This article has been accepted for publication in IEEE Transactions on Information Forensics and Security. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2025.3607261

Unveiling Privacy Risks in the Long Tail: Membership Inference in Class Skewness

Hailong Hu, Jun Pang, Yantao Li, and Huafeng Qin

***Abstract*—Real-world datasets often exhibit long-tailed distributions, raising important questions about how privacy risks evolve when machine learning (ML) models are applied to such data. In this work, we present a comprehensive analysis of membership inference attacks in long-tailed scenarios, revealing significant privacy vulnerabilities in tail data. We begin by examining standard ML models trained on long-tailed datasets and identify three key privacy risk effects: amplification, convergence, and polarization. Building on these insights, we extend our analysis to state-of-the-art long-tailed learning methods, such as foundation model-based approaches, offering new perspectives on how these models respond to membership inference attacks across head to tail classes. Finally, we investigate the privacy risks of ML models trained with differential privacy in long-tailed scenarios. Our findings corroborate that, even when ML models are designed to improve tail class performance to match head classes and are protected by differential privacy, tail class data remain particularly vulnerable to membership inference attacks.**

***Index Terms*—Membership inference, class skewness, long-tailed learning, privacy preservation.**

I. Introduction

Real-world multi-class datasets often exhibit long-tailed distributions, where most classes (i.e., tail classes) contain only a few samples, while a small subset of classes (i.e., head classes) have a large amount of data [38]. These long-tailed distributions are ubiquitous across various domains, including image recognition, medical diagnosis, fraud detection, and multilingual text processing [58], [60]. For example, in medical diagnosis, rare diseases such as specific cancer subtypes are far less frequent compared to common non-cancerous conditions, naturally resulting in imbalanced medical datasets with long-tailed distributions [49]. Standard machine learning (ML) models, typically designed for balanced datasets, tend to perform well on head classes but often fail to effectively generalize to tail classes (for example, see Figure 1).

Long-tailed learning has been proposed to address the challenges posed by long-tailed datasets, aiming to improve the representation performance on tail classes without compromising the accuracy of head classes [58]. Early approaches focus on data-level strategies, such as re-sampling techniques where tail classes are over-sampled or head classes are under-sampled to re-balance the dataset [7], [31]. Subsequent research has shifted toward algorithm-level solutions that emphasize representation learning to develop more discriminative features for tail classes. By leveraging the representational power of deep neural networks, various loss functions [4], [20], [39], [40], such as class-balanced loss [12] and logit adjustment loss [33], have been designed to directly address data imbalance during model training. In the current era of foundation models, recent studies leverage large-scale pre-trained models to tackle long-tailed learning challenges. Foundation models, such as CLIP [37] and Vision Transformer [15], trained on extensive and diverse datasets, possess rich feature representations that enhance tail class performance by providing better initializations and transferable features. Building on foundation models, various fine-tuning techniques [9], [21], [24], such as Low-Rank Adapter (LoRA) [22] and Visual Prompt Tuning (VPT) [24], have been integrated into long-tailed learning frameworks to improve computational efficiency and recognition performance, especially for tail classes.

Despite significant advancements in long-tailed learning, the privacy risks remain largely unexplored. Previous studies [18], [19] have investigated memorization issues in standard ML models trained on long-tailed datasets using influence estimation methods. However, these studies are limited to standard ML models and do not extend to ML models that incorporate long-tailed learning techniques. As foundation models gain prominence, it remains unclear how privacy risks evolve when long-tailed learning built on these architectures achieves high performance, where the accuracy on tail classes rivals that of head classes. To address this gap, we conduct a systematic privacy analysis of long-tailed learning through the lens of membership inference attacks [44]. Membership inference, a primary type of privacy attack, seeks to determine whether a data point is included in the training set, providing a practical lower bound on the leakage of an ML model's training set [36], [45]. In this work, by focusing on membership inference attacks, we analyze privacy risks from an attacker's perspective, addressing three key questions to provide new insights into membership leakage in long-tailed scenarios.

Q1: How do standard ML models perform in terms of privacy risks when trained on long-tailed datasets? We begin by conducting an in-depth investigation of membership

This work was supported in part by the National Natural Science Foundation of China under Grants 61976030 and 62072061, the Natural Science Foundation of Chongqing under Grant CSTB2024NSCQ-MSX1118, and the Chongqing Municipal Education Commission under Grants KJZD-M202500804 and KJQN202500812. (Corresponding authors: Yantao Li and Huafeng Qin).

Hailong Hu and Huafeng Qin are with the National Research Base of Intelligent Manufacturing Service, Chongqing Technology and Business University, Chongqing 400067, China (e-mail: huhailong@ctbu.edu.cn, qinhuafengfeng@163.com).

Jun Pang is with the Faculty of Science, Technology and Medicine and the Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg, Esch-sur-Alzette 4365, Luxembourg (e-mail: jun.pang@uni.lu).

Yantao Li is with the College of Computer Science, Chongqing University, Chongqing 400044, China (e-mail: yantaoli@cqu.edu.cn).

Our code is available at: https://github.com/HailongHuPri/LTMI.

inference attacks against standard ML models trained on long-tailed datasets (see Section IV). Within this context, we define standard ML models as those that do not employ specific techniques to address the challenges posed by long-tailed distributions. Through our analysis, we uncover novel privacy risk effects hiding in tail classes from three different perspectives, namely *amplification*, *convergence*, and *polarization* effects.

From a class distribution perspective, the amplification effect reveals that as class size decreases from head to tail classes, the true positive rate (TPR) at a low false positive rate (FPR) significantly increases. This makes tail classes considerably more vulnerable to membership inference attacks compared to head classes. For example, in the CIFAR10-IF500 dataset (see Figure 1c), the TPR@0.1% FPR for tail class 9 reaches 100%, while for head class 0, it hovers around 1%, with an increase of 100 times.

From an attack difficulty perspective, the convergence effect shows that as FPR decreases (indicating a greater attack difficulty), the TPR for head classes drops sharply, while TPR for tail classes remains relatively stable or declines only slightly. This leads to TPR convergence for tail classes, thereby increasing their privacy risks. For instance, when FPR decreases by 50 times, from 5.0% to 0.1% in the CIFAR10-IF500 (see Figure 2c), the TPR for head class 0 decreases by 12 times, from 12.50% to 1.00%, while the TPR for tail class 9 remains steady.

From an imbalance degree perspective, the polarization effect shows that as the imbalance factor increases in long-tailed datasets, the TPR for head classes drops significantly, while the TPR for tail classes rises dramatically. This polarizes membership leakage risks, intensifying the vulnerability of tail classes to membership inference. For example, in CIFAR10-IF500, with an imbalance factor of 500 (see Figure 3), the TPR@0.1% FPR for head class 0 drops by 17 times, from 17% in the balanced CIFAR10 dataset to 1%, while for tail class 9, it increases by 11 times, from 9% to 100%.

Q2: When long-tailed learning techniques are employed to enhance model generalization of tail classes, how does this impact privacy risks? Building on the identified privacy risks in long-tailed datasets, we extend our investigation to assess the membership vulnerabilities in state-of-the-art long-tailed learning methods (see Section V-A). We conduct extensive experiments on foundation model-based long-tailed learning using six different fine-tuning techniques, selected for their proven effectiveness in addressing tail class challenges. The foundation model leverages Vit-B/16 [15] as the backbone, with fine-tuning methods including Classifier fine-tuning, AdaptFormer [9], Adapter [21], Bias tuning [55], LoRA [22] and VPT [24]. Our evaluation demonstrates that while foundation model-based long-tailed learning significantly improves accuracy across all classes, including tail classes, it does not substantially mitigate the privacy risks associated with tail classes. Compared to head classes, the reduction in privacy risks for tail classes is only marginal. For instance, in the CIFAR10-IF500 dataset, the test accuracy exceeds 90% across all classes. However, the TPR@0.1% FPR for tail class 9 drops by only 0.2 times, from 100% to 80%, while the TPR@0.1% FPR for head class 0 decreases by 10 times.

Furthermore, we explore the privacy risks associated with loss function-based long-tailed learning (see Section V-B), as loss function design has received increasing attention for tackling long-tailed challenges. We investigate six different loss functions, including balanced Softmax loss [39], class balanced loss [12], focal loss [40], logit adjustment loss [33], label distribution disentangling loss [20], and label distribution aware margin loss [4], using cross-entropy loss as the baseline. Again, while loss function-based long-tailed learning methods improve the performance of tail classes, they show limited effectiveness in reducing privacy risks for these classes.

Q3: When differential privacy mechanisms are applied to ML models trained on long-tailed datasets, how effective are they in mitigating privacy risks? Differential privacy [16] is the gold standard defense mechanism for mitigating membership inference attacks. To understand its effectiveness in long-tailed scenarios, we systematically evaluate privacy risks for ML models trained with differentially private stochastic gradient descent (DPSGD) [1]. First, we analyze membership inference risks in standard ML models trained with DPSGD (see Section VI-A). Our experimental results show that DPSGD significantly reduces privacy risks across all classes. For instance, under a privacy budget of $\epsilon \approx 3$, the TPR@0.1% FPR drops by about 8 times for both head class 0 and tail class 9. However, tail classes continue to exhibit higher privacy risks compared to head classes. Specifically, TPR@0.1% FPR for tail class 9 is 12.62%, which is 120 times higher than the 0.13% for head class 0. This persistence of the amplification effect under DPSGD indicates that while DPSGD mitigates risk, privacy risks for tail classes are not fully eliminated. Furthermore, this mitigation comes at the cost of model utility, with most classes showing near-zero accuracy, except for a few head classes like class 0 (see Figure 7).

