CoLA: A Choice Leakage Attack Framework to Expose Privacy Risks in Subset Training

Qi Li[1,2,3], Cheng-Long Wang[*,1,2], Yinzhi Cao[4], Di Wang[1,2]
[1]King Abdullah University of Science and Technology
[2] Provable Responsible AI and Data Analytics Lab
[3]National University of Singapore [4]Johns Hopkins University

Abstract

Training models on a carefully chosen portion of data rather than the full dataset is now a standard preprocess for modern ML. From vision coreset selection to large-scale filtering in language models, it enables scalability with minimal utility loss. A common intuition is that training on fewer samples should also reduce privacy risks. In this paper, we challenge this assumption. We show that subset training is not privacy free: the very choices of which data are included or excluded can introduce new privacy surface and leak more sensitive information. Such information can be captured by adversaries either through side-channel metadata from the subset selection process or via the outputs of the target model. To systematically study this phenomenon, we propose CoLA (Choice Leakage Attack), a unified framework for analyzing privacy leakage in subset selection. In CoLA, depending on the adversary's knowledge of the side-channel information, we define two practical attack scenarios: Subset-aware Side-channel Attacks and Black-box Attacks. Under both scenarios, we investigate two privacy surfaces unique to subset training: (1) Training-membership MIA (TM-MIA), which concerns only the privacy of training data membership, and (2) Selection-participation MIA (SP-MIA), which concerns the privacy of all samples that participated in the subset selection process. Notably, SP-MIA enlarges the notion of membership from model training to the entire data-model supply chain. Experiments on vision and language models show that existing threat models underestimate subset-training privacy risks: the expanded privacy surface leaks both training and selection membership, extending risks from individual models to the broader ML ecosystem.

1 Introduction

The scale of modern datasets has made training on the full corpus increasingly impractical. To address this, practitioners routinely employ subset training, where only a carefully chosen ratio of data is used. This paradigm is adopted not only for efficiency but also to improve data quality, since selection can remove redundancy and noise while retaining informative samples. Subset training spans diverse applications: coreset selection (Bachem et al., 2015; Munteanu et al., 2018; Mirzasoleiman et al., 2020) in vision, dataset pruning (Sorscher et al., 2022; Yang et al., 2022), active learning (Sener and Savarese, 2018; Ducoffe and Precioso, 2018) in general ML, and large-scale deduplication (Lee et al., 2022), filtering (Rae et al., 2021), and sampling (Gunasekar et al., 2023; Wettig et al., 2024) in language model pretraining.

While subset training is widely celebrated for these benefits, its privacy implications remain underexplored (Zhao and Zhang, 2025). A common intuition suggests that fewer training samples should imply less privacy leakage (Dong et al., 2022). Yet this reasoning overlooks an important fact: *the choices made during subset selection themselves encode signals about which data were included and which were excluded.* These signals can be inherited through shifts in the data distribution or model behavior, making them exploitable by adversaries.

We ask the fundamental question: *Does subset training actually reduce privacy leakage?* Our answer is *no*. We show that subset training introduces new attack surfaces: not only is the included data that used for training compromised, but the excluded data discarded from training can also become vulnerable due to correlations introduced by the selection mechanism. In other words, due to the data-oriented nature of the subset selection process, beyond the training data leakage emphasized by traditional MIA (Shokri et al., 2017; Hu et al., 2022), the choice signals further extend privacy risks from individual models to the broader data–model supply chain. Accordingly, we define two complemen-

*Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 16137–16149

tary privacy surfaces: *Training-membership MIA (TM-MIA)*, which resembles traditional MIA by focusing on the membership of training data, and *Selection-participation MIA (SP-MIA)*, a privacy surface tailored to subset training that focuses on membership at the data selection level.

To systematically study membership leakage under these privacy surfaces, we propose **CoLA (Choice Leakage Attack)**, a framework that leverages choice signals in a principled way to conduct attacks across different surfaces. CoLA captures risks under two complementary settings: (i) a *Subset-aware Side-channel* setting, where the adversary has access to the target model's outputs and selection metadata (e.g., the selection algorithm and the inclusion ratio); and (ii) a *Black-box* setting, where the adversary observes only model outputs and is aware that subsetting may have been used, without knowing any selection metadata. Extensive results show that for both privacy surfaces under these two attack settings, CoLA can substantially strengthen the attack performance. In short, subset training does not guarantee privacy; it enlarges the attack surface of modern ML pipelines and highlights the need to protect privacy across the entire data–model supply chain. We summarize our contributions as follows:

- We provide the first systematic definition and exploration of the membership leakage problem under subset training. This novel attack scenario reveals a severe privacy risk in the subset selection process: not only is the privacy of training data compromised, but the data excluded during selection is also at risk.
- We propose CoLA (Choice Leakage Attack), a framework that leverages choice signals in a principled way for more reliable membership inference, while seamlessly unifying diverse attack settings and surfaces.
- Experiments across both vision and language models confirm the broad capability of CoLA. For example, in the black-box setting, the AUC of CoLA on Pythia-160M surpasses 80% under SP-MIA.

2 Related Works

Subset training and data-efficient learning. A large body of research has explored how to reduce the cost of large-scale training by operating on subsets of data. Coreset selection constructs small but representative subsets that approximate training on the full data (Bachem et al., 2015; Mirzasoleiman et al., 2020). Dataset pruning removes redundant or low-value samples to improve efficiency and generalization (Sorscher et al., 2022; Yang et al., 2022; Tan et al., 2024; Ren et al., 2025b,a; Hu et al., 2024a; Zhang et al., 2025b; Liang et al., 2022a). Active learning queries the most informative examples to reduce annotation cost (Agarwal et al., 2020; Margatina et al., 2021). In large-scale language models, deduplication and filtering pipelines are routinely applied to eliminate noise and improve training quality (Lee et al., 2022; Gao et al., 2020). These techniques have been extensively studied for efficiency and utility, but their privacy consequences remain largely underexplored.

Membership inference attacks. Membership inference attacks (MIAs) are among the most widely studied privacy threats in machine learning. Early work by Shokri et al. (2017) proposed shadow models to train attack classifiers distinguishing members from nonmembers. Subsequent methods exploited confidence scores, loss values, or gradients (Yeom et al., 2018; Carlini et al., 2022a). MIAs have been demonstrated in supervised learning, federated learning, and large language models (Nasr et al., 2018; Hu et al., 2022; Wang et al., 2025a; Li et al., 2025), motivating defenses such as differential privacy (Abadi et al., 2016; Wang et al., 2025b; Zhang et al., 2025a; Xiang et al., 2024) and adversarial regularization (Nasr et al., 2018). This body of work reveals how models trained on fixed datasets can memorize and leak sensitive information. However, they primarily focus on constructing membership signals in a one-shot manner, with these signals being tightly coupled to a specific model. We find such model-oriented signal less effective in the context of subset training. Leveraging the unique characteristics of the subset selection process, we instead construct membership signals in a data-oriented manner.

Synthetic data and privacy. Synthetic data generation has been studied as a way to train models without exposing raw datasets, with the promise of stronger privacy (Hu et al., 2024b; Tan et al., 2025; Li et al., 2023; Liang et al., 2022b). However, subsequent research has shown that synthetic datasets can still leak sensitive information about the original data, including membership and attributes (Stadler et al., 2022; van Breugel et al., 2023; Zhao and Zhang, 2025; Huai et al., 2019). Rather than analyzing risks inherent in *synthetic*

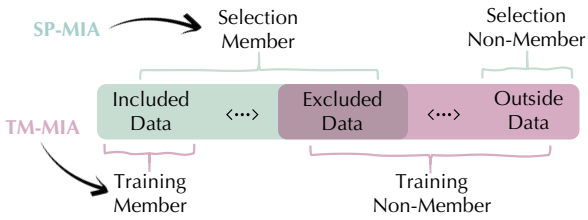


Figure 1: Privacy surfaces under subset training.

data generation pipelines, we turn to *subset training with real data*, where high-fidelity samples remain but the selection process itself exposes a distinct and overlooked channel of privacy leakage.

3 Problem Setting

3.1 Membership inference under subset training

Let $D_0 \subseteq \mathcal{X} \times \mathcal{Y}$ denote the original dataset that undergoes a subset selection procedure. A selector $\mathrm{Sel}(\cdot; r)$ with a given selection ratio r partitions D_0 into two disjoint sets: the **included data** I used for training, and the **excluded data** E that are discarded:

$$(I, E) = \mathrm{Sel}(D_0; r), \quad \text{with} \quad I \cap E = \varnothing,\ I \cup E = D_0,\ \frac{|I|}{|D_0|} = r. \tag{1}$$

Following the standard MIA pipeline (Shokri et al., 2017), we further denote by O the **outside data** that never enter the selection process. A model f_θ is trained solely on I. This partition naturally induces two types of membership inference task:

Training-membership MIA (TM-MIA). This attack takes the model itself as the attack surface and membership is defined solely by the training data. A sample x is a member if $x \in I$ and non-member if $x \in E \cup O$. This forms a natural and widely adopted threat model, as the model is the most direct output of the ML system. This setting is consistent with conventional MIAs (Shokri et al., 2017; Carlini et al., 2022a).

Selection-participation MIA (SP-MIA). However, when the attack surface is enlarged to the entire data–model pipeline, membership expands from only the training data to a much larger portion of all collected data. As shown in Figure 1, we refer to the collected data as selection members, where a sample x is a member if $x \in I \cup E$ and a non-member if $x \in O$. Its membership cannot be explained by direct model memorization, but instead reveals *choice leakage*, a side-channel signal from the subset selection process of the data-model supply chain. Such choice leakage risk is severe as it exposes a system's selection preferences. Once the data–model supply chain is exposed to privacy risks, the entire pipeline, from raw data to model outputs, becomes vulnerable to malicious manipulation. **To our knowledge, this is the first work to systematically investigate this perspective.**

Both tasks can be framed as binary hypothesis tests over a scoring function $s : \mathcal{D}_0 \to \mathbb{R}$, which measures the likelihood of a sample x belonging to the respective member set. Given $\mathcal{D}_0 = I \cup E \cup O$, the member–nonmember partitions are:

$$\mathcal{M}_{\mathrm{TM}} = I, \qquad \mathcal{N}_{\mathrm{TM}} = E \cup O, \tag{2}$$

$$\mathcal{M}_{\mathrm{SP}} = I \cup E, \quad \mathcal{N}_{\mathrm{SP}} = O. \tag{3}$$

The goal is to design a scoring function $s(x)$ that distinguishes $\mathcal{M}$ from $\mathcal{N}$ under both definitions.

3.2 Adversary knowledge

Subset training changes not only the definition of membership but also the adversary's potential knowledge and capabilities. We consider two complementary scenarios:

Subset-aware side-channel attacks. In line with the common assumption in prior MIAs, the adversary can query the deployed model f_θ and observe its outputs (e.g., prediction labels or confidence scores). In addition, it has access to *side information about the selection process*, such as the strategy used (e.g., coreset selection, pruning, filtering) or the approximate inclusion ratio. Such an assumption is realistic: pruning papers routinely report retained percentages to justify efficiency–utility trade-offs, active learning and coreset methods describe selection strategies for reproducibility, and large-scale LLM pipelines release dataset cards documenting filtering heuristics, inclusion ratios, or deduplication statistics (Cohen-Addad et al., 2021; Biderman et al., 2023a; Dubey et al., 2024; Yang et al., 2024). In many real-world pipelines, the type of selection/curation strategy is also not secret: it is often exposed through product documentation, APIs, or service deliverables (e.g., dataset curation platforms like Roboflow (Roboflow, 2024) explicitly document dedup/curation functions, and data-quality/data-broker services such as Experian (Experian, 2026) and Acxiom (Acxiom, 2026) publicly describe deduplication and filtering workflows). Crucially, this information reflects only high-level rules, not the exact membership of individual samples. Our attack targets precisely this

gap: even when only the selection algorithm or ratio is public, an adversary can exploit this side-channel to infer which specific samples were included or excluded, thereby exposing *choice leakage* in subset training.

Black-box attacks. Here the adversary can only query the deployed model f_θ and observe its outputs. The entire subset selection stage is hidden, so the adversary must rely solely on the observable behavior of the trained model or the intrinsic data-specific information. This setting captures the most restrictive and widely assumed threat model in prior MIA research (Hu et al., 2022).

3.3 Real-World Impact

Our threat model poses concrete privacy risks in real-world data pipelines. In many settings, inferring that a record appeared in the candidate pool, even if it was later excluded from training, can already disclose sensitive participation in data collection, such as HIV screening, clinical trial recruitment, welfare applications, or credit pre-approval. Such participation signals may themselves expose individuals to stigma, discrimination, or reputational harm. At the same time, when the adversary has partial knowledge of the selection process, a 'present-but-rejected' signal can further narrow plausible private attributes, such as medical markers, risk scores, or financial status. Beyond individual leakage, these signals may also reveal what kinds of samples the pipeline tends to retain or discard, thereby enabling more targeted and lower-cost poisoning attacks. The risk therefore extends beyond memorization to a broader attack surface, i.e., the data-model supply chain threat.

4 Method

4.1 Challenges of Membership Inference under Subset Training

In conventional MIA, success comes from exploiting overfitting: models tend to assign systematically higher confidence to their training data than to non-members. Under subset training, however, this signal becomes entangled. Figure 2 illustrates this using the LiRA attack signal from (Carlini et al., 2022a) on a model trained on I selected from D_0 by Glister (Killamsetty et al., 2021b). The dataset used here is CIFAR10 and the model is ResNet18. Since the selector is designed to make training on I approximate the effect of training on $I \cup E$, the confidence distributions of included, excluded, and outside samples exhibit more complex overlaps: **(i)** I concentrates at high confidence, E shifts lower, while outside data often show a bimodal distribution; **(ii)** in TM-MIA, I and $E \cup O$ remain partly separated but overlap substantially at high confidence; **(iii)** in SP-MIA, the distribution of $I \cup E$ largely overlaps with that of outside data, making the groups difficult to distinguish. This overlap complexity shows that model-oriented signals are no longer sufficient under subset training, highlighting the need for data-oriented alternatives.

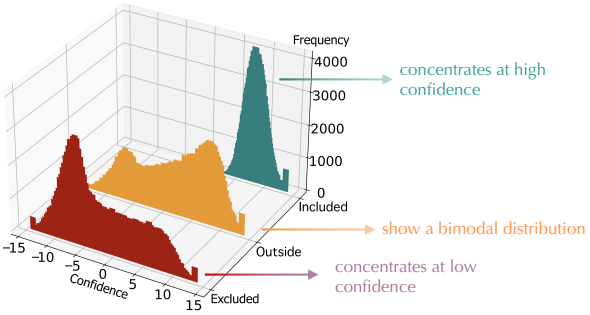


Figure 2: Signal distributions under subset training.

4.2 Choice Leakage Attack

Motivation. Just as models can overfit to their training data, subset selectors can *overfit at the selection level*: by design they preferentially reselect examples that match their implicit criteria (e.g., high informativeness, low noise, or strong representativeness). This persistent re-selection introduces a stable bias in the choice process that itself serves as a reliable membership signal. We exploit this *inclusion stability*, the tendency of a sample to be repeatedly chosen across multiple trials, as the core signal for our attack.

Specifically, we approximate many different candidate combinations by constructing a series of overlapping subsets ("windows") $\{W_i \subseteq D_0\}_{i=1}^{m}$, where m is the number of windows, to capture inclusion-stable samples. Each W_i represents one plausible candidate set the selector might face; by examining the selector's decisions on a sample across these windows, we reveal whether it is consistently favored.

Subset-aware side-channel attack. In the side-channel setting, the adversary knows both the selector $\mathrm{Sel}(\cdot; r)$ and the selection ratio $r \in (0, 1]$. For each window W_i, we run $\mathrm{Sel}(\cdot; r)$ and record whether $x \in W_i$ is selected by the selector, and get its evidence $e(x, W_i)$ in the current window:

$$e(x, W_i) = \mathbb{1}[x \in \mathrm{Sel}(W_i; r)]. \quad (4)$$

Suppose in the window construction, x appears in n out of m windows; by aggregating the selection evidence across these windows, we obtain its *inclusion count*:

$$t(x) = \sum_{i=1}^{n} e(x, W_i), \tag{5}$$

where $t(x)$ is the number of times x is selected, For fair comparison, the windows are constructed as sliding windows with fixed intervals and cyclic wrapping (details are provided in Section 5), thus each data appears in exactly the same number of windows. Hence, the exposure count n is constant across all x and serves only as a scaling factor in our score function. This also highlights the motivation behind our multi-shot membership signal: rather than relying on a single output, choice leakage signal is derived from *how consistently a sample is selected across different selections.* The membership score $s_{\text{Side}}(x)$ is obtained by aggregating evidence across windows:

$$s_{\text{side}}(x, n, r) = w\big(t(x);\, n, r\big), \tag{6}$$

where w is a monotone weighting function. From a statistical perspective, if each inclusion is a Bernoulli trial, then $t(x) \sim \text{Binomial}(n(x), p(x))$ where $p(x)$ is the probability of a data to be included. Given the selection ratio r, the expected inclusion count under random choice is $r \cdot n(x)$. We can therefore design w as a smooth monotone mapping centered around $r \cdot n(x)$:

$$w\big(t(x);\, n(x), r\big) = \frac{\sigma\big(\kappa(t(x) - r \cdot n(x))\big)}{Z(n(x), r)}, \qquad \sigma(u) = \frac{1}{1 + e^{-u}}, \qquad \kappa > 0. \tag{7}$$

where κ controls the slope and Z is a normalization constant (depending only on $n(x), r$) that does not affect relative ranking. Since the ratio $r \in (0, 1]$ and each sample has the same exposure count n. Without loss of generality, we therefore adopt the following simplified scoring function:

$$w(t(x); n) \;= \sigma\big(t(x) - \tfrac{n}{2}\big) = \frac{1}{1 + e^{-(t(x) - \frac{n}{2})}}. \tag{8}$$

This formulation monotonically amplifies scores of samples with high inclusion counts and constrains the range by n, which makes scores comparable across windows. Finally, under both TM-MIA and SP-MIA, the decision is made by thresholding:

$$\hat{y}(x) = \mathbb{1}[s_{\text{side}}(x) \geq \tau], \tag{9}$$

where τ is a decision threshold. Samples that are more stably selected as included data across windows will receive higher scores and are thus more likely to be classified as training members.

Black-box attack. In this setting, the subset selection process remains a black box to the adversary, and no direct selection metadata is available. Guided by our general motivation of *inclusion stability* (samples that are repeatedly reselected across plausible candidate sets reveal membership), we infer stable inclusion by identifying samples that consistently act as geometric representatives across windows. Specifically, for each window we perform unsupervised embedding clustering to locate representative samples. Formally, let $f(\cdot)$ be an embedding model. For each window $W_i \subseteq \mathcal{D}_I$, we compute embeddings $f(x), x \in W_i$, and perform k-means clustering (Ahmed et al., 2020) in the embedding space. Each sample $x \in W_i$ is then assigned to a cluster $c(x; W_i)$, and we measure its distance to the corresponding cluster centroid $d(x, W_i) = \|f(x) - c(x; W_i)\|_2$. The distance is used to serve as the evidence:

$$e(x, W_i) = \mathbb{1}\big[d(x, W_i) \leq Q_{0.5}\big(W_i\big)\big], \tag{10}$$

where $Q_{0.5}(\cdot)$ is the median distance among all samples in W_i. The formal definitions of the inclusion count and exposure count follow the same formulation as in Eq. 5, with the only difference that the evidence $e(x, W_i)$ is redefined as Eq. 10 under the current black-box setting.

Here, to capture multi-shot stability, since the evidence for each data now related the distance to its centroid in each window W_i, we apply a weighted score function which reveals not only the inclusion count but also the actual distance it receives:

$$s_{\text{black}}(x) = w(t(x); n)/\bar{d}(x), \tag{11}$$

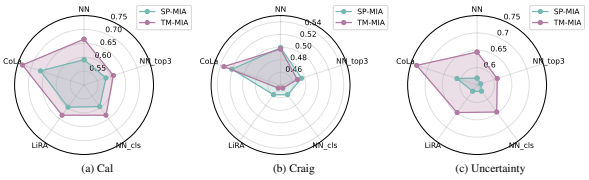
where $\bar{d}(x) = \frac{1}{t(x)} \sum_{i:\, x \in W_i} d(x, W_i)$ denotes the average clustering distance of sample x across the windows in which it is included. This design ensures that samples consistently close to centroids across many windows receive higher scores. The weighting function $w(t; n)$ follows the same formulation as in Eq. 8. Finally, similar to the side-channel setting, membership is determined by thresholding:

$$\hat{y}(x) = \mathbb{1}[s_{\text{black}}(x) \geq \tau]. \tag{12}$$

This unsupervised formulation enables membership inference even without any knowledge of the

Table 1: Results for vision models under the subset-aware side-channel attack setting. Results are averaged over 9 coreset selection methods. *Intensity* denotes the selection ratio r (Light: $r = 0.2$, Medium: $r = 0.4$, Heavy: $r = 0.6$, Extensive: $r = 0.8$). Best results per row are in bold.

Intensity	Setting	NN		NN_top3		NN_cls		LiRA		CoLA	
		AUC	TPR@5%FPR	AUC	TPR@5%FPR	AUC	TPR@5%FPR	AUC	TPR@5%FPR	AUC	TPR@5%FPR
Light	SP-MIA	51.23 ±2.56	5.83 ±1.34	51.77 ±3.02	3.34 ±4.45	51.59 ±2.66	6.37 ±2.22	51.26 ±4.85	6.43 ±3.70	**61.39 ±2.48**	**14.24 ±2.02**
	TM-MIA	64.00 ±12.15	12.13 ±5.96	61.57 ±15.05	8.93 ±13.52	67.24 ±16.18	19.30 ±17.14	69.86 ±22.08	15.74 ±18.19	**83.77 ±2.44**	**42.19 ±4.51**
Medium	SP-MIA	52.33 ±3.56	6.31 ±1.48	53.59 ±4.30	3.56 ±2.28	54.51 ±4.51	6.64 ±1.58	54.99 ±4.69	5.31 ±0.42	**81.93 ±3.50**	**42.66 ±5.81**
	TM-MIA	59.84 ±12.84	10.80 ±4.97	60.37 ±10.53	2.91 ±3.32	66.84 ±11.79	12.51 ±5.85	62.96 ±13.69	4.61 ±2.52	**88.53 ±2.55**	**60.10 ±7.62**
Heavy	SP-MIA	52.21 ±3.83	12.20 ±15.48	53.20 ±4.85	2.80 ±2.43	53.53 ±5.94	12.37 ±15.44	53.69 ±5.68	4.26 ±1.77	**96.86 ±2.60**	**88.60 ±5.51**
	TM-MIA	55.00 ±9.59	19.31 ±26.18	52.40 ±10.78	1.67 ±2.04	57.44 ±11.06	19.63 ±26.72	52.61 ±11.77	2.81 ±2.32	**89.06 ±1.90**	**60.36 ±5.87**
Extensive	SP-MIA	55.64 ±5.31	7.59 ±1.68	59.56 ±6.39	4.00 ±2.92	56.66 ±5.90	7.60 ±1.87	61.54 ±8.36	5.09 ±2.68	**92.20 ±6.94**	**91.86 ±7.23**
	TM-MIA	61.41 ±6.63	10.99 ±2.61	60.13 ±12.20	4.21 ±4.15	62.80 ±7.52	11.27 ±2.73	59.66 ±12.03	4.63 ±3.77	**80.74 ±8.23**	**49.76 ±6.98**



(a) Cal (b) Craig (c) Uncertainty

Figure 3: The MIA performance of different attack surface on vision models under black-box setting.

underlying subset selection metadata. The inclusion stability-based pipeline of CoLA naturally unifies different attack surfaces within a single framework, thereby facilitating coordinated attacks.

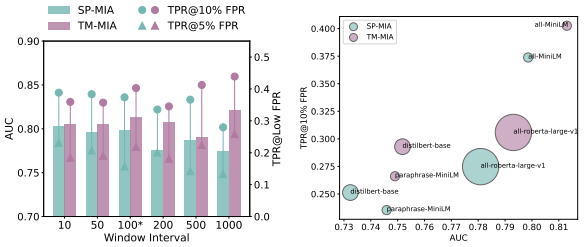
5 Experiments

5.1 Setups

Models and Datasets. We conduct experiments on both vision and language models. For the vision side, without loss of generality, we use ResNet-18 trained on CIFAR-10. We evaluate the performance on both subset-aware side-channel attacks and black-box attacks. For language models, since training multiple LMs from scratch is computationally expensive, we restrict our study to black-box attacks. The detailed data and model usage in our experiments can be found in Appendix A.2.

Subset Selection Methods. For vision models, we select nine representative dataset pruning methods from different categories. Specifically, we include decision boundary based methods such as DeepFool (Ducoffe and Precioso, 2018) and Contrastive Active Learning (Cal) (Margatina et al., 2021); the bi-level optimization based method Glister (Killamsetty et al., 2021b); error based methods including Forgetting (Toneva et al., 2018) and GraNd (Paul et al., 2021); the uncertainty based method Least Confidence (denoted as Uncertainty) (Coleman et al., 2020); the gradient matching based method Craig (Mirzasoleiman et al., 2020); and geometry based methods such as Contextual Diversity (Agarwal et al., 2020) and Herding (Welling, 2009). These methods cover a broad range of perspectives on dataset pruning, from boundary sensitivity to optimization criteria, error contribution, uncertainty, gradient alignment, and geometric diversity. The selection ratio is set to 0.2, 0.4, 0.6, and 0.8. For language models, as discussed in the previous subsection, we adopt a commonly used data filtering strategy that has been systematically studied in (Meeus et al., 2024; Duan et al., 2024), and consider two deduplication strengths, namely '13_0.8' and '13_0.2'.

Baseline MIA Methods. For vision models, we consider four baselines: NN, NN_top3, and NN_Cls (Shokri et al., 2017; Salem et al., 2018), which use the model's output logits, the top-3 logits, and the combination of logits with class labels as membership signals, respectively, as well as LiRA (Carlini et al., 2022a), which fits Gaussian distributions and leverages the likelihood to infer membership. The shadow model used in each baseline method is set to 8. For language models, we consider six baselines, including the loss (Yeom et al., 2018), Lower (lowercase) (Carlini et al., 2021), Min-K% (minkprob) (Shi et al., 2023), Min-K%++ (minkplusplus) (Zhang et al., 2024), Pac



(a) Window size. (b) Embedding model.

Figure 4: Ablation studies on (left) the window size and (right) the embedding model for MIA performance.

(pac_10) (Ye et al., 2024), and the Golden baseline Bag_of_Words (bow) (Meeus et al., 2024). Here, bow serves as a performance reference: methods performing below it are regarded as ineffective.

Evaluation Metrics. In most MIA studies (Hisamoto et al., 2020; Carlini et al., 2022b; Li et al., 2025), attack performance is typically evaluated by aggregating over all possible thresholds using the AUC score. We adopt the same practical evaluation metric in our experiments. We also report True Positive Rate at low False Positive Rate (TPR@Low FPR) (Carlini et al., 2022a), which is an important metric in MIAs and measures the detection rate at a meaningful threshold.

5.2 Results under Subset-aware Side-channel Attacks

A subset-aware side-channel attack is a type of attack specific to the subset selection process. Its success indicates that current practices of disclosing meta information about subset selection are unsafe and can lead to privacy leakage.

Table 1 reports the average MIA results across different coreset selection methods we consider for vision models. As shown, in the relatively simple TM-MIA setting, baseline methods can still perform reasonably well, which is expected since this setting closely resembles traditional MIAs (Shokri et al., 2017; Hu et al., 2022) for which these baselines were originally designed.

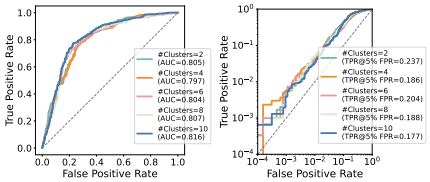
However, in the SP-MIA setting that is unique to subset training, baseline methods largely fail (AUC close to 50%), indicating their inability to effectively distinguish between included and excluded data. Fundamentally, this stems from the fact that baseline methods rely heavily on model outputs; as illustrated in Figure 2, included and excluded data exhibit output distributions that are highly similar to other data, resulting in poor separability. However, this does not mean that privacy cannot be compromised under SP-MIA. In contrast, CoLA achieves strong performance in both TM-MIA and SP-MIA settings, thanks to its multi-shot, data-centric membership signal that tightly aligns with the subset selection process and captures fine-grained data interactions, thereby enabling better separability. Moreover, we observe that as the selection ratio (*Intensity*) increases, the risk of privacy leakage becomes more severe, highlighting the significant vulnerability of the subset selection process as a potential side channel.

In the black-box attack setting, we study both vision and language models. Language model subset selection often relies on heuristic semantic filtering or deduplication, rather than the formally defined selection algorithms and ratios common in vision, which makes it naturally suited to black-box analysis. In this scenario, the adversary has no access to any meta information about the selection procedure. Consequently, a successful membership inference attack under these conditions indicates that the subset selection process itself—much like model training—can implicitly reveal private information about the data. This implies that privacy risks arising from subset selection must be addressed proactively: mitigating them requires careful design choices and safeguards before the selection process is executed.

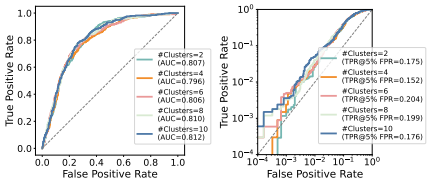
The results for vision models and language models are shown in Figure 3 and Figure 7, respectively. For vision models, we adopt three representative selection methods: Cal (Margatina et al., 2021), Craig (Mirzasoleiman et al., 2020), and Uncertainty (Coleman et al., 2020). As illustrated in Figure 3, under the black-box setting, SP-MIA remains more challenging than TM-MIA. Moreover, CoLA consistently outperforms the baselines by about 5% in AUC across all experiments, demonstrating strong attack capability. For language models, this contrast is even more pronounced. As shown in Figure 7, all baseline methods except CoLA perform worse than the bow baseline, indicating that they essentially fail in the context of subset selection MIA. Furthermore, while SP-MIA and TM-MIA results are relatively close for CoLA, the baselines exhibit a sharp gap, with SP-MIA close to random guessing (AUC around 50%), and TM-MIA reaches only about 60%.

5.3 Ablation Studies.

Influence of Window Construction. In Figure 4a, we present an ablation study on the influence of window interval, conducted with Pythia-

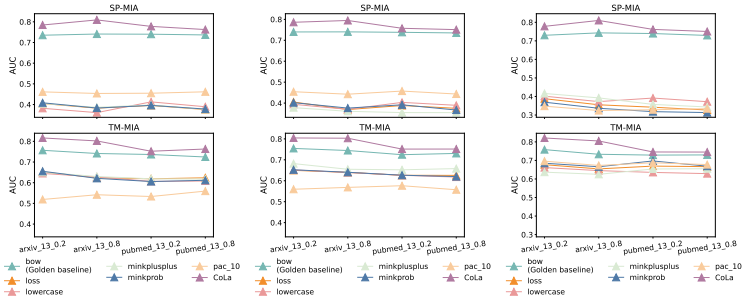


(a) TM-MIA



(b) SP-MIA

Figure 5: The effect of varying the number of clusters used for embedding clustering in the black-box setting. Beyond the default choice of 5, we further consider values between 2 and 10 and report the corresponding AUC curves and TPR@5% FPR.



(a) Pythia-70M (b) Pythia-160M (c) GPT-Neo-125M

Figure 7: For language models, all baseline methods except CoLA perform worse than the *bow* baseline.

Setting	ResNet18-CIFAR100		VGG19-CIFAR10		VGG19-CIFAR100	
	AUC	TPR@5%FPR	AUC	TPR@5%FPR	AUC	TPR@5%FPR
SP-MIA	67.28 ±1.36	19.05 ±1.17	64.98 ±2.03	17.15 ±2.12	70.31 ±1.64	21.23 ±1.81
TM-MIA	85.53 ±2.07	38.46 ±1.43	81.43 ±1.43	40.35 ±1.65	86.67 ±2.36	42.10 ±1.39

Table 2: Subset-aware side-channel attacks under different vision models and datasets.

160m on the arxiv_ngram_13_0.8 dataset. Several observations can be made: first, regardless of the window interval size, the performance under SP-MIA is consistently lower than that under TM-MIA, highlighting its greater challenge. Second, the choice of window interval size does not substantially affect the performance of CoLA. In SP-MIA, increasing the size reduces the exposure count n of each data sample, which makes the inclusion signal coarser and leads to a slight performance drop. However, this drop remains marginal.

Influence of Embedding Model. As a data-centric MIA method, CoLA achieves a clear decoupling from the target model. As discussed earlier, it derives the membership signal by reallocating data combinations based on overfitting at the selection level. For language data, the inherent inconsistency in format and length requires the use of a dedicated embedding model in this reallocation process. To examine the effect of embedding model choice, we conduct an ablation study beyond the default all-MiniLM-L6-v2, considering three alternatives: paraphrase-MiniLM-L6-v2 (paraphrase-MiniLM), distilbert-base-nli-stsb-mean-tokens (distilbert-base), and all-roberta-large-v1. The results are shown in Figure 4b, where the circle size indicates the parameter scale of each embedding model. We observe that different embedding models have a noticeable impact on inference performance, particularly on TPR at low FPR. Moreover, larger model size does not necessarily translate into better performance, highlighting the importance of choosing an appropriate embedding model. Nevertheless, the results remain generally acceptable across all choices (with AUC consistently above 70% and TPR@10% FPR above 25%). How to customize embedding models for MIA under subset selection is a meaningful question, which we leave for future work.

Results under Different Vision Models and Datasets. In Table 2, we conduct subset-aware side-channel attack on the CIFAR-100 dataset with the VGG19 model to verify whether CoLA remains reliable across different vision datasets and models. The selection ratio here is set to 0.2. As can be observed, CoLA consistently works well across various vision model–dataset combinations, revealing its general applicability. Specifically, attacks on VGG19 are more pronounced than on ResNet18 under the same setting, and CIFAR-100 is more vulnerable than CIFAR-10. Moreover, the observation that SP-MIA is more challenging than TM-MIA is consistent with previous findings.

Influence of Clustering. In Figure 5, we study the effect of varying the number of clusters used for embedding clustering in the black-box setting. Beyond the default choice of 5, we consider values between 2 and 10 and report the corresponding AUC curves and TPR@5% FPR. The results show that, for both SP-MIA and TM-MIA, the clustering number has only a marginal effect on performance.

6 Conclusion

In this work, we take the first step toward systematically understanding the privacy risks of subset training. Contrary to the common intuition that training on fewer samples should reduce privacy leakage, we demonstrate that the very choices made during subset selection can themselves become exploitable signals, exposing both included and excluded data to membership inference. To capture this phenomenon, we introduced CoLA, a unified framework that leverages choice patterns to construct robust membership signals. Across both vision and language models, under both subset-aware side-channel and black-box settings, CoLA consistently reveals that subset training does not mitigate but instead amplifies privacy leakage. Our findings highlight that privacy risks extend beyond model outputs to the data–model supply chain itself.

Limitations

Despite CoLA's strong performance, we did not explore many alternative data selection methods, which remains an important direction for future work. We partly address this through consistent protocols, ablations, and experiments across multiple models and datasets, showing that the observed gains are stable.

Acknowledgments

Di Wang and Cheng-Long Wang are supported in part by the funding BAS/1/1689-01-01,RGC/3/7125-01-01, FCC/1/5940-20-05, FCC/1/5940-06-02, and King Abdullah University of Science and Technology (KAUST) – Center of Excellence for Generative AI, under award number 5940 and a gift from Google. Dr. Yinzhi Cao is supported in part by the National Science Foundation (NSF) under grants 23-17185 and 23-19742. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of NSF.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318.

Acxiom. 2026. Acxiom infobase.

Sharat Agarwal, Himanshu Arora, Saket Anand, and Chetan Arora. 2020. Contextual diversity for active learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pages 137–153. Springer.

Mohiuddin Ahmed, Raihan Seraj, and Syed Mohammed Shamsul Islam. 2020. The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, 9(8):1295.

Olivier Bachem, Mario Lucic, and Andreas Krause. 2015. Coresets for nonparametric estimation-the case of dp-means. In *International Conference on Machine Learning*, pages 209–217. PMLR.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023a. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 2397–2430. PMLR.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, and 1 others. 2023b. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. If you use this software, please cite it using these metadata.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. 2022a. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1897–1914. IEEE.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. 2022b. Membership inference attacks from first principles. In *43rd IEEE Symposium on Security and Privacy, SP 2022, San Francisco, CA, USA, May 22-26, 2022*, pages 1897–1914. IEEE.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, and 1 others. 2021. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pages 2633–2650.

Vincent Cohen-Addad, David Saulpic, and Chris Schwiegelshohn. 2021. A new coreset framework for clustering. In *STOC '21: 53rd Annual ACM SIGACT Symposium on Theory of Computing, Virtual Event, Italy, June 21-25, 2021*, pages 169–182. ACM.

C Coleman, C Yeh, S Mussmann, B Mirzasoleiman, P Bailis, P Liang, J Leskovec, and M Zaharia. 2020. Selection via proxy: Efficient data selection for deep learning. In *International Conference on Learning Representations (ICLR)*.

Tian Dong, Bo Zhao, and Lingjuan Lyu. 2022. Privacy for free: How does dataset condensation help privacy? In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 5378–5396. PMLR.

Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. 2024. Do membership inference attacks work on large language models? *arXiv preprint arXiv:2402.07841*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 82 others. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Melanie Ducoffe and Frederic Precioso. 2018. Adversarial active learning for deep networks: a margin based approach. *arXiv preprint arXiv:1802.09841*.

Experian. 2026. Data quality management software & solutions.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, and 1 others. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. Textbooks are all you need. *CoRR*, abs/2306.11644.

Sorami Hisamoto, Matt Post, and Kevin Duh. 2020. Membership inference attacks on sequence-to-sequence models: Is my data in your machine translation system? *Transactions of the Association for Computational Linguistics*, 8:49–63.

Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S Yu, and Xuyun Zhang. 2022. Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)*, 54(11s):1–37.

Lijie Hu, Chenyang Ren, Huanyi Xie, Khouloud Saadi, Shu Yang, Zhen Tan, Jingfeng Zhang, and Di Wang. 2024a. Dissecting representation misalignment in contrastive learning via influence function. *arXiv preprint arXiv:2411.11667*.

Yuzheng Hu, Fan Wu, Qinbin Li, Yunhui Long, Gonzalo Munilla Garrido, Chang Ge, Bolin Ding, David Forsyth, Bo Li, and Dawn Song. 2024b. Sok: Privacy-preserving data synthesis. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 4696–4713.

Mengdi Huai, Di Wang 0015, Chenglin Miao, Jinhui Xu, and Aidong Zhang. 2019. Privacy-aware synthesizing for crowdsourced data. In *IJCAI*, pages 2542–2548.

Krishnateja Killamsetty, Sivasubramanian Durga, Ganesh Ramakrishnan, Abir De, and Rishabh Iyer. 2021a. Grad-match: Gradient matching based data subset selection for efficient deep model training. In *International Conference on Machine Learning*, pages 5464–5474. PMLR.

Krishnateja Killamsetty, Durga Sivasubramanian, Ganesh Ramakrishnan, and Rishabh Iyer. 2021b. Glister: Generalization based data subset selection for efficient and robust learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8110–8118.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, Dublin, Ireland. Association for Computational Linguistics.

Qi Li, Xingyu Li, Xiaodong Cui, Keke Tang, and Peican Zhu. 2023. Hept attack: heuristic perpendicular trial for hard-label attacks under limited query budgets. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 4064–4068.

Qi Li, Runpeng Yu, and Xinchao Wang. 2025. Vid-sme: Membership inference attacks against large video understanding models. *arXiv preprint arXiv:2506.03179*.

Yunji Liang, Yuchen Qin, Qi Li, Xiaokai Yan, Luwen Huangfu, Sagar Samtani, Bin Guo, and Zhiwen Yu. 2022a. An escalated eavesdropping attack on mobile devices via low-resolution vibration signals. *IEEE Transactions on Dependable and Secure Computing*, 20(4):3037–3050.

Yunji Liang, Yuchen Qin, Qi Li, Xiaokai Yan, Zhiwen Yu, Bin Guo, Sagar Samtani, and Yanyong Zhang. 2022b. Accmyrinx: Speech synthesis with non-acoustic sensor. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(3):1–24.

Katerina Margatina, Giorgos Vernikos, Loïc Barrault, and Nikolaos Aletras. 2021. Active learning by acquiring contrastive examples. pages 650–663.

Matthieu Meeus, Igor Shilov, Shubham Jain, Manuel Faysse, Marek Rei, and Yves-Alexandre de Montjoye. 2024. Sok: Membership inference attacks on llms are rushing nowhere (and how to fix it). *arXiv preprint arXiv:2406.17975*.

Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. 2020. Coresets for data-efficient training of machine learning models. In *International Conference on Machine Learning*, pages 6950–6960. PMLR.

Alexander Munteanu, Chris Schwiegelshohn, Christian Sohler, and David Woodruff. 2018. On coresets for logistic regression. *Advances in Neural Information Processing Systems*, 31.

Milad Nasr, Reza Shokri, and Amir Houmansadr. 2018. Machine learning with membership privacy using adversarial regularization. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, CCS '18, page 634–646, New York, NY, USA. Association for Computing Machinery.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. 2021. Deep learning on a data diet: Finding important examples early in training. *Advances in Neural Information Processing Systems*, 34:20596–20607.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, H. Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, and 61 others. 2021. Scaling language models: Methods, analysis & insights from training gopher. *CoRR*, abs/2112.11446.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Chenyang Ren, Yifan Jia, Huanyi Xie, Zhaobin Xu, Tianxing Wei, Liangyu Wang, Lijie Hu, and Di Wang. 2025a. Attributing data for sharpness-aware minimization. *arXiv preprint arXiv:2507.04059*.

Chenyang Ren, Huanyi Xie, Shu Yang, Meng Ding, Lijie Hu, and Di Wang. 2025b. Evaluating data influence in meta learning. *arXiv preprint arXiv:2501.15963*.

Roboflow. 2024. Add tags to images.

Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. 2018. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. *arXiv preprint arXiv:1806.01246*.

Ozan Sener and Silvio Savarese. 2018. Active learning for convolutional neural networks: A core-set approach.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2023. Detecting pretraining data from large language models. *arXiv preprint arXiv:2310.16789*.

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

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari Morcos. 2022. Beyond neural scaling laws: beating power law scaling via data pruning. *Advances in Neural Information Processing Systems*, 35:19523–19536.

Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. 2022. Synthetic data - anonymisation groundhog day. In *31st USENIX Security Symposium, USENIX Security 2022, Boston, MA, USA, August 10-12, 2022*, pages 1451–1468. USENIX Association.

Bowen Tan, Zheng Xu, Eric P. Xing, Zhiting Hu, and Shanshan Wu. 2025. Synthesizing privacy-preserving text data via finetuning *without* finetuning billion-scale LLMs. In *Forty-second International Conference on Machine Learning*.

Haoru Tan, Sitong Wu, Fei Du, Yukang Chen, Zhibin Wang, Fan Wang, and Xiaojuan Qi. 2024. Data pruning via moving-one-sample-out. *Advances in Neural Information Processing Systems*, 36.

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2021. Augmented SBERT: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 296–310, Online. Association for Computational Linguistics.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. 2018. An empirical study of example forgetting during deep neural network learning. In *International Conference on Learning Representations*.

Boris van Breugel, Hao Sun, Zhaozhi Qian, and Mihaela van der Schaar. 2023. Membership inference attacks against synthetic data through overfitting detection.

In *International Conference on Artificial Intelligence and Statistics, 25-27 April 2023, Palau de Congressos, Valencia, Spain*, volume 206 of *Proceedings of Machine Learning Research*, pages 3493–3514. PMLR.

Cheng-Long Wang, Qi Li, Zihang Xiang, Yinzhi Cao, and Di Wang. 2025a. Towards lifecycle unlearning commitment management: Measuring sample-level unlearning completeness. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 6481–6500.

Liangyu Wang, Junxiao Wang, Jie Ren, Zihang Xiang, David E Keyes, and Di Wang. 2025b. Private training large-scale models with efficient dp-sgd. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Max Welling. 2009. Herding dynamical weights to learn. In *Proceedings of the 26th annual international conference on machine learning*, pages 1121–1128.

Alexander Wettig, Aatmik Gupta, Saumya Malik, and Danqi Chen. 2024. Qurating: selecting high-quality data for training language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Zihang Xiang, Tianhao Wang, and Di Wang. 2024. Preserving node-level privacy in graph neural networks. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 4714–4732. IEEE.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Shuo Yang, Zeke Xie, Hanyu Peng, Min Xu, Mingming Sun, and Ping Li. 2022. Dataset pruning: Reducing training data by examining generalization influence. In *The Eleventh International Conference on Learning Representations*.

Wentao Ye, Jiaqi Hu, Liyao Li, Haobo Wang, Gang Chen, and Junbo Zhao. 2024. Data contamination calibration for black-box llms. *arXiv preprint arXiv:2405.11930*.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pages 268–282. IEEE.

Jiaming Zhang, Mingxi Lei, Meng Ding, Mengdi Li, Zihang Xiang, Difei Xu, Jinhui Xu, and Di Wang. 2025a. Towards user-level private reinforcement learning with human feedback. *arXiv preprint arXiv:2502.17515*.

Jingyang Zhang, Jingwei Sun, Eric Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Frank Yang, and Hai Li. 2024. Min-k%++: Improved baseline for detecting pre-training data from large language models. *arXiv preprint arXiv:2404.02936*.

Lin Zhang, Lijie Hu, and Di Wang. 2025b. Mechanistic unveiling of transformer circuits: Self-influence as a key to model reasoning. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1387–1404.

Yunpeng Zhao and Jie Zhang. 2025. Does training with synthetic data truly protect privacy? In *The Thirteenth International Conference on Learning Representations*.

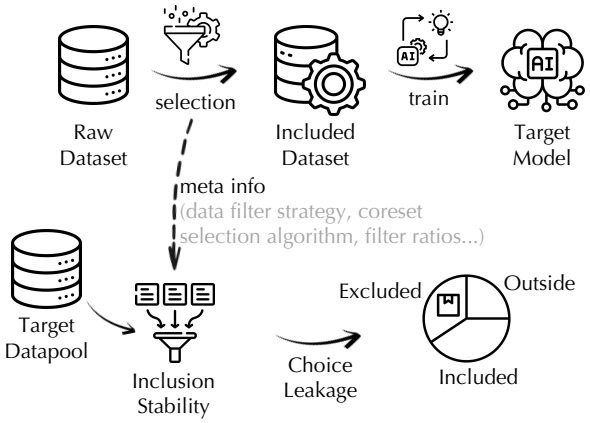


Figure 8: Choice Leakage Attack (CoLA) across the data–model supply chain. CoLA augments conventional MIA by exploiting subset selection metadata leaked along the data–model supply chain. By identifying which samples are more likely to pass selection, it not only strengthens membership inference but also enables adversaries to craft tailored threats.

A Appendix

A.1 The Privacy Threats behind Data-Model Supply Chain

As shown in Figure 8, the data–model supply chain describes the pipeline from raw data collection, through subset selection and model training, to the deployment of a target model. In this process, subset selection plays a central role: only a fraction of the raw dataset is included for training, while others are excluded or remain outside. The metadata of this selection process (e.g., filtering strategies, coreset algorithms, or filter ratios) introduces new privacy surfaces. Such information can inadvertently leak "choice signals" that reveal which samples are more likely to be included in training, thereby extending the privacy risk beyond conventional training data exposure. CoLA (Choice Leakage Attack) directly exploits this vulnerability by using selection metadata to strengthen membership inference. It targets the data–model supply chain by identifying samples more likely to survive selection, thereby increasing membership leakage and enabling more targeted attacks. Once this supply chain is exposed, the entire pipeline from raw data to model outputs becomes more vulnerable.

A.2 Details of the Model and Dataset Usage

Leveraging the rich open-source models in NLP and following the setup in (Meeus et al., 2024), we use deduplicated models from the Pythia (Biderman et al., 2023b) and GPT-Neo (Black et al., 2021) families, specifically pythia-70m, pythia-160m, and gpt-neo-125m, all trained on the MIMIR dataset (Gao et al., 2020; Duan et al., 2024). From the MIMIR dataset, we select two subsets, arXiv and PubMed Central, and evaluate each under two split settings: 'arxiv_ngram_1_0.8', 'arxiv_ngram_13_0.2', 'pubmed_central_ngram_13_0.8', and 'pubmed_central_ngram_13_0.2', where '13_0.8' denotes removing non-member examples that share $>$ 80% 13-gram overlap with members.

In the black-box attacks for vision models, we derive embeddings from the activations just before the final linear layer of a shadow model that shares the target model's architecture. The shadow model is trained using the GradMatch method (Killamsetty et al., 2021a) (distinct from the MIA methods evaluated in our paper) with a selection rate of 0.5. For language models, due to the various lengths of each sequence, we obtain fixed-dimensional embeddings using a dedicated embedding model; by default we use 'all-MiniLM-L6-v2' (Reimers and Gurevych, 2019; Thakur et al., 2021).

For CoLA, the default interval is set to 500 for vision models and 100 for language models, with the window size to be 20,000 and 1,000, respectively. In black-box attacks, the number of clusters is fixed at 5. Ablation studies are provided in Section 5.3.

A.3 Ethics Statement

This work focuses on understanding privacy risks in subset training through systematic analysis of membership inference attacks (MIAs). Our study is purely methodological and does not involve human subjects or personally identifiable information. All datasets used are publicly available benchmark datasets, and we complied with their intended use and licensing terms. We emphasize that the proposed Choice Leakage Attack (CoLA) is presented as a research contribution to highlight potential vulnerabilities in modern training pipelines, not to enable misuse. Our findings are intended to inform the community about inherent privacy risks and to guide the development of stronger defenses.