-152 and an LSTM decoder. Their metric-based attack (MB-M^{4}I) generates a caption from the target and compares it to the ground truth using ROUGE and BLEU, while their feature-based attack (FB-M^{4}I) embeds images and captions into a shared space with a custom multimodal feature extractor (MFE) and uses Euclidean distance as the membership signal; they report attack success rates up to 94.83% on MS-COCO, Flickr8k, and IAPR TC-12. Li et al. [7] extended this to modern large vision-language models (LVLMs) by introducing the VL-MIA benchmark over LLaVA-1.5, MiniGPT-4, and LLaMA-Adapter V2, together with MaxRényi-$K\%$, a target-free entropy-based metric that operates on either modality and uses cross-modal token slices to probe image membership through the text channel.

Miyamoto et al. [9] cast doubt on many of these reported gains. Their OpenLVLM-MIA benchmark equalizes the distribution of member and non-member images across three LVLM training stages, and under this control state-of-the-art attacks including Perplexity, Min-K% Probability, and MaxRényi collapse to near-random performance. The implication is that much of the apparent signal in prior LVLM-MIA results was driven by dataset-construction artifacts rather than genuine memorization, and that meaningful progress requires attacks whose signals are tied to content rather than distribution.

Our Task 2 method responds directly to this critique. We modernize the M^{4}I framework in two ways: we replace the custom MFE with off-the-shelf CLIP (ViT-B/32) [10], and we replace the original linear SVM with XGBoost [2]. We combine MB-M^{4}I features (ROUGE-L and token overlap between ground-truth and VLM-generated text) with FB-M^{4}I features (CLIP cosine similarities between image, ground-truth, and generated text, plus their gap). Empirically, the cross-modal alignment signal in CLIP space dominates: image-to-text CLIP similarity accounts for roughly half of the classifier's importance, while features derived from the VLM's own generations are marginal. This reinforces the central message of OpenLVLM-MIA: in modern instruction-tuned VLMs, membership appears to be a property of how naturally aligned an image-text pair is in a strong pretrained cross-modal space, rather than of the target VLM's output.

2.3 Theoretical and Methodological Foundations

Finally, Haghifam et al. [5] study MIA from a sample-complexity angle. Working in the tractable setting of Gaussian mean estimation, they show that successful sample-based attacks may require $\Omega(n + n^2\rho^2)$ auxiliary samples when the data covariance is unknown, far more than the $O(n)$ typically assumed in empirical audits. This suggests that much of the pessimism around LLM-MIA [3, 1] and VLM-MIA [9] may partly reflect auditor rather than model limitations, and it argues for attacks that exploit richer side information about the data distribution. Our feature-fusion approach can be read in this light: by stacking multiple weak signals with different distributional assumptions, we aim to extract membership information that no single statistic captures in isolation.

Synthesizing these threads, prior work establishes two facts that shape our design: (i) MIAs on modern LLMs succeed only in narrow regimes, typically finetuning or strong memorization, and require calibration against richer signals than raw loss; and (ii) MIAs on VLMs are fragile under bias controls and rely more on cross-modal alignment than on per-token likelihoods. Our contribution is

to instantiate the same methodological principle — combining multiple weak, complementary signals through a supervised meta-classifier — in both modalities, adapted to each setting's geometry.

3 Data

Our study evaluates membership inference attack performance across two distinct datasets with differnet modalities: text-only, and vision-langauge model.

3.1 Medical Clinical Notes (COLX 531)

For the language modeling task, we investigated a dataset consisting of medical clinical notes, specifically focusing on patient discharge summaries and detailed case reports. Each record contains three primary features: a unique identifier, the text of the medical report (which includes patient demographics, medical history, treatments, and outcomes), and a binary `is_member` label. This binary flag indicates whether the sample was utilized during the fine-tuning of the target SmolLM2 models. The dataset exhibits a consistent distribution of text lengths across its splits, as detailed in Table 1.

Split	N	Mean Length	Max	Min
Train	50,000	2016.88	5650	752
Validation	10,000	2010.69	4886	836
Test	15,000	2021.14	5382	787

Table 1: Descriptive statistics of text character lengths across the medical clinical notes dataset splits.

3.2 Multimodal Data (COLX 585)

The vision-language task targets the `UBC-SLIME/colx_585_vlm` dataset, which focuses on evaluating instruction-tuned vision-language models. Each sample consists of five primary features: a unique identification string, raw image bytes, a binary `is_member` label, and a text string containing a conversational prompt paired with a model response.

A critical feature of this dataset is the categorical `type` metadata (e.g., *seen_img_unseen_txt*), which enables a granular analysis of memorization across different modality exposures. The descriptive statistics for the text lengths, calculated based on the combined character count of the prompt and response strings, are summarized in Table 2.

Split	N	Mean	Med.	Max	Min	SD
Train	6,000	261.93	246	813	98	89.24
Validation	1,200	258.05	241	737	104	86.61
Test	6,000	247.19	232	1079	82	91.35

Table 2: Descriptive statistics of text lengths across the vision-language dataset splits.

4 MIAs on Language Model

4.1 Methods

4.1.1 Task Formulation

Given a fine-tuned causal language model f_θ and a text sample x, the goal of a Membership Inference Attack (MIA) is to produce a scalar membership score $s(x) \in \mathbb{R}$ such that $s(x)$ is stochastically larger when x was used in fine-tuning than when it was not. The classification is evaluated by AUC-ROC and TPR at a fixed FPR of 0.1.

4.1.2 Target Models and Architecture

All methods operate on one or both of the following model pairs, where each fine-tuned model is paired with its pre-fine-tuning base checkpoint:

- **SmolLM2-135M pair**: fine-tuned `UBC-SLIME/colx_531_smollm2-135m` and its base checkpoint `HuggingFaceTB/SmolLM2-135M`.
- **SmolLM2-360M pair**: fine-tuned `UBC-SLIME/colx_531_smollm2-360m` and its base checkpoint `HuggingFaceTB/SmolLM2-360M`.

Models variants are decoder-only Transformer language models [13]. The 135M model uses 30 transformer layers, each with multi-head self-attention, 9 heads, head dimension 64, hidden size 576, and a gated MLP with SiLU activation. The 360M model uses 32 layers with hidden size 960 and 15 attention heads. Both use Rotary Position Embeddings (RoPE), RMSNorm for layer normalization, and a vocabulary of 49,152 tokens with tied input/output embeddings. All inference is performed in `eval()` mode with `torch.no_grad()`, using `bfloat16` on CUDA. Sequence inputs are truncated to a maximum of 1,024 tokens.

4.1.3 Method 1: Raw Loss Baseline

The simplest membership proxy is the sequence-level cross-entropy loss of the fine-tuned model. For a tokenized input $\mathbf{x} = (x_1, \ldots, x_T)$, the model computes:

$$\mathcal{L}(x) = -\frac{1}{T}\sum_{t=1}^{T} \log p_\theta(x_t \mid x_{<t}) \tag{1}$$

The membership score is $s(x) = -\mathcal{L}(x)$: lower loss implies higher membership probability. This baseline requires no reference model and no labeled data.

4.1.4 Method 2: Reference Model (Likelihood Ratio)

Raw loss conflates the inherent linguistic difficulty of a document with the memorization signal. To isolate the latter, we compare the fine-tuned model's loss against the same architecture's base (pre-fine-tuning) checkpoint [**?**]:

$$s_{\text{ref}}(x) = \mathcal{L}_{\text{base}}(x) - \mathcal{L}_{\text{ft}}(x) \tag{2}$$

A positive score indicates that fine-tuning disproportionately reduced the loss for x, which is the expected signature of memorization. Both models use the same tokenizer; inference follows the same truncation and device settings as the baseline.

4.1.5 Method 3: Min-K% Prob

Rather than summarizing loss as a sequence mean, Min-K% Prob [**?**] focuses on the tail of the per-token log-probability distribution. For a given sequence, we compute per-token log-probabilities $\{l_1, \ldots, l_{T-1}\}$ from the fine-tuned model, select the $k = \lfloor (T-1) \cdot K \rfloor$ tokens with the *lowest* log-probability, and define:

$$s_{\text{MinK}}(x) = \frac{1}{k} \sum_{i \in \text{bottom-}k} l_i \tag{3}$$

The intuition is that even the hardest-to-predict tokens in a memorized sequence retain relatively high log-probability compared to unseen text. We set $K = 0.20$ (bottom 20% of tokens).

4.1.6 Method 4: Casing Attack

The Casing Attack exploits fine-grained stylistic memorization: a model trained on text with specific capitalization should exhibit elevated loss when that text is fully lowercased. For each sample, we extract four scalar features using the fine-tuned 135M model: the loss on the original text ($f_1 = \mathcal{L}(x_{\text{orig}})$), the loss on the lowercased text ($f_2 = \mathcal{L}(x_{\text{lower}})$), the absolute loss increase due to lowercasing ($f_3 = f_2 - f_1$), and the loss ratio ($f_4 = f_2/(f_1 + \varepsilon)$). These four features are standardized with `StandardScaler` and passed to a **Gradient Boosting Classifier** trained on labeled train-split samples. The GBC is configured with 200 estimators, maximum tree depth 4, learning rate 0.1, and subsample ratio 0.8. The model outputs $\hat{p}(y = 1 \mid \mathbf{f})$ as the membership score via `predict_proba`, and the cost function minimized during training is the binomial deviance (log-loss).

4.1.7 Method 5: Improved Neighborhood Attack (Dual-Model Feature Fusion)

Our strongest method combines all four models into a supervised feature-based classifier. For each sample x, we extract a **16-dimensional feature vector** spanning five categories:

Base losses. Per-sample cross-entropy loss from each of the four models: $\ell_{\text{ft1}}, \ell_{\text{ft2}}, \ell_{\text{base1}}, \ell_{\text{base2}}$.

Reference ratios. For each model pair ($i \in \{1, 2\}$):

$$r_i^{\Delta} = \ell_{\text{base}_i} - \ell_{\text{ft}_i} \tag{4}$$

$$r_i^{\times} = \ell_{\text{ft}_i} / (\ell_{\text{base}_i} + \varepsilon) \tag{5}$$

Neighborhood features. We generate $N = 5$ perturbed neighbors $\{\tilde{x}_j\}$ per sample using **token-drop perturbation**: for each neighbor, a random 10% of whitespace-tokenized words are removed. For each fine-tuned model i, we compute:

$$g_i = \frac{1}{N}\sum_{j=1}^{N} \ell_{\text{ft}_i}(\tilde{x}_j) - \ell_{\text{ft}_i}(x) \tag{6}$$

$$\sigma_i = \text{std}\big[\ell_{\text{ft}_i}(\tilde{x}_1), \ldots, \ell_{\text{ft}_i}(\tilde{x}_N)\big] \tag{7}$$

where g_i represents the neighborhood gap, and σ_i is the standard deviation of the neighbor losses.

Cross-model features.

$$d_{\text{ft}} = \ell_{\text{ft1}} - \ell_{\text{ft2}} \tag{8}$$

$$d_{\text{gap}} = g_1 - g_2 \tag{9}$$

where d_{ft} measures the loss difference between the fine-tuned models, and d_{gap} captures the neighborhood gap difference.

Text-level features. These include the total word count $|\mathbf{w}|$ and the lexical diversity, which is calculated as $\frac{|\text{unique}(\mathbf{w})|}{|\mathbf{w}|}$.

All 16 features are standardized with `StandardScaler`. A **Gradient Boosting Classifier** with 300 estimators, max tree depth 4, learning rate 0.05, and subsample ratio 0.8 is trained on train-split labels and outputs $\hat{p}(y = 1 \mid \mathbf{f})$. The cost function is binomial deviance (log-loss).

4.2 Experiments

4.2.1 Experimental Settings

Baselines. Method 1 (raw loss) serves as the primary baseline. It requires no labeled data, no reference model, and no fine-tuning, making it the natural lower bound for all subsequent methods.

Reference model (Method 2). The reference model requires loading two 135M-parameter models simultaneously but no labeled supervision. This ablates the contribution of loss normalization over the raw-loss baseline.

Min-K% Prob (Method 3). We evaluate $K \in \{0.20\}$ on the validation split. Only the fine-tuned 135M model is needed; no labeled data is required.

Casing Attack (Method 4). The GBC classifier is trained on all labeled train-split samples (50,000 examples) and evaluated on the validation split. We report feature importances to assess which casing signal drives performance.

Improved Neighborhood Attack (Method 5). Four models are loaded simultaneously. Features are extracted for all three splits (train / validation / test). The GBC classifier is trained on train-split labels only; validation performance is reported without any hyperparameter tuning on the validation set. We conduct a feature importance analysis after training to identify the dominant signals.

4.2.2 Analysis

Effect of loss normalization. The reference model (Method 2) improves AUC by +8.6 points and TPR@FPR=0.1 by +6.6 points over the raw-loss baseline. This confirms that normalizing against a base model effectively removes the confound of document-level linguistic complexity.

Min-K% Prob vs. Reference Model. Both methods improve over the baseline but differ in what they measure: Min-K% Prob focuses on per-token tail surprisal within a single model, while the Reference Model operates at the sequence level across two models. The Reference Model is stronger (+5.4 AUC), suggesting that cross-model normalization captures more signal than token-level tail statistics for this domain.

Failure of the Casing Attack. The Casing Attack yields near-baseline performance (AUC 0.5887), indicating that SmolLM2-135M does not memorize surface-level capitalization patterns to a degree detectable via lowercasing perturbation. The largest GBC feature importance falls on `loss_orig` rather than `loss_diff`, meaning the classifier degrades to near-raw-loss behavior.

Neighborhood Attack and cross-model signals. The Improved Neighborhood Attack achieves a dramatic jump to AUC 0.91. Feature importance analysis reveals that the top three features account for 86% of total importance (Table 3), and all are cross-model reference signals, not neighborhood gaps.

Rank	Feature	Importance
1	`loss_diff_ft`	33.1%
2	`ref_diff_2`	30.0%
3	`ref_ratio_2`	23.1%
	Neighborhood features (`nb_gap`)	$\sim$0.5%

Table 3: Top feature importances for the Improved Neighborhood Attack (GBC, 300 estimators).

The dominant feature, `loss_diff_ft` ($\ell_{\text{ft1}} - \ell_{\text{ft2}}$), captures a systematic difference in how the 135M and 360M fine-tuned models memorize training samples. Members exhibit a characteristic loss gap between the two model sizes that non-members do not. The 360M reference pair (`ref_diff_2`, `ref_ratio_2`) contributes over 53% of the signal, suggesting that larger models produce stronger memorization signatures. The classical neighborhood gap features contribute minimally ($\sim$0.5%), confirming that for fine-tuned LLMs, distributional-level memorization outweighs verbatim token-sequence memorization.

5 Method and Experiments for MIAs on Visual Language Model

5.1 Methods

5.1.1 Task Formulation

Given a fine-tuned vision-language model f_θ and a multimodal sample (x_v, x_t) consisting of an image x_v and its associated text x_t, the goal of a Membership Inference Attack (MIA) is to produce a scalar membership score $s(x_v, x_t) \in \mathbb{R}$ such that $s(x_v, x_t)$ is stochastically larger when the image-text pair was used in fine-tuning than when it was not. As in the text-only setting, the classification is evaluated by AUC-ROC and TPR at a fixed FPR of 0.1.

5.1.2 Target Models and Architecture

All methods operate on a single model pair, where the fine-tuned model is paired with its pre-fine-tuning base checkpoint:

- **SmolVLM-256M pair**: fine-tuned `UBC-SLIME/colx_585_vlm` and its base checkpoint `HuggingFaceTB/SmolVLM-256M-Instruct`.

SmolVLM is a vision-language model that couples a SigLIP-based vision encoder with a LLaMA-family causal language model through a modality projection layer. The vision encoder processes input images (resized to a longest edge of 512 pixels) using 12 transformer layers with hidden size 768, 12 attention heads, patch size 16, intermediate size 3072, GELU activation, and LayerNorm ($\epsilon = 10^{-6}$). Visual tokens are projected into the language model's embedding space via a pixel-shuffle downsampling layer (scale factor 4) and concatenated with the text token embeddings.

The language model component uses 30 transformer layers with hidden size 576, 9 attention heads (3 key-value heads, head dimension 64), and a gated MLP with SiLU activation and intermediate size 1536. It employs Rotary Position Embeddings (RoPE, $\theta = 100{,}000$) with a maximum context length of 8192, RMSNorm ($\epsilon = 10^{-5}$) for layer normalization, and a vocabulary of 49,280 tokens. All inference is performed in `eval()` mode with `torch.no_grad()`, using `bfloat16` on CUDA and `float32` on CPU/MPS. Text inputs are formatted using SmolVLM's chat template with SDPA attention, where the user turn contains the image and prompt and the assistant turn contains the target text.

5.1.3 Method 1: Raw Loss Baseline

The simplest membership proxy for VLMs extends the text-only loss baseline to multimodal inputs. Given an image x_v and a tokenized text sequence $x_t = (x_1, \ldots, x_T)$ formatted as a user–assistant conversation via the SmolVLM chat template, the model computes the cross-entropy loss over the assistant (caption) tokens only:

$$\mathcal{L}(x_v, x_t) = -\frac{1}{T}\sum_{t=1}^{T} \log p_\theta(x_t \mid x_{<t}, x_v) \quad (10)$$

where the conditioning on x_v reflects the image tokens injected into the sequence via the vision encoder and projection layer. The membership score is $s(x_v, x_t) = -\mathcal{L}(x_v, x_t)$: lower loss implies higher membership probability.

A Logistic Regression classifier with balanced class weights is trained on the scalar loss feature extracted from the training split (6,000 samples) and evaluated on the validation split (1,200 samples) using AUC-ROC and TPR@FPR=0.1. This baseline requires no reference model and no labeled data beyond the binary membership indicator.

5.1.4 Method 2: Neighborhood Attack

This method measures local loss curvature around each sample by generating perturbed neighbors in both modalities and comparing losses from the fine-tuned and base models. For each sample, 5 text neighbors are created by randomly dropping 10% of words from the assistant response, and 5 image neighbors are generated by cycling through Gaussian noise ($\sigma = 25$), MAE-style patch masking (15% of 32×32 patches), and random crop-and-resize (70–90% crop). The core signal is the neighborhood gap:

$$g(x) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}(x_v, \tilde{x}_t^{(i)}) - \mathcal{L}(x_v, x_t) \quad (11)$$

where $\tilde{x}_t^{(i)}$ are perturbed neighbors. Members sit at sharper local minima, producing larger gaps. Analogous gaps are computed for image perturbations and the base model, yielding relative gaps that isolate finetuning-specific memorization. Combined with cross-modal binding features (product and sum of text and image gaps) and text statistics, the full 17-dimensional feature vector is fed into an XGBoost classifier.

5.1.5 Method 3: Multi-Feature Min-K% Prob

This method extends the raw loss baseline by extracting seven features from per-token log-probabilities: mean loss, perplexity, Min-K% Prob (bottom 20% tokens), maximum single-token loss, standard deviation of token losses, zlib-normalized loss ratio, and token count. The zlib ratio divides loss by the compressed byte length of the text, serving as a complexity calibration proxy that prevents inherently simple texts from being falsely flagged as memorized. All features are standardized and fed into a Logistic Regression classifier with balanced class weights.

5.1.6 Method 4: Lightweight Complexity Calibration

This method uses the same seven features as Method 3 but replaces Logistic Regression with an XGBoost classifier (300 estimators, max depth 4, learning rate 0.05). The motivation is that nonlinear interactions between token-level statistics — such as sequences with high perplexity but low zlib ratio — may be diagnostic of membership but invisible to a linear classifier. This single-model approach avoids the computational overhead of reference model inference while relying on zlib compression as an intrinsic calibration signal in place of a base model comparison.

5.1.7 Method 5: M^{4}I-CLIP Attack

This method adapts the metric-based and feature-based attacks from M^{4}I [6] to the VLM setting using CLIP as the cross-modal feature extractor. For the metric-based branch, text is generated from the finetuned SmolVLM via greedy decoding and compared to the ground-truth answer using ROUGE-L F1 and token overlap ratio. For the feature-based branch, OpenAI CLIP ViT-B/32 encodes the image, ground-truth text, and generated text into a shared embedding space, from which four cosine similarities are computed: image–text, image–generated, text–generated, and a derived gap $\Delta_{\text{clip}} = \text{sim}(x_v, \hat{x}_t) - \text{sim}(x_v, x_t)$. All seven features are fed into an XGBoost classifier. A fast CLIP-only variant omits VLM generation entirely and combines three CLIP features with precomputed loss-based statistics, achieving nearly identical Kaggle AUC (0.80537 vs 0.80560), indicating that the membership signal resides primarily in CLIP-space alignment rather than model generation behavior.

5.1.8 Method 6: Neighborhood + MIA Combined Attack

This method combines two complementary signals: a three-layer MIA attack measuring absolute memorization, and a dual-model neighborhood attack measuring perturbation sensitivity. The MIA component extracts three features: the loss ratio $r = \mathcal{L}_{\text{base}}/(\mathcal{L}_{\text{ft}} + \epsilon)$ between the base and finetuned models (univariate AUC = 0.7929), caption contrast (loss increase when the original caption is replaced with N shuffled captions), and corruption sensitivity (loss change under Gaussian blur and noise). The neighborhood component generates 5 text neighbors (10% word dropout) and 5 image neighbors (Gaussian noise, patch masking, crop-resize), computing perturbation gaps for both models and their difference to isolate finetuning-specific memorization. Three cross-modal binding features capture joint sensitivity across modalities. All 20 features are standardized and fed into a Logistic Regression classifier with balanced class weights, chosen over XGBoost for its superior TPR@FPR=0.1 (0.3884 vs 0.3719). CLIP features were evaluated but excluded due to collinearity with the loss ratio, which diluted high-confidence predictions.

5.1.9 Method 7: Hybrid MIA + LiRA

This method supersets the M^{4}I-CLIP approach by combining four signal families. Signals A (metric-based) and B (CLIP embeddings) are retained from Method 5, extended with ROUGE-1, ROUGE-2, BLEU-1, and BLEU-2. Signal C adds a LiRA cross-model comparison: for each model (finetuned and base), six token-level features are extracted (loss, perplexity, Min-K% Prob, max token loss, std token loss, zlib ratio), from which eight cross-model differentials are derived including loss_ratio, loss_diff, min_k_diff, and norm_loss_improvement $= (\mathcal{L}_b - \mathcal{L}_f)/\mathcal{L}_b$. Signal D measures generation consistency by producing multiple sampled outputs at temperature 1.0 and computing pairwise ROUGE-L similarity; members are expected to produce more consistent descriptions. All 33 features are fed into an XGBoost classifier (500 estimators, max depth 4). The method achieves validation AUC of 0.8061, comparable to M^{4}I-CLIP but below the Neighborhood + MIA method (0.8486), likely because the LiRA cross-model features are collinear with the CLIP alignment signal.

5.2 Experiments

All experiments use the `UBC-SLIME/colx585_group_project_data` dataset (6,000 train, 1,200 validation, 6,000 test), evaluated on AUC-ROC and TPR@FPR=0.1.

Progression Across Milestones. Our approach evolved through three stages. In Milestone 4, the raw loss baseline achieved AUC = 0.5008, establishing that aggregate loss alone is insufficient for VLM membership inference. In Milestone 5, three methods explored token-level statistics and perturbation sensitivity: Multi-Feature Min-K% Prob (AUC = 0.5127), Lightweight Complexity Calibration (AUC = 0.5605), and Neighborhood Attack

(AUC = 0.5988). The key finding was that text-only token statistics are near-random on VLMs, while methods incorporating image perturbation or base model comparison carry genuine signal. In Milestone 6, shifting to multimodal feature extraction produced large gains: M^{4}I-CLIP (AUC = 0.8093), Ultimate Hybrid MIA (AUC = 0.8061), and Neighborhood + MIA Combined (AUC = 0.8486, best overall).

M^{4}I-CLIP Variants. We evaluated a full variant (7 features, requiring VLM generation, Kaggle AUC = 0.80560) and a fast CLIP-only variant (10 features, no generation, Kaggle AUC = 0.80537). The $\Delta = 0.00023$ confirms that the membership signal resides in static CLIP image–text alignment rather than generation behavior, while the fast variant runs $\sim$32$\times$ faster. Per-feature analysis shows `clip_sim_gap` (AUC = 0.8991) and `clip_img_text_sim` (AUC = 0.8553) dominate, while generation-dependent features like ROUGE-L contribute minimally (AUC $\approx$ 0.52).

CLIP Integration with Other Methods. Adding CLIP features to the Neighborhood + MIA method (Method 6) reduced TPR@FPR=0.1 from 0.3884 to 0.2727 due to collinearity with `loss_ratio`, motivating separate submissions rather than feature merging.

6 Results

This section provides a comprehensive analysis of the results obtained for both the language model and vision-language model tasks. The findings are organized chronologically by milestone to illustrate the iterative progression from baseline metrics towards more refined methodologies.

6.1 MIAs on Language Model (COLX 531)

Table 4 reports the validation-split performance for all methods evaluated across Milestones 1 through 3 on the medical clinical notes dataset.

Milestone	Method	AUC	TPR@FPR=0.1
M1	Raw Loss (Baseline)	0.5839	0.1552
M2	Reference Model (Loss Ratio)	0.6701	0.2214
M2	Min-K% Prob (K=20%)	0.6166	0.1682
M2	Casing Attack	0.5887	0.1566
M3	**Neighborhood Attack (GBC)**	**0.9071**	**0.6966**

Table 4: Validation-set performance across all LLM membership inference attack methods. Best results are in bold.

The initial evaluation of membership inference demonstrates that raw sequence-level cross-entropy loss is an insufficient predictor for finetuned medical notes, yielding an AUC of only 0.5839. The primary reason for this failure is that raw loss does not distinguish between inherent linguistic difficulty of a document from the actual memorization signal. The Reference Model in Milestone 2 improved the AUC to 0.6701 by normalizing against a base model, which effectively removed document-level complexity.

However, the significant performance leap to an AUC of 0.9071 in Milestone 3 reveals that cross-model reference signals are the dominant predictors of membership. Feature importance analysis confirmed that the systematic loss gap between the finetuned models provided the most reliable evidence of memorization, vastly outperforming verbatim token-sequence memorization features.

6.2 MIAs on Vision-Language Model (COLX 585)

The progression for the vision-language task followed a similar trajectory, starting with near-random performance and culminating in a robust classifier through multimodal feature extraction. The validation-split performance is detailed in Table 5.

Milestone	Method	AUC	TPR@FPR=0.1
M4	Raw Loss (Baseline)	0.5008	0.1200
M5	Multi-Feature Min-K% Prob	0.5127	0.0800
M5	Lightweight Complexity Calibration	0.5605	0.1800
M5	Neighborhood Attack	0.5988	0.1867
M6	M4I-CLIP Attack	0.8093	0.3000
M6	Hybrid MIA + LiRA	0.8061	0.2633
M6	**Neighborhood + MIA Combined**	**0.8486**	**0.3884**

Table 5: Validation-set performance across all VLM membership inference attack methods. Best results are in bold.

The results acquired in Milestones 4 and 5 indicate that text-only token statistics are near-random when applied to VLMs, with the raw loss baseline failing to separate members from non-members. The substantial gain observed in Milestone 6 arises from shifting the analytical focus to cross-modal alignment. Analysis of the MI-CLIP variants demonstrates that the membership signal in modern instruction-tuned VLMs resides primarily in static image-text alignment within the CLIP embedding space rather than in the generative text output behavior. The combined Neighborhood and MIA approach achieved the highest overall validation AUC of 0.8486 (with a corresponding Kaggle test score of 0.8119) because it fused absolute memorization signals with perturbation sensitivity.

6.3 Error Analysis and Interpretations

The reason high-AUC results were achievable only after Milestone 3 and Milestone 6 is due to the successful isolation of the membership signal from distributional noise. Early text-only attempts failed because they could not distinguish between hard-to-learn training texts and inherently difficult non-training samples. By using probabilistic variation and reference-based calibration, inspired by frameworks such as SPV-MIA and Min-K%++, the meta-classifiers were able to identify local maxima in the likelihood space that are characteristic of training data.

In the VLM context, the results indicate that membership is a structural property of how naturally aligned an image-text pair is within a cross-modal space, providing an orthogonal leakage channel compared to text-only generation. The scores for the language model task did not reach perfect separability due to the inherently high n-gram overlap between member and non-member medical notes, which blurs the boundary for detection algorithms relying on token sequences.

7 Conclusion

This study evaluates the susceptibility of specialized, fine-tuned language and vision-language models to membership inference attacks. Through an iterative experimental process, we demonstrated that while raw sequence-level loss serves as a poor indicator of membership, the fusion of multiple individually weak signals into a supervised meta-classifier provides a robust auditing framework. Our methodology successfully lifted the validation AUC from near-random baselines to 0.9071 for medical text and 0.8486 for multimodal image-text pairs. The primary discovery of this work is that the most potent leakage signals are modality-specific: cross-model loss gaps for large language models and cross-modal alignment discrepancies for vision-language models.

7.1 Limitations

The primary limitation of this research is the significant computational cost associated with extracting high-dimensional features for vision-language tasks. The requirement of a highly time extensive extraction window for a 6,000-sample test set posed a practical barrier, necessitating complex incremental checkpointing strategies to ensure progress persistence across recycled cloud runtimes. Furthermore, the high degree of n-gram overlap and standardized terminology within medical clinical notes creates an inherently fuzzy boundary between member and non-member records, limiting the precision of token-sequence based detection. Additionally, using multiple models to capture diverse membership signals introduces overhead that makes the approach difficult to scale to time-sensitive auditing scenarios without specialized hardware.

7.2 Future Work

Future research should focus on reducing the computational cost of the auditing pipeline by developing faster and more efficient frameworks that do not rely on resource-intensive extraction processes. A promising direction involves replacing heavy embedding models with lighter alternatives that still capture the most relevant membership signals, removing the need for large external models entirely. Additionally, exploring more varied ways of modifying input text could help uncover weaknesses that simple word removal misses, especially in domains where text tends to follow predictable patterns. Finally, if the same methodology were applied to larger models in future work, it would be valuable to examine whether the membership signals identified here remain consistent as model size and complexity grow.

Generative AI Usage

Generative AI was used as a collaborative assistant throughout this project, functioning similarly to a teaching assistant rather than a solution provider. We found that while these tools are helpful for guidance and support, they are prone to errors and do not always implement complex methodologies reliably, and therefore all AI-assisted work was manually reviewed and verified by the team.

Specifically, GenAI was consulted to assist with low-level, boilerplate code writing, such as extraction loops and general scripting scaffolding, as well as debugging and improving the robustness of our implementation by helping identify and resolve technical errors. All higher-level design decisions and methodology-specific implementations were handled and reviewed by us directly.

Additionally, GenAI was used to support the writing process, checking grammar, clarity, tone, and ensuring our intended meaning was communicated accurately throughout the report.

References

[1] Bowen Chen, Namgi Han, and Yusuke Miyao. 2025. A statistical and multi-perspective revisiting of the membership inference attack in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*.

[2] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 785–794.

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

[4] Wenjie Fu, Huandong Wang, Chen Gao, Guanghua Liu, Yong Li, and Tao Jiang. 2024. Practical membership inference attacks against fine-tuned large language models via self-prompt calibration. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[5] Mahdi Haghifam, Adam Smith, and Jonathan Ullman. 2025. The sample complexity of membership inference and privacy auditing. *arXiv preprint arXiv:2508.19458*.

[6] Pingyi Hu, Zhuoran Wang, Ruoxi Sun, Hu Wang, and Minhui Xue. 2022. M^4I: Multi-modal models membership inference. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[7] Zhan Li, Yongtao Wu, Yihang Chen, Francesco Tonin, Elias Abad Rocamora, and Volkan Cevher. 2024. Membership inference attacks against large vision-language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

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

[9] Ryo Miyamoto, Xinru Fan, Fuma Kido, Tetsuya Matsumoto, and Hayato Yamana. 2025. OpenLVLM-MIA: A controlled benchmark revealing the limits of membership inference attacks on large vision-language models. *arXiv preprint arXiv:2510.16295*.

[10] 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. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*.

[11] Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2024. Detecting pretraining data from large language models. In *International Conference on Learning Representations (ICLR)*.

[12] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008.

[14] Roy Xie, Junlin Wang, Ruomin Huang, Minxing Zhang, Bhuwan Dhingra, and Rong Ge. 2024. RECALL: Membership inference via relative conditional log-likelihoods. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.