To date, our investigation reveals that foundation model-based long-tailed learning achieves substantial improvements in tail class accuracy while DPSGD effectively mitigates membership inference attacks. Leveraging the complementary strengths of these approaches, we propose an innovative integration of foundation model-based long-tailed learning with DPSGD, referred to as fine-tuning DPSGD. Additionally, we further assess the privacy risks of the fine-tuning DPSGD model (see Section VI-B). Our experimental results demonstrate that fine-tuning DPSGD significantly enhances utility while maintaining privacy under a given privacy budget. For example, under a privacy budget of $\epsilon \approx 3$, the fine-tuning DPSGD model achieves the same TPR@0.1% FPR as the DPSGD model trained from scratch, while achieving a substantial increase in test accuracy, from 0% to 93% (see Figure 8). However, the amplification effect on tail classes remains. The TPR@0.1% FPR for tail class 9 rises to 25%, 250 times compared to 0.1% for head class 0, highlighting the need for more advanced DPSGD training techniques, which we leave for future work.

Contributions. In summary, our contributions are threefold:

(1) We conduct a systematic privacy risk analysis of ML models in long-tailed scenarios, covering standard ML models, foundation model-based long-tailed learning, loss

function-based long-tailed learning, and models trained with differential privacy mechanisms.

(2) We reveal three privacy risks hidden in the tail data of ML models, demonstrating that membership leakage in tail classes remains a significant concern, even when test accuracy exceeds 90% in state-of-the-art foundation model-based long-tailed learning.

(3) We achieve a new state-of-the-art performance for privacy-preserving long-tailed learning, showing that while privacy risks and test performance can be simultaneously improved in long-tailed scenarios, the amplified privacy risks for tail classes necessitate novel privacy mechanisms.

II. Related Work

Long-tailed learning. Long-tailed data refers to datasets where a few head classes have many data samples, while many tail classes have few data samples. The objective of long-tailed learning is to improve the performance of tail classes so that it matches or closely approaches that of head classes [58]. Early methods in long-tailed learning usually focus on rebalancing the dataset distribution through data resampling techniques. Typical examples include under-sampling from the head classes to reduce their dominance and over-sampling from the tail classes to increase their representation [7], [17], [31]. Beyond data resampling, data augmentation methods have been explored to address long-tailed data issues. These methods involve synthetically increasing the diversity of tail class samples through various transformations or by generating new data points [48], [56]. With the success of deep neural networks, researchers shifted their focus toward exploring novel neural architectures, loss functions, and training protocols to optimize the performance of tail classes. For instance, Ross et al. [40] proposed a novel focal loss function that re-weights the loss based on prediction probabilities. Kang et al. [27] developed a decoupled training technique where the learning process is divided into two stages: representation learning and classification. In the era of foundation models, long-tailed learning has achieved new state-of-the-art advancements. Relying on the generalization capabilities of the foundation model, researchers have been able to develop novel prompt learning and fine-tuning techniques to improve performance on long-tailed data [14], [43]. However, existing works predominantly focus on improving model accuracy across head and tail classes, and privacy considerations have been largely overlooked in the context of long-tailed learning. In this work, we will provide a comprehensive privacy analysis to reveal risks of membership leakage in long-tailed learning.

Membership inference attacks. Membership inference attacks aim to determine whether a specific data point was included in an ML model's training set [6], [29], [41], [44], [51]. Shokri et al. [44] introduced the first membership inference attack on ML models, where a shadow model-based method was developed to reveal membership information of ML models. Since then, numerous attack methods have been proposed based on different threat models, such as reducing the number of shadow models [41], utilizing label information [11], [30], exploiting gradient information [35], and analyzing loss trajectories [32]. However, these efforts have predominantly concentrated on membership inference attacks against classification models trained on balanced datasets. In contrast, our study investigates membership inference attacks against ML models trained on long-tailed datasets, filling a crucial gap in the current literature and extending the understanding of these vulnerabilities within real-world long-tailed distributions.

While several works have sought to explore the memorization issues in standard ML models trained on long-tailed data [18], [19], they have largely overlooked the privacy vulnerabilities in long-tailed learning, where the performance of the tail classes has been significantly improved. There is one work [46] studying membership inference attacks against standard ML models trained on long-tailed data. Similarly, this work does not consider membership inference vulnerabilities in long-tailed learning and a significant limitation of this work lies in its use of average-case accuracy as the primary evaluation metric. However, this metric has been shown to inadequately capture worst-case privacy vulnerabilities in membership inference attacks [2], [5]. In addition, several works study membership inference attacks against recommender systems in which user data exhibit a long-tail distribution [57], [59]. However, they do not consider scenarios involving long-tailed learning and differential privacy. Furthermore, although some works have investigated the fairness concerns of standard ML models trained on long-tailed datasets in the context of differential privacy [3], [46], our work reveals privacy issues and analyzes the relationship between accuracy and privacy risks of tail classes. In other words, our work provides a comprehensive privacy analysis in long-tailed scenarios through the lens of membership inference attacks. Unlike previous studies, we not only identify and quantify the privacy vulnerabilities faced by models trained on long-tailed data, but also offer an in-depth exploration of the relationship between accuracy and privacy risks for both head and tail classes. Furthermore, we extend our analysis to examine membership leakage in both cutting-edge foundation model-based long-tailed learning and classical long-tailed learning methods. Finally, our work also pioneers the integration of foundation models with differential privacy and existing long-tailed strategies, achieving state-of-the-art performance in terms of both utility and privacy.

III. Preliminaries

A. Machine Learning Notations

Consider a training dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$, where each sample $x_i \in \mathbb{R}^d$ represents a d-dimensional input vector, and $y_i \in \{1, 2, \dots, C\}$ denotes the corresponding class label where C represents the total number of classes, and N is the total number of samples. We define $D_k \subset D$ as the subset of training samples belonging to class k, with $|D_k|$ indicating the number of samples in class k. Without loss of generality, we assume classes are indexed in descending order of sample size. Thus, for all $k \in \{1, 2, \dots, C-1\}$, the condition $|D_k| \geq |D_{k+1}|$ holds, ensuring that class D_1 has the largest number of samples and class D_C the smallest.

For a balanced dataset, the number of samples per class is approximately equal, i.e., $|D_k| \approx \frac{N}{C}$ for all k. In contrast, for a long-tailed dataset, we introduce an imbalance factor β, which quantifies the degree of imbalances as the ratio of the largest to the smallest class sizes: $\beta = \frac{\max_k |D_k|}{\min_k |D_k|} = \frac{D_1}{D_C}$. In such datasets, the number of samples in class k can be calculated by an exponential distribution:

$$|D_k| = |D_1| \times \beta^{-\frac{k-1}{C-1}}, \tag{1}$$

where $|D_1|$ represents the number of samples of the largest class. Common values for β include 50, 100, and 500. Based on class size, we categorize classes with $|D_k| \geq 100$ as head classes and those with $|D_k| \leq 20$ as tailed classes.

The objective of a classification model trained on D is to learn a function $f : \mathbb{R}^d \rightarrow \{1, 2, \ldots, C\}$, such that, for an input x_i, the predicted label $\hat{y}_i = f(x_i)$ closely approximates the true label y_i.

B. Membership Inference Attacks

Membership inference attacks aim to determine whether a specific target sample x was part of a machine learning model's training dataset D. Given a trained classification model f, an adversary seeks to devise an attack algorithm $\mathcal{A}$ to infer whether x was included in D.

In this work, we adopt the Likelihood Ratio Attack (LiRA) introduced by Carlini et al. [5], due to its demonstrated effectiveness in membership inference attacks. LiRA operates by first training shadow models to replicate the behavior of the target model f. These shadow models are divided into two categories: those trained on datasets that contain the target sample x (referred to as IN models) and those trained on datasets without the target sample x (referred to as OUT models). Then, LiRA models the outputs of the IN and OUT shadow models using two Gaussian distributions, $\mathcal{N}(\mu_{x,\text{in}}, \sigma^2_{x,\text{in}})$ and $\mathcal{N}(\mu_{x,\text{out}}, \sigma^2_{x,\text{out}})$, respectively. The likelihood of a target sample x is computed based on these distributions, leading to the following likelihood ratio:

$$\mathcal{A}(f, x) = \frac{\mathcal{N}(f(x) \mid \mu_{x,\text{in}}, \sigma^2_{x,\text{in}})}{\mathcal{N}(f(x) \mid \mu_{x,\text{out}}, \sigma^2_{x,\text{out}})}, \tag{2}$$

where $f(x)$ denotes the output of the classification model for the target sample x. The adversary uses this ratio to make the final membership decision.

The effectiveness of membership inference attacks is typically evaluated using metrics such as accuracy, logarithmic-scale ROC curve, and true positive rate (TPR) at low false positive rate (FPR). In this work, we primarily report TPR at 0.1% FPR (denoted as TPR@0.1% FPR), as this metric effectively captures privacy risks in the worst-case scenarios.

C. Threat Model

A threat model outlines the assumptions and capabilities of adversaries. In this work, our threat model aligns with the common practices widely used in prior works on membership inference attacks [5], [41], [44]. Specifically, we assume that adversaries have black-box access to the target ML model, allowing them to query the model with input samples and observe its outputs, such as the logit scores. Additionally, we assume that adversaries possess shadow datasets that follow the same distribution as the training set of the target model. Furthermore, adversaries are assumed to have the knowledge of the target model, such as the model's architectures and training protocols.

D. Datasets

In this work, we conduct experiments on the CIFAR10 and CIFAR100 datasets, as they allow us to control for imbalance factors and computational efficiency, facilitating a more comprehensive analysis. These datasets are also widely used in the fields of privacy and long-tailed learning [2], [5], [19], [28]. Specifically, we utilize CIFAR10 and CIFAR100 to construct both balanced and long-tailed versions. For long-tailed datasets, we label them according to the dataset name with the imbalance factor, such as CIFAR10-IF100. Furthermore, we set $D_1 = 2,500$ for CIFAR10 and $D_1 = 250$ for CIFAR100. The number of samples in subsequent classes is computed using Equation 1. For balanced datasets, we maintain a total sample size approximately equal to that of the long-tailed datasets while ensuring an equal number of samples across all classes.

IV. Membership Inference on Long-tailed Data

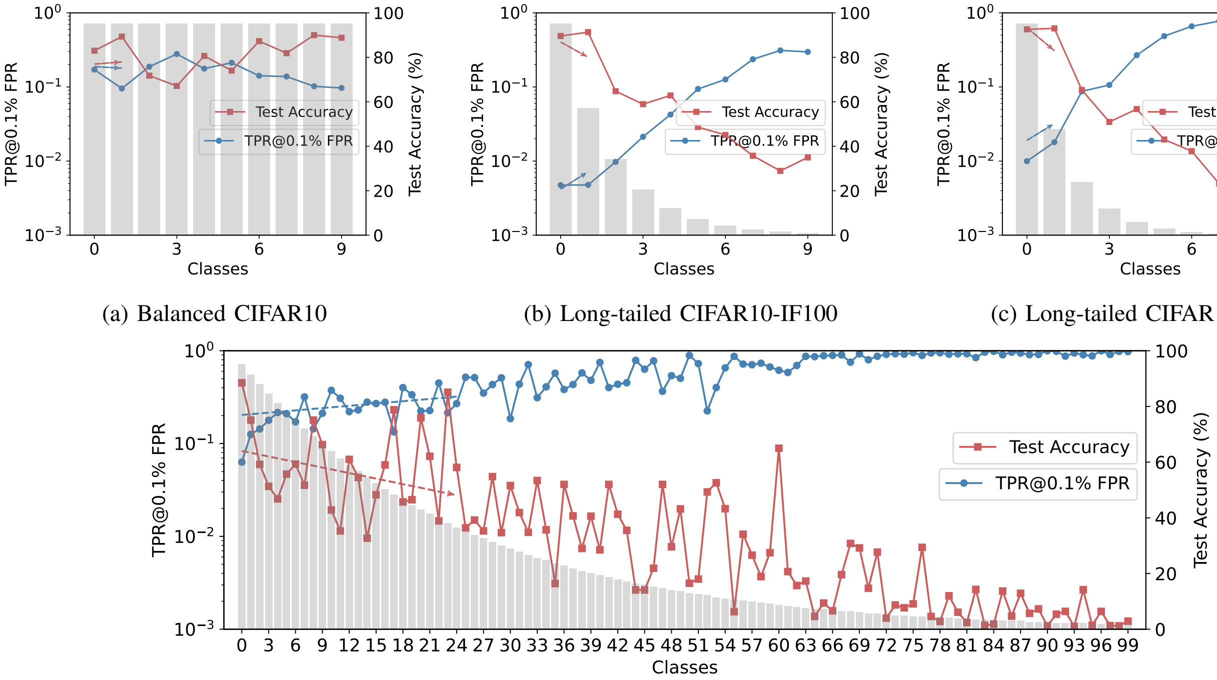
In this section, we explore membership inference attacks against standard ML models trained on long-tailed data, which seeks to answer question Q1. We begin by outlining experimental settings and conclude with our findings, which reveal three novel privacy effects: amplification, convergence, and polarization.

In this work, a standard ML model refers to one trained on a dataset — whether balanced or long-tailed — without employing any specialized mechanisms to address the unique challenges posed by long-tailed distributions.

A. Experimental Settings

Target models. We utilize the WideResNet architecture [54] with a widening factor of 4 and a depth of 16 as the target model. Optimization is performed using stochastic gradient descent (SGD) with a learning rate of 0.1 and a weight decay of 5×10^{-4}. The models are trained using cross-entropy loss for 200 epochs, and the version achieving the highest test accuracy is selected as the final target model. In this work, these models are also referred to as standard ML models.

Attack configurations. Following the experimental setup of previous work [5], we evaluate the test accuracy of target models using the official test splits of CIFAR10 and CIFAR100. The remaining data are split equally: one half for training the models (serving as member samples) and the other half as nonmember samples. We randomly choose 64 subsets and train a total of 64 ML models. One of these models is designated as the target model, while the remaining 63 models serve as shadow models. Attack performance is evaluated using an equal number of member and nonmember samples, and TPRs at low FPRs are reported. All reported values are averaged over 10 different target models.



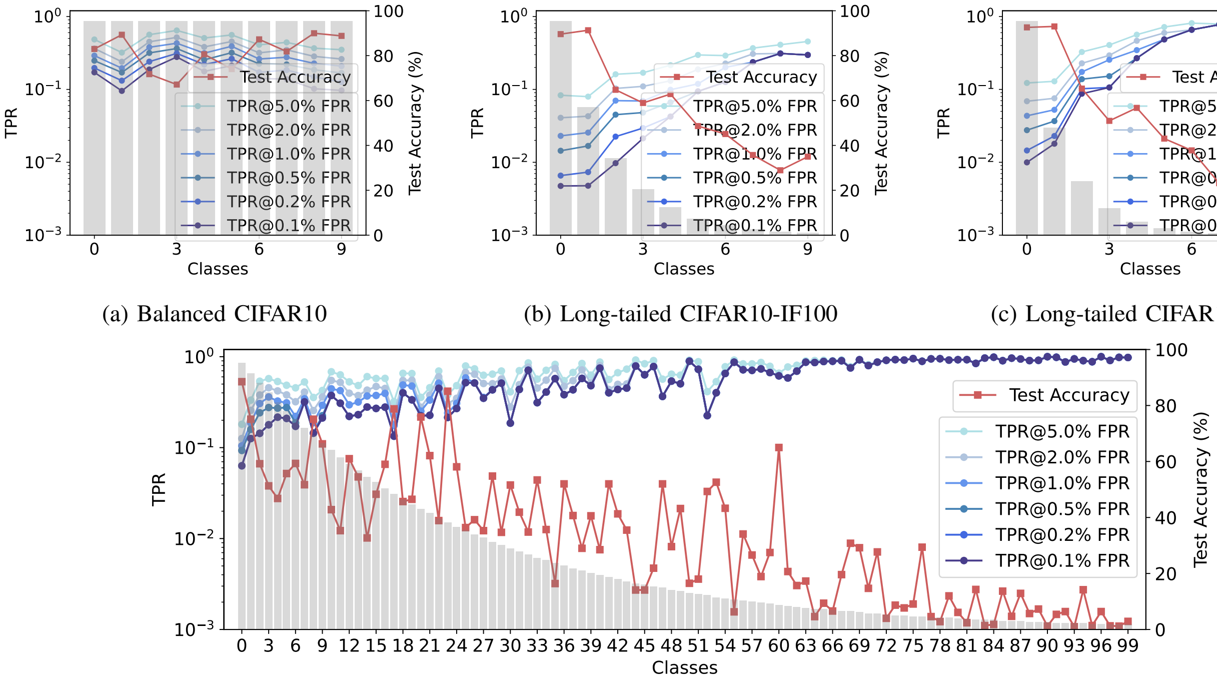
(a) Balanced CIFAR10

(b) Long-tailed CIFAR10-IF100

(c) Long-tailed CIFAR10-IF500

(d) Long-tailed CIFAR100-IF50

Fig. 1: **Attack performance and test accuracy across different classes.** In CIFAR10, both test accuracy and TPR@0.1% FPR exhibit balanced performance across classes (Figure 1a). However, in Figures 1b and 1c, test accuracy drops sharply while TPR@0.1% FPR rises significantly from head to tail classes (i.e., from class 0 to class 9), indicating increased privacy risks. A similar trend is observed in CIFAR100-IF50, and results on the balanced CIFAR100 are provided in the Supplementary Material (Figure A.1). The gray bar represents the number of samples in each class. The reported TPR@0.1% FPR and test accuracy are averaged over 10 target models.



(a) Balanced CIFAR10

(b) Long-tailed CIFAR10-IF100

(c) Long-tailed CIFAR10-IF500

(d) Long-tailed CIFAR100-IF50.

Fig. 2: **Attack performance and test accuracy across different FPRs.** In the balanced dataset (Figure 2a), TPR decreases fairly uniformly as FPR decreases. However, in the long-tailed dataset, TPR drops significantly in head classes compared to tail classes. Results for the balanced CIFAR100 are shown in the Supplementary Material (Figure A.2).

B. Results

1) Attack performance on different classes: Figure 1 illustrates the attack performance and test accuracy of target models across different classes for both balanced and long-tailed versions of the CIFAR10 and CIFAR100 datasets. In long-tailed datasets, we observe a significant drop in model accuracy from head to tail classes, accompanied by an increase in TPR values. Conversely, in the balanced datasets, both TPR and accuracy exhibit relatively stable fluctuations in all classes.

As shown in Figure 1a, in the balanced CIFAR10 dataset, test accuracy stabilizes approximately at 80%, while TPR@0.1% FPR remains around 16%, indicating consistent performance across all classes without significant variation. In contrast, in the long-tailed dataset, such as CIFAR10-IF100 (Figure 1b), TPR@0.1% FPR shows a substantial increase, rising from 0.47% in class 0 to 29.70% in class 9. This is accompanied by a significant decline in test accuracy, from 89.59% in class 0 to 34.99% in class 9. Furthermore, as shown in Figure 1c, in the long-tailed CIFAR10-IF500 dataset with an imbalance factor of 500, TPR@0.1% FPR escalates dramatically from 1.00% in class 0 to an alarming 100.00% in class 9. In contrast, test accuracy exhibits a sharp decline, from 92.59% in class 0 to a mere 3.50% in class 9. Similar trends are observed in the CIFAR100 dataset, which contains 100 classes. As depicted in Figure 1d, in the long-tailed CIFAR100-IF50, as the number of samples per class decreases from class 0 to class 99, accuracy declines sharply from 88.50% to 2.90%, while TPR@0.1% FPR rises markedly from 6.31% to 98.00%. In contrast, in the balanced CIFAR100 dataset, test accuracy and TPR@0.1% FPR fluctuate around 39.14% and 57.49%, respectively, as shown in Figure A.1 in the Supplement.

We refer to this observation as the tail class privacy risk amplification effect, or simply the *amplification effect*.

> **Amplification Effect**
>
> In a long-tailed dataset, as the number of samples per class decreases from head to tail classes, the TPR at low FPR shows a notable increase, thereby amplifying membership leakage risks of tail classes.

2) Attack performance on different FPRs: Figure 2 shows the attack performance and test accuracy of target models across different FPRs on both balanced and long-tailed versions of CIFAR10 and CIFAR100 datasets. A lower FPR indicates a higher level of attack difficulty, as it means that adversaries must operate with fewer allowable mistakes. Overall, we observe that in long-tailed datasets, as FPR decreases, TPR drops sharply in head classes but remains relatively high in tail classes, leading to convergence. In contrast, in balanced datasets, TPR declines more uniformly across all classes as FPR decreases.

For example, as demonstrated in Figure 2b, in the CIFAR10-IF100 dataset, TPR for head class 0 drops sharply from 8.31% to 0.47% as FPR decreases from 5.00% to 0.10%. In contrast, TPR for tail class 9 decreases modestly from 45.45% to 29.70% as FPR reduces from 5.00% to 2.00%, stabilizing at around 29.70% as FPR continues to drop to 0.10%. Despite a 50-fold decrease in the FPR, TPR for head class 0 falls approximately 20-fold, while for tail class 9, it decreases by only 0.65-fold. In the CIFAR10-IF500 dataset (Figure 2c), these differences are even more pronounced. TPR for head class 0 drops from 12.25% to 1.00% as FPR declines from 5.00% to 0.10%, while TPR for tail class 9 remains nearly constant at 100.00%, showing minimal reduction even with decreasing FPR.

A similar trend is observed in the CIFAR100-IF50 dataset, as shown in Figure 2d. For head class 0, TPR decreases significantly from 18.05% to 6.31%, a 3-fold reduction, as FPR decreases from 5.00% to 0.10%. In contrast, TPR for tail class 99 remains stable around 98.00%. In fact, TPR across all tail classes (from class 60 to class 99, each with fewer than 20 samples) remains nearly constant, showing minimal sensitivity to decreasing FPR.

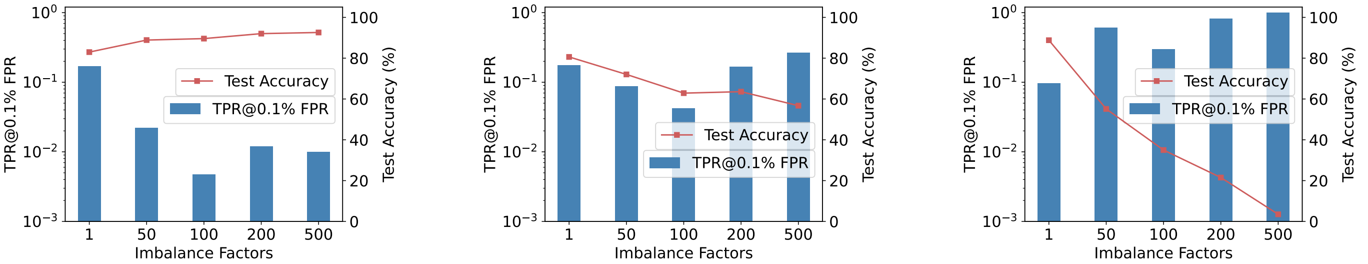
We refer to this observation as the tail class privacy risk convergence effect, or simply the *convergence effect*.

> **Convergence Effect**
>
> In a long-tailed dataset, as the FPR decreases, the TPR of head classes drops significantly, while the TPR of tail classes decreases only slightly or remains relatively stable, resulting in a convergence of TPR among the tail classes.

3) Attack performance on different imbalance factors: Figure 3 illustrates the attack performance and test accuracy of target models across different imbalance factors on the CIFAR10 dataset. We focus on three classes: class 0 with thousands of samples (head class), class 4 with hundreds of samples (medium class), and class 9 with tens of samples (tail class). Overall, as the imbalance factor increases, we observe distinct trends across these class groups: The TPR of head classes decreases while their accuracy improves slightly; the TPR of medium classes increases slightly with a minor decline in accuracy; and the TPR of tail classes rises significantly as their accuracy sharply drops.

Specifically, as shown in Figure 3a, the TPR@0.1% FPR for head class 0 decreases significantly from 17.14% in the balanced dataset (imbalance factor of 1) to 2.22% at an imbalance factor of 50, and further drops to just 1.00% at an imbalance factor of 500. In contrast, as illustrated in Figure 3c, the TPR for tail class 9 rises sharply with imbalance factors, rising from 9.69% at an imbalance factor of 1 to 60.69% at an imbalance factor of 50, and eventually reaching 100.00% at an imbalance factor of 500. Interestingly, even when the accuracy for class 9 falls to just 3.50% at an imbalance factor of 500, the TPR remains extremely high. Similarly, at an imbalance factor of 200, the TPR reaches approximately 1.89%, despite class 9 accuracy being only 21.51%. For medium class 4, as depicted in Figure 3b, the TPR initially declines from 17.68% at an imbalance factor of 1 to 8.80% at an imbalance factor of 50, but then fluctuates with increasing imbalance, eventually rising to 26.84% at an imbalance factor of 500.

We term this the imbalance-induced tail class privacy risk polarization effect, or simply the *polarization effect*.



(a) Class 0

(b) Class 4

(c) Class 9

Fig. 3: **Attack performance and test accuracy across different imbalance factors.** Three classes in CIFAR10 are selected based on sample size. As imbalance factors increase, TPR@0.1% FPR decreases significantly for class 0, remains relatively stable for class 4, and increases sharply for class 9.

> **Polarization Effect**
>
> In long-tailed datasets, as the imbalance factor increases, the TPR of head classes declines significantly, while the TPR of tail classes rises sharply, leading to a polarization of membership leakage risks between head and tail classes.

Takeaway. In summary, we answer question Q1 by identifying three privacy risk effects related to tail classes in standard ML models trained on long-tailed datasets, each from a unique perspective. From the perspective of class distribution, the amplification effect reveals that membership leakage risks are significantly higher for tail classes than for head classes, thereby amplifying privacy risks in tail classes. From the perspective of attack difficulty, the convergence effect indicates that as attack scenarios become stricter — requiring fewer mistakes by adversaries (i.e., lower FPR) — TPR in head classes decreases significantly more than in tail classes, resulting in a convergence of TPR values in tail classes. From the perspective of imbalance degrees, the polarization effect illustrates that as the degree of imbalance increases, TPR in head classes decreases while TPR in tail classes increases, leading to polarization between head and tail classes.

V. Membership Inference in Long-tailed Learning

In this section, building on the insights discussed in Section IV, we investigate membership inference attacks against long-tailed learning, which focuses on answering question Q2. We begin by exploring foundation model-based long-tailed learning in Section V-A, given its state-of-the-art performance. Additionally, in Section V-B, we investigate various loss function-based methods for long-tailed learning, considering their foundational significance and widespread application in this field.

A. Foundation Model-based Long-tailed Learning

In the era of foundation models, fine-tuning methods for addressing long-tailed learning have gained significant attention. In this subsection, we introduce common fine-tuning methods for foundation model-based long-tailed learning and present our experimental results.

1) Fine-tuning: Foundation models, such as CLIP [37] and ViT [15], are typically trained on extensive datasets, with CLIP, for instance, utilizing a dataset of 400 million image-text pairs. Leveraging the Transformer architecture [15], [47], these models demonstrate exceptional generalization capabilities; however, with parameter sizes ranging from millions to billions, they are highly computationally intensive. To address this, fine-tuning methods have been proposed to effectively adapt these models for various downstream applications.

Figure 4 provides an overview of the components of a foundation model. The process begins by dividing the input image into m patches, which are then passed through an embedding layer to produce corresponding token representations. These embedded tokens are fed sequentially through a series of transformer blocks, which progressively refine the token representations, ultimately generating high-level feature representations. These extracted features serve as the input to a classifier for the final task-specific prediction. In this work, we systematically explore six different fine-tuning methods, as described below:

(1) Traditional fine-tuning. An intuitive method is to fine-tune the last k layers of Transformer blocks, similar to fine-tuning ResNet-based models. However, when applied to foundation models, this method is computationally expensive due to the large number of parameters. In this work, we adopt a lightweight classifier fine-tuning method that freezes all Transformer blocks and only fine-tunes the classifier.

In addition to the traditional fine-tuning, parameter-efficient methods have been introduced to reduce the computational costs by limiting the number of trainable parameters. In this work, we consider five widely-used parameter-efficient methods.

(2) Bias tuning. Zaken et al. [55] proposed fine-tuning only the bias components of the foundation model, effectively preserving learned representations while minimizing the number of parameters requiring updates. Specifically, given a projection function at one layer $XW + b$, where W is the weight matrix and b is the bias vector, bias tuning focuses on optimizing b while keeping W fixed.

(3) Visual prompt tuning (VPT). Jia et al. [24] proposed adding learnable prompts into the input space for fine-tuning while keeping the Transformer blocks frozen. Formally, given

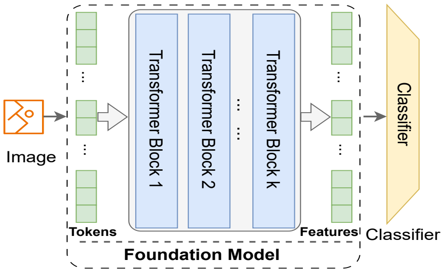


Fig. 4: Components of a foundation model.

an input X at a layer, a learnable prompt p is prepended, resulting in a modified input $X' = [p, X]$. In this work, we adopt a deep version of VPT, where prompts are added to the input space of every Transformer layer.

(4) Adapter. Houlsby et al. [21] introduced adapter modules into each Transformer layer. For a given input X, the adapter applies layer normalization LN and uses two weight matrices, $W_{\text{down}} \in \mathbb{R}^{d\times r}$ and $W_{\text{up}} \in \mathbb{R}^{r\times d}$ (where $r \ll d$). The output is computed as: $\text{Adapter}(X) = \text{ReLU}(\text{LN}(X)W_{\text{down}})W_{\text{up}}$. The adapter serves as a bottleneck, allowing only the adapter parameters to be updated during fine-tuning, while the rest of the model remains fixed.

(5) Low-Rank Adapter (LoRA). Hu et al. [22] proposed the low-rank adapter, which introduces trainable rank decomposition matrices into each Transformer layer. Given a weight matrix $W_0 \in \mathbb{R}^{d\times k}$, the updated weights can be expressed as: $W' = W_0 + W_{\text{down}}W_{\text{up}}$, where $W_{\text{down}} \in \mathbb{R}^{d\times r}$ and $W_{\text{up}} \in \mathbb{R}^{r\times k}$ with $r \ll \min(d, k)$. During fine-tuning, LoRA keeps the pre-trained W_0 frozen and only updates the rank decomposition matrices W_{down} and W_{up}.

(6) AdaptFormer. Chen et al. [9] introduced AdaptFormer, a parallel structure that replaces the traditional sequential adapter. Formally, given a traditional adapter $\text{Adapter}(X)$, the final output is computed as: $\text{MLP}(\text{LN}(X)) + s \cdot \text{Adapter}(X) + X$, where s is a scaling parameter, either set manually or learned during fine-tuning.

2) Long-tailed learning: In addition to the fine-tuning methods discussed, effective long-tailed learning with foundation models often integrates additional strategies to further enhance the performance of tail classes.

As shown in Figure 4, foundation model-based long-tailed learning typically involves selecting appropriate model backbones, classifier types, and loss functions during fine-tuning, as well as incorporating post-processing methods during prediction. In this work, we adopt the LIFT framework proposed by Shi et al. [43] for foundation model-based long-tailed learning. Specifically, we employ ViT-B/16 [15] as the foundation model backbone. For classification, we utilize a cosine classifier, defined as:$z_k = \sigma \cdot \frac{w_k^\top f}{\|w_k\|_2 \|f\|_2}$, where z_k represents the prediction for class k, w_k is the corresponding class weight, f is the feature vector, and σ is a scaling hyperparameter. Compared to a linear classifier, this cosine classifier more effectively balances feature representations by normalizing the classifier weights. During fine-tuning, we use logit adjustment (LA) loss (see Table I), as it has been shown to outperform cross-entropy loss by promoting more balanced feature learning across the class distribution. In Section V-B, we provide a detailed analysis of the effects of different loss functions on accuracy and privacy. During prediction, we apply random perturbations to each input to improve the model's generalization performance further.

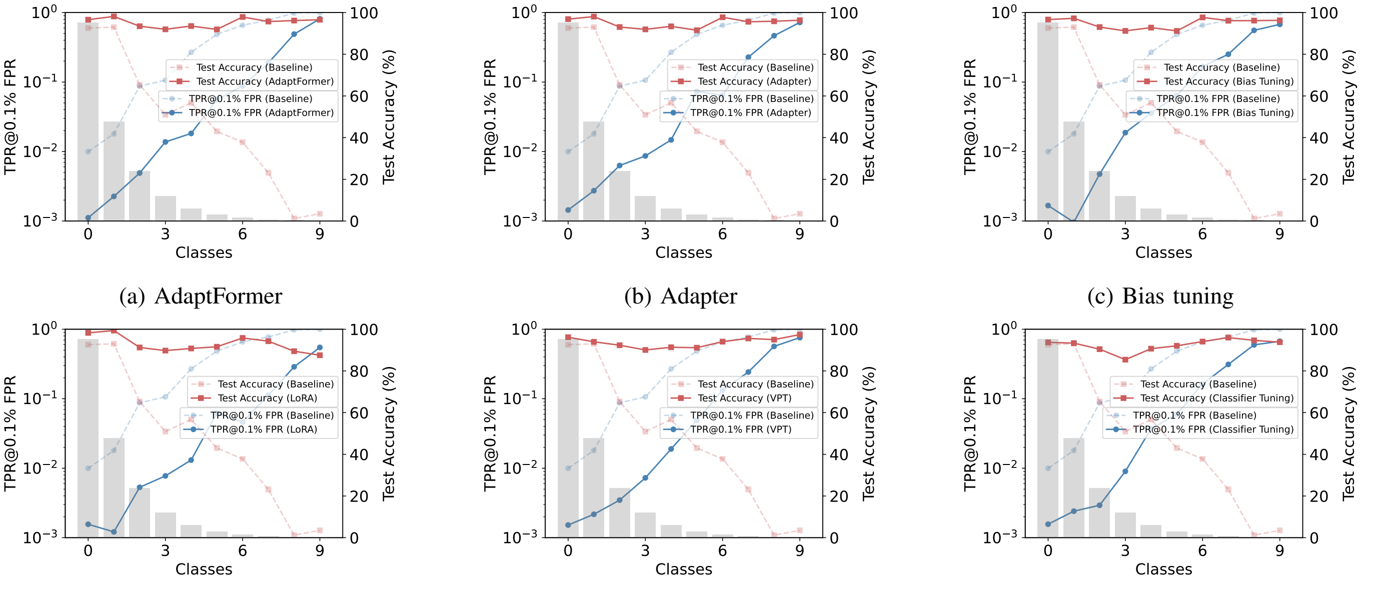
3) Experimental settings: For target models, we train six types of models using six different fine-tuning methods, including traditional classifier fine-tuning and five parameter-efficient fine-tuning methods: Adaptformer, Adapter, Bias tuning, LoRA, and VPT. We adopt all recommended hyperparameters from the LIFT framework proposed by Shi et al [43]. The cosine classifier is initialized using semantic knowledge from CLIP, and the number of training epochs is set to 10. For membership inference attacks, we follow the same settings and procedures illustrated in Section IV-A.

4) Results: Figure 5 presents a comprehensive analysis of foundation model-based long-tailed learning across six different fine-tuning methods on the CIFAR10-IF500 dataset. The baseline, depicted by the dashed lines, represents the WideResNet model trained from scratch. Several key observations are as follows:

(a) Test accuracy. Foundation model-based long-tailed learning demonstrates significant test accuracy improvements across all classes and all fine-tuning methods compared to the baseline. For instance, as shown in Figure 5, the most significant gains are observed in tail classes, such as class 9, where accuracy increases from 3.50% to over 90%, nearly approaching head class performance levels. While head classes see only modest increases in test accuracy, they remain stable at around 90%, underscoring the positive impact of foundation model-based long-tailed learning on model generalization, particularly for tail classes.

(b) TPR@0.1% FPR. Across all classes and fine-tuning methods, foundation model-based long-tailed learning shows a general reduction in TPR@0.1% FPR compared to the baseline, likely due to improved model generalization. The most obvious decreases occur in head classes; for instance, in class 0, TPR@0.1% FPR drops by approximately 10 times, from 1.00% to 0.1%, effectively reaching random guessing levels for adversaries. However, the decrease in TPR@0.1% FPR for tail classes is less significant. For example, in the worst-case scenario for adversaries, TPR in LoRA decreases only by about half, from 100% to 54%, indicating that the tail class privacy risk amplification effect persists, with membership leakage for tail classes still up to 540 times greater than for head classes. Additional results on CIFAR100-IF50 are included in Supplementary Material (Figure A.3).

Takeaway. In summary, we tackle question Q2 by investigating the state-of-the-art foundation model-based long-tailed learning. We reveal that foundation model-based long-tailed learning is highly effective for enhancing overall accuracy across both head and tail classes, demonstrating strong generalization capabilities. However, it does not adequately address membership leakage risks in tail classes. Our work highlights the privacy vulnerabilities in tail classes, emphasizing the need for strategies that can enhance generalization while simultaneously protecting against privacy attacks in long-tailed learning.



(a) AdaptFormer

(b) Adapter

(c) Bias tuning

(d) LoRA

(e) VPT

(f) Classifier fine-tuning

Fig. 5: **Attack performance and test accuracy in foundation model-based long-tailed learning on CIFAR10-IF500.** All fine-tuning methods enhance test accuracy for all classes but do not effectively reduce TPR for tail classes.

B. Loss Function-based Long-tailed Learning

In long-tailed learning, the design of loss functions has received significant attention due to their critical role in improving model performance on highly imbalanced datasets. In this section, we systematically examine the privacy risks of long-tailed learning across different loss functions.

1) Introduction: Standard ML models are typically trained using the cross-entropy (CE) loss function, which is extensively applied in various classification tasks. However, the CE loss function does not account for class imbalance in long-tailed datasets, often resulting in poor performance on tail classes, as models tend to favor head classes with more training samples. To address this, researchers have developed various loss functions that explicitly or implicitly balance each class's contribution during training.

Consider a data sample x with predicted logits z and corresponding softmax probabilities p. let z_k and p_k denote the predicted logits and probabilities for class k, respectively, π_k represent the sample frequency for class y, defined as $\pi_k = \frac{|D_k|}{|D|}$. In this work, we explore six widely used loss functions, summarized in Table I.

(1) Balanced Softmax (BS) loss [39]. BS loss adjusts the predicted logits by scaling them with the corresponding label frequencies π_k, to mitigate class imbalance effectively.

(2) Class Balanced (CB) loss [12]. CB loss introduces a re-weighting mechanism that is inversely proportional to the effective number of samples for each class. This effective number is calculated with an exponential function of the training label counts.

(3) Focal loss [40]. Focal loss addresses class imbalance by applying a modulating factor $(1 - p_k)^\gamma$, which reduces the loss for well-classified samples (often from head classes) and emphasizes difficult-to-classify examples, typically found in tail classes.

(4) Logit Adjustment (LA) loss [33]. LA loss incorporates a margin parameter m_i for each class, adjusting the softmax operation to improve robustness against imbalanced data by expanding margins for tail classes.

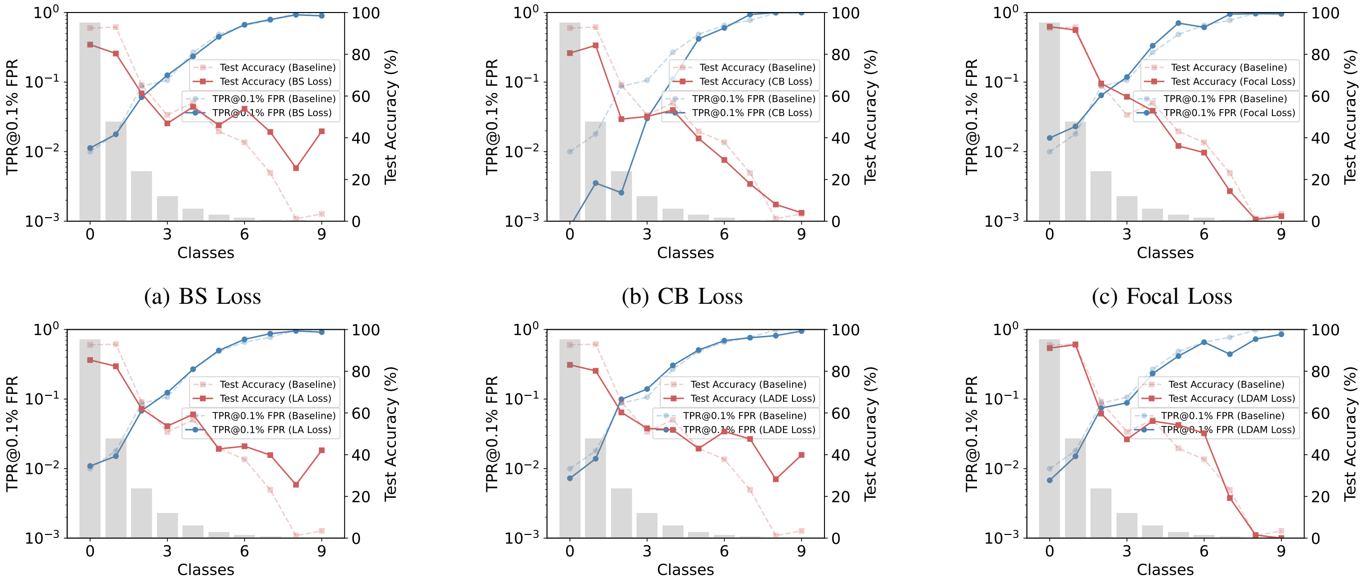
(5) Label Distribution Disentangling (LADE) loss [20]. LADE loss combines BS loss with a regularization term to disentangle logits from the source label distribution, enhancing the model's generalization to tail classes.

(6) Label Distribution Aware Margin (LDAM) loss [4]. LDAM loss applies class-dependent adjustments based on the training label frequencies, providing larger margins for tail classes to address class imbalance and boost performance on underrepresented classes.

TABLE I: Loss formulations. m_k is a margin parameter for class k. γ and τ are hyperparameters. L_{LADER} is a regularization term.

Loss	**Formulation**
CE Loss	$L_{\text{CE}} = -\log(p_k)$
BS Loss	$L_{\text{BS}} = -\log\left(\frac{\pi_k \exp(z_k)}{\sum_{j=1}^{C} \pi_j \exp(z_j)}\right)$
CB Loss	$L_{\text{CB}} = -\frac{1-\gamma}{1-\gamma^{\lvert D_k \rvert}}(\log(p_k))$
Focal Loss	$L_{\text{Focal}} = -(1-p_k)^\gamma \log(p_k)$
LA Loss	$L_{\text{LA}} = -\log\left(\frac{\exp(z_k - \tau \cdot m_k)}{\sum_{j=1}^{C} \exp(z_j - \tau \cdot m_j)}\right)$
LADE Loss	$L_{LADE} = L_{BS} + \alpha \cdot L_{LADER}$
LDAM Loss	$L_{LDAM} = -\log\left(\frac{\lvert D_k \rvert^{\frac{1}{4}} \exp(z_k)}{\sum_{j=1}^{C} \lvert D_j \rvert^{\frac{1}{4}} \exp(z_j)}\right)$

2) Experimental settings: In this section, we adopt the WideResNet architecture as the target model. The models are trained with various loss functions, as detailed in Table I. Hyperparameters related to each loss function are chosen



(a) BS Loss (b) CB Loss (c) Focal Loss

(d) LA Loss (e) LADE Loss (f) LDAM Loss

Fig. 6: **Attack performance and test accuracy in loss function-based long-tailed learning on CIFAR10-IF500.** While accuracy for tail classes has increased, their TPR shows only a minimal decrease.

according to the guidelines provided in their respective original papers and codes. All other experimental settings, such as data preprocessing and optimization strategies, remain consistent across all configurations. The setup for membership inference attacks follows the procedure outlined in Section IV-A.

3) Results: Figure 6 displays attack performance and test accuracy for loss function-based long-tailed learning on CIFAR10-IF500. Six different loss functions are evaluated, with the baseline model, trained using cross-entropy loss, serving as a benchmark. Overall, while loss functions designed for long-tailed datasets (such as CB, LADE, and LDAM) improve classification accuracy, particularly for tail classes, they do not consistently reduce the associated privacy risks. To illustrate, consider the BS loss function shown in Figure 6a. For tail class 9, test accuracy improves substantially by 40%, from approximately 3% to 43% compared to the baseline. However, TPR@0.1% FPR remains at baseline levels, nearing 100%, indicating the privacy risks for tail classes remain unmitigated. In fact, the amplification effect — where privacy risks intensify for tail classes — persists across all evaluated loss functions. Additional results for loss function-based long-tailed learning on CIFAR100-IF50 are included in Supplementary Material (Figure A.4).

Takeaway. In summary, we seek to answer question Q2 through analyzing the classic loss function-based long-tailed learning. We show that similar to foundation model-based long-tailed learning, loss function-based long-tailed learning improves accuracy for tail classes, demonstrating a beneficial effect. However, the privacy risks of tail classes remain high in these loss functions designed for long-tailed learning remain high, comparable to those observed with the standard cross-entropy loss function.

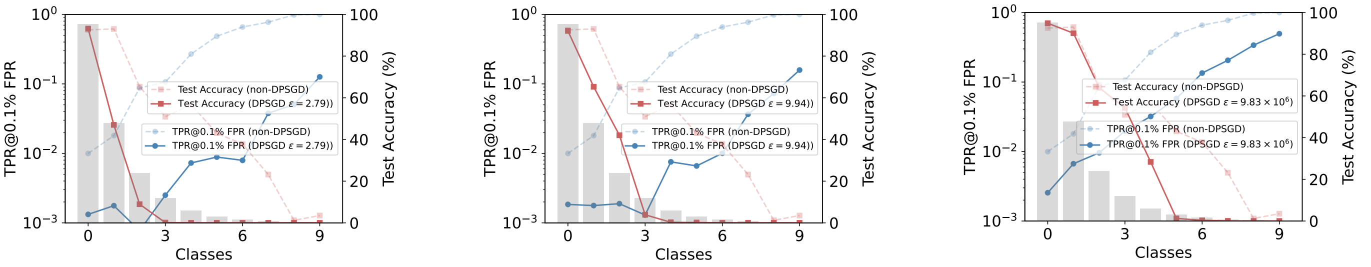
VI. Membership Inference in Long-tailed Data with Differential Privacy

In this section, we focus on examining the privacy risks of ML models trained with differential privacy in long-tailed scenarios, which aims to address question Q3. We start with investigating whether the privacy vulnerabilities identified in previous sections persist under differential privacy. Furthermore, we present a privacy analysis of an innovative approach that combines differential privacy with foundation model-based long-tailed learning.

A. Models Trained from Scratch using DPSGD

1) Introduction: Differential privacy [16] is a primary defense mechanism against privacy attacks, providing strong theoretical guarantees for data protection. In this work, we focus on the differentially private stochastic gradient descent (DPSGD) [1], as it is widely adopted for training ML models in both academic research and industrial applications. While several heuristic defenses have been proposed to mitigate membership inference attacks [8], [10], [23], [34], recent work [2] demonstrates that DPSGD remains the most robust and effective defense. Moreover, unlike heuristic methods, DPSGD offers formal privacy guarantees under the differential privacy framework, enabling rigorous theoretical analysis of privacy preservation.

2) Experimental settings: We implement DPSGD using the Opacus library [53], a widely used framework for training ML models with differential privacy. We adopt the augmentation multiplicity technique, which has demonstrated improvements in the utility-privacy trade-off in recent works [13], [42]. We utilize the WideResNet architecture and train three versions of models using DPSGD, each with different privacy budgets ϵ: $\epsilon \approx 3$, $\epsilon \approx 10$, and $\epsilon \approx 10^7$. The first two budgets,



(a) $\epsilon \approx 3$

(b) $\epsilon \approx 10$

(c) $\epsilon \approx 10^7$

Fig. 7: **Attack performance and test accuracy across different privacy budgets on CIFAR10-IF500.** Results for CIFAR100-IF50 are presented in the Supplement (Figure A.5). TPR@0.1% FPR and test accuracy are averaged over 10 target models.

$\epsilon \approx 3$ and $\epsilon \approx 10$, represent typical privacy-preserving settings commonly used in prior works [1], [42]. In contrast, the extreme budget of $\epsilon \approx 10^7$ simulates a scenario with minimal privacy constraints, allowing us to assess privacy risks under high utility conditions. For experiments conducted under standard privacy settings, we apply a noise multiplier of 1.2, while for the high-utility setting, the noise multiplier is significantly reduced to 0.00625. Across all configurations, the privacy parameter δ is fixed at 10^{-5}. The gradient clipping norm and augmentation factor are set to 1.0 and 8, respectively. We use cross-entropy loss and optimize the models with SGD, with a learning rate of 4. Training is terminated once the target privacy budget is reached.

3) Results: Figure 7 illustrates attack performance and test accuracy across three different privacy budgets using DPSGD on CIFAR10-IF500. In each scenario, the baseline is the target model trained without DPSGD. Overall, we observe the following key points:

(a) As expected, compared to the baseline, TPR@0.1% FPR for models trained with DPSGD decreases significantly across all classes, which indicates that DPSGD can effectively mitigate privacy risks. However, this also results in a marked decline in overall test accuracy, dropping to approximately 0% in most classes and leading to considerable utility loss.

(b) Despite using DPSGD, the privacy risk amplification effect in tail classes remains pronounced. For example, in the model with $\epsilon \approx 3$ (Figure 7a), TPR@0.1% FPR increases by about 100 times, from 0.13% in head class 0 to 12.62% in tail class 9. This demonstrates that adversaries targeting head classes achieve only random guessing performance, while their ability to infer membership in tail classes is amplified approximately 120 times relative to random guessing. This highlights ongoing privacy risks in tail classes.

(c) Test accuracy decreases across almost all classes, as DPSGD inherently causes utility loss. However, test accuracy for the head class (class 0) remains consistently high, exceeding 90%, with no substantial decline. In contrast, the remaining classes experience significant drops. For example, accuracy in tail class 9 falls to 0%, indicating a failure in classification.

(d) Both tail and head classes show a notable decrease in TPR@0.1% FPR. For example, when $\epsilon \approx 3$, TPR@0.1% FPR for head class 0 decreases by about 8 times, from 1.00% to approximately 0.13%, approaching random guessing levels. For tail class 9, TPR decreases by about 8 times as well, from 100% to 12.62%.

(e) As the privacy budget increases (from Figure 7a to Figure 7c), head class accuracy improves noticeably. However, this increase in utility comes at the cost of heightened privacy risks in tail classes. For instance, when ϵ increases to 10^7, head class accuracy (e.g., classes 0, 1, 2, 3) approaches that of the non-private model, but tail class privacy risks (e.g., classes 6, 7, 8, 9) escalate, as indicated by significantly increased TPR values.

Takeaway. In summary, we approach question Q3 by means of studying models trained with DPSGD. We demonstrate that while DPSGD effectively reduces privacy risks, the amplification effect persists, with tail classes remaining more vulnerable to membership inference attacks even under DPSGD. In addition, as the privacy budget increases, head classes experience gains in both privacy and utility, while tail classes face increased privacy risks and utility loss. This highlights a fundamental challenge within the DPSGD mechanism — achieving balanced privacy protection across all data classes in long-tailed datasets.

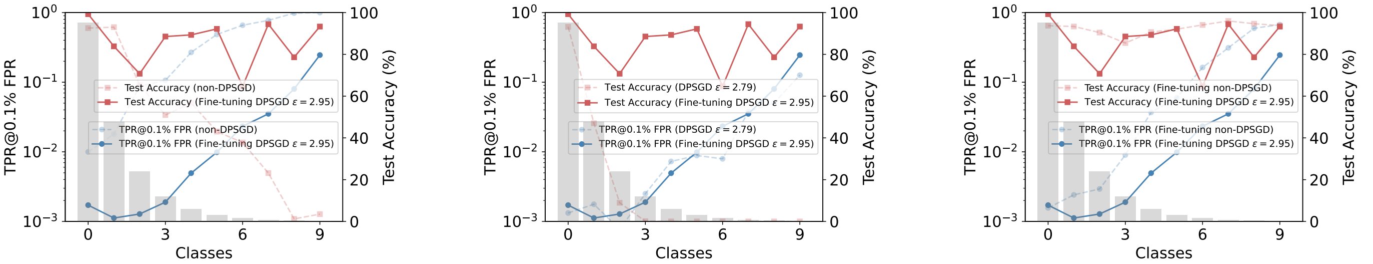
B. *Fine-tuning Models using DPSGD*

1) Introduction: In Section V-A, we show that test accuracy for all classes improves via foundation model-based long-tailed learning. However, privacy risks for tail classes remain rather high. Similarly, in Section VI-A, we illustrate that while DPSGD effectively mitigates membership inference attacks across all classes, it results in low accuracy for most classes. In this section, we present the first exploration of combining foundation model-based long-tailed learning with DPSGD to address both utility performance and privacy concerns.

By integrating these two complementary methods, we seek to determine whether a model can achieve comparable levels of privacy and utility across both head and tail classes in long-tailed scenarios. This also provides insights into how DPSGD interacts with foundation model-based long-tailed learning to mitigate privacy risks across long-tailed datasets.

2) Experimental settings: We use the ViT-B/16 as the foundation model backbone and apply the classifier fine-tuning method. DPSGD is implemented using the Opacus library, with the privacy budget ϵ set to 2.95. We employ the LA loss function with a learning rate of 0.005, a noise multiplier of 1.2, and an augmentation factor of 1. All other DPSGD and

This article has been accepted for publication in IEEE Transactions on Information Forensics and Security. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2025.3607261



(a) Fine-tuning DPSGD vs. Non-DPSGD

(b) Fine-tuning DPSGD vs. DPSGD

(c) Fine-tuning, DPSGD vs. non-DPSGD

Fig. 8: **Attack performance and test accuracy on foundation model-based long-tailed learning with DPSGD on CIFAR10-IF500.** Results for CIFAR100-IF50 are presented in the Supplementary Material (Figure A.6).

foundation model hyperparameters follow the specifications in Section VI-A and Section V-A.

3) Results: Figure 8 presents a comparative analysis of attack performance and test accuracy for three methods: foundation model-based long-tailed learning with DPSGD (fine-tuning DPSGD), foundation model-based long-tailed learning without DPSGD (fine-tuning non-DPSGD) and the WideResNet model with DPSGD on CIFAR10-IF500. Overall, the experimental results demonstrate that integrating foundation model-based long-tailed learning with DPSGD significantly improves both utility and privacy preservation.

(a) As shown in Figure 8a, the fine-tuning DPSGD model not only significantly achieves higher accuracy but also substantially reduces privacy risks, far outperforming the model trained from scratch without DPSGD. Accuracy for head class 0 reaches approximately 99%, with TPR@0.1% FPR dropping to around 0.17%. In contrast, for tail class 9, accuracy improves by 90%, from roughly 3% to 93%, with TPR@0.1% FPR decreasing by 75%, from 100% to approximately 25%.

(b) As depicted in Figure 8b, with a privacy budget of approximately 3, the fine-tuning DPSGD model significantly enhances utility across all classes compared to the model trained from scratch with DPSGD. Both models exhibit similar TPR values at 0.1% FPR, this is somehow expected, given their shared DPSGD mechanism and similar privacy budgets. Notably, the fine-tuning DPSGD model shows a substantial improvement in overall accuracy, especially in classes 2 through 9. This demonstrates the dual advantages of integrating the two complementary methods, enhancing accuracy while maintaining robust privacy. While accuracy variations exist across different classes in the fine-tuning DPSGD model, these are attributed to the inherent gradient clipping and noise addition in DPSGD. We plan to explore more advanced private training strategies to mitigate this issue in future work.

(c) As described in Figure 8c, where both models utilize foundation model-based long-tailed learning with the same classier fine-tuning, TPR@0.1% FPR in fine-tuning DPSGD decreases significantly across all classes, with only a modest reduction in accuracy. This outcome reflects the characteristic trade-off in DPSGD, highlighting the balance between privacy and utility, a principle analogous to the No Free Lunch Theorem in optimization [52].

4) Overhead analysis: To evaluate the computational overhead, we compare the training time of a standard DPSGD model trained from scratch with that of a fine-tuned DPSGD model. The standard DPSGD model adopts the WideResNet architecture. As it is trained from scratch, the number of trainable parameters is equal to the total number of parameters in the network, which amounts to 2.75 million. The fine-tuned DPSGD model leverages the ViT-B/16 as its backbone. A small subset of 7.68 thousand parameters is fine-tuned in the fine-tuned DPSGD model, while it contains a total of 149.63 million parameters. Both models are trained with a target privacy budget of approximately $\epsilon = 3$. The experiments are conducted on a workstation running Ubuntu 22.04.3 LTS, equipped with a single NVIDIA A100 80GB PCIe GPU.

We report the average training time over five independent runs for both configurations. The standard DPSGD model takes 83.93 seconds on average, whereas the fine-tuned DPSGD model reduces this overhead, taking 40.05 seconds. Recall that under this privacy budget, the accuracy of the fine-tuned DPSGD model across all classes, particularly the tail classes, is significantly improved, as shown in Figure 8b. This highlights that the fine-tuning DPSGD model offers both computational efficiency and improved model utility in privacy-preserving scenarios.

Takeaway. In summary, we answer question Q3 by diving into the privacy analysis of fine-tuning models using DPSGD. We show that foundation model-based long-tailed learning with DPSGD preserves privacy while improving accuracy across all classes. However, privacy risks remain higher for tail classes compared to head classes, motivating future research into specialized private training strategies for long-tailed data or the development of a new privacy preservation mechanism.

VII. Discussion

In this section, we analyze the effect of shadow dataset distribution on membership inference attacks and investigate why samples from tail classes are more vulnerable. We further present experimental results on text classification tasks, examine the effects of confidence intervals on attack performance, and discuss potential mitigation strategies.

A. Effect of Shadow Dataset Distribution

In this section, we investigate how the distribution of shadow datasets affects membership leakage risks across different classes in the long-tailed scenarios. While most prior

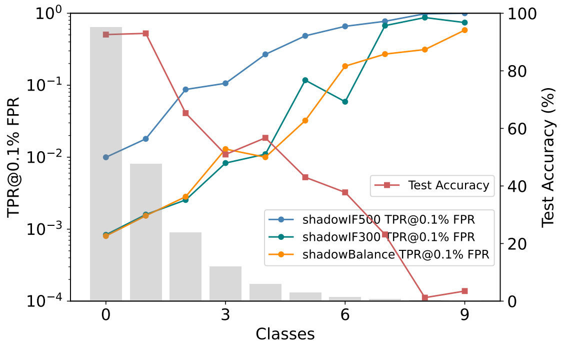


Fig. 9: **Effect of shadow dataset distribution.** The attack performance shows a decrease with increasing distributional discrepancies between the shadow and target datasets. Head classes exhibit a greater decrease in membership inference risk compared to tail classes.

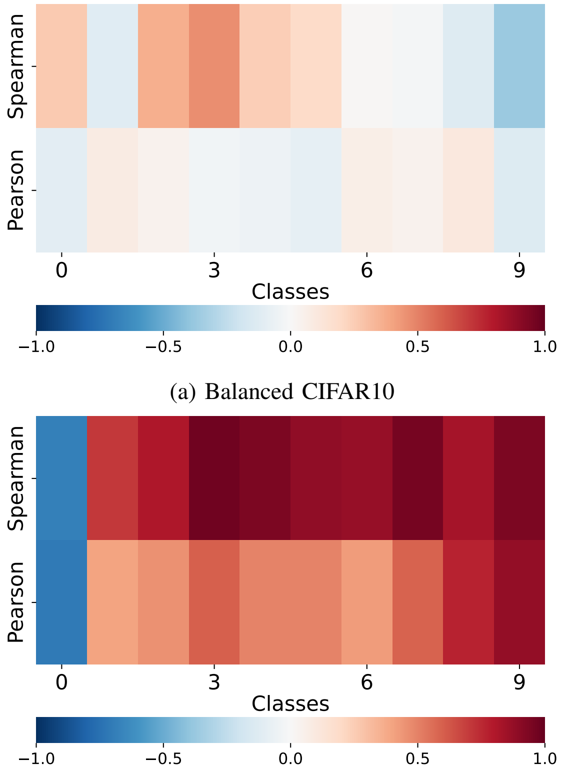
works [5], [41], [44] assume that the shadow datasets share the same distribution as the target dataset, we also examine how distributional discrepancies between shadow and target datasets influence membership leakage in long-tailed scenarios, particularly from head classes to tail classes.

We choose the model trained on the CIFAR10-IF500 dataset as the target model. To assess the effect of distribution mismatch in shadow datasets, we construct three distinct shadow datasets with varying class distributions: 1) shadowIF500: this dataset has the same distribution of the target model (imbalance factor of 500) and serves as the baseline. 2) shadowIF300: this dataset adopts a long-tailed distribution with a different imbalance factor of 300, representing a moderate distributional shift. 3) shadowBalance: this dataset follows a balanced distribution (imbalance factor of 1), where each class contains an equal number of samples. This configuration simulates a scenario where adversaries lack prior knowledge of the target dataset's distribution and default to using a balanced shadow dataset.

Results. Figure 9 illustrates the effect of shadow dataset distribution on membership inference performance. We observe that the TPRs for head classes decline substantially as the distribution discrepancies increase. In contrast, the TPRs for tail classes exhibit only marginal decreases. This shows that distribution mismatches in shadow datasets mainly reduce attack effectiveness on head classes, while tail classes remain consistently vulnerable. Overall, tail class samples exhibit higher membership leakage compared to head class samples, even when the shadow datasets have distributional discrepancies from the target dataset. Additional results under different FPRs are provided in the Supplementary Material (Figure A.7).

B. Why Are Samples from Tail Classes More Vulnerable?

In this section, we investigate why samples from tail classes exhibit higher membership inference risks compared to those from head classes in long-tailed scenarios. Prior work [50] shows that high importance data samples in a dataset are more vulnerable to membership inference attacks. Motivated by this, we analyze the increased membership inference risks in tail classes through the lens of data importance.



(a) Balanced CIFAR10

(b) Long-tailed CIFAR10-IF500

Fig. 10: **Correlation between Shapley values and membership scores across classes on CIFAR10.** In the balanced dataset, samples do not exhibit strong positive correlations across classes. In contrast, in the long-tailed dataset, samples from tail classes show stronger positive correlation than those from head classes. Pearson and Spearman denote the Pearson and Spearman correlation coefficients, respectively.

We quantify data importance using Shapley values, where a higher value indicates a greater contribution to the model's performance. Specifically, we adopt the efficient and scalable KNN-Shapley method [25] to compute the Shapley values. To quantify membership inference vulnerability, we adopt Equation 2 to compute membership scores, which serve as a proxy metric derived from membership inference attacks. We then assess the correlation between the Shapley values and membership scores using Pearson and Spearman correlation coefficients. A strong positive correlation provides evidence that data importance is a key explanatory factor for membership inference vulnerability.

Results. Figure 10 illustrates the correlation between Shapley values and membership scores across head and tail classes in CIFAR10. We observe that samples from tail classes in the long-tailed CIFAR10-IF500 dataset exhibit stronger positive correlations between membership scores and Shapley values compared to those in the balanced CIFAR10 dataset. This indicates that in tail classes, samples with higher Shapley values, i.e., those more influential to the model, are also more susceptible to membership inference attacks.

In other words, samples from tail classes are more likely to be inferred as members because they are more important to the model. Their higher Shapley values reflect their greater

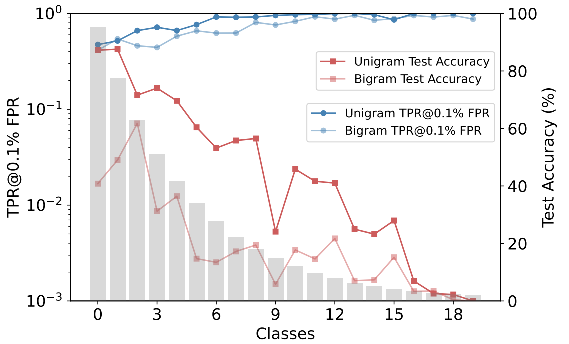
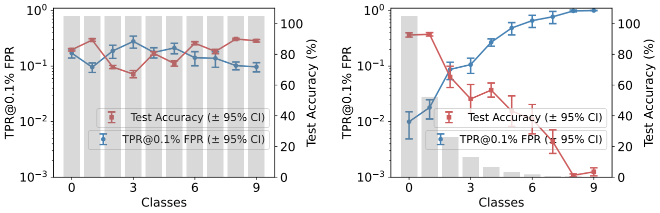


Fig. 11: **Results on text classification.**



(a) Balanced CIFAR10

(b) CIFAR10-IF500

Fig. 12: **Effects of confidence intervals on CIFAR10.** CI refers to the confidence interval.

influence on the model's decisions. This aligns with prior work [50], which shows that samples with higher Shapley values are more vulnerable to membership inference attacks. In addition, the scarcity of samples in tail classes naturally leads to greater memorization. This is because the model has few examples from which to learn under tail classes. To achieve near-optimal generalization, it is compelled to rely heavily on each individual sample, leading to a degree of memorization or overfitting on these specific data points. This is consistent with theoretical insights from [18], which demonstrate that memorization of certain examples is often necessary for achieving almost optimal generalization performance.

Overall, class skewness in the long-tail scenarios acts as a membership inference vulnerability amplifier. By forcing the model to treat scarce tail-class data as highly influential points, it simultaneously increases their importance to the model's predictive accuracy and their susceptibility to membership leakage. Additional results on CIFAR100 are provided in the Supplementary Material (Figure A.8).

C. Results on Text Classification

In addition to image classification, we extend our study to text classification to demonstrate the generalizability of our findings in natural language tasks.

Specifically, we utilize the 20 Newsgroups dataset [26], which comprises approximately 20,000 documents across 20 categories. For our experiments, we create a long-tailed version with an imbalance factor of 50, where the largest class contains 600 samples. The corresponding shadow datasets follow the same class distribution. For text preprocessing, we convert the documents into numerical feature vectors using Term Frequency-Inverse Document Frequency (TF-IDF) representations. To capture different levels of textual granularity, we experiment with two n-gram configurations: 1) Unigrams, which consider individual words, and 2) Bigrams, which capture consecutive two-word phrases. The resulting TF-IDF vectors are then input into a four-layer fully connected neural network with an architecture of 256-128-64-20. ReLU activation functions and Batch Normalization are applied after each hidden layer. Dropout regularization is used with rates of 0.4 after the first two layers and 0.3 after the third.

Results. Figure 11 presents the membership inference attack performance on the 20 Newsgroups dataset. Similar to image classification, tail class samples exhibit significantly higher TPR values compared to head classes, indicating greater membership leakage in the tail regions. We also investigate the effect of different n-gram configurations. While changing from unigrams to bigrams affects the model's test accuracy, the TPR@0.1% FPR remains high and continues to increase from head to tail classes. This indicates that, regardless of feature granularity, the long-tail distribution inherently amplifies membership inference risks in text classification tasks.

D. Effects of Confidence Intervals

Figure 12 presents the TPR@0.1%FPR and test accuracy with 95% confidence intervals on both balanced CIFAR10 and long-tailed CIFAR10-IF500 datasets. All results are averaged over 10 runs to ensure statistical robustness. We observe that both TPR@0.1%FPR and test accuracy exhibit narrow confidence intervals, indicating stable membership inference performance and classification accuracy across classes.

In the balanced CIFAR10, the confidence intervals for TPRs remain consistently tight with minimal variance among classes. In contrast, for CIFAR10-IF500, tail classes exhibit much smaller confidence intervals than head classes, indicating that membership leakage on tail class samples is not only higher but also more stable. Additional results on CIFAR100 are provided in the Supplementary Material (Figure A.9).

E. Potential Mitigation

The exacerbated privacy risks in long-tailed scenarios highlight the urgent need for mitigation. In this section, we discuss several potential strategies.

Privacy equalization via dataset rebalancing. A direct and effective mitigation strategy is to rebalance the training dataset to alleviate class skewness. Our experiments (See Figure 1) show that the TPR values of membership inference attacks on a balanced dataset remain uniform across classes, in stark contrast to the heightened risks observed in long-tailed settings. This means that mitigating class skewness at the data level can reduce class-specific privacy vulnerabilities.

Privacy diffusion via class taxonomy expansion. Another mitigation strategy involves intentionally expanding the class taxonomy to redistribute privacy risks. Given that membership leakage is more pronounced in tail classes, the purposeful introduction of new auxiliary classes, designed to function as additional tail classes, can diffuse the concentration of privacy vulnerabilities. By expanding the class taxonomy, the model's dependency on existing tail class samples can be reduced, thus lowering per-class vulnerability. This risk redistribution mechanism offers a novel perspective for managing privacy leakage, particularly in application domains where the incorporation of auxiliary semantic classes is feasible.

Privacy reinforcement via differentially private fine-tuning. Leveraging foundation models combined with DPSGD private training presents a promising pathway for mitigating privacy risks. Our experiments in Section VI-B show that this method provides formal privacy guarantees and reduces the membership inference vulnerabilities of tail classes. However, despite these improvements, challenges still remain. For instance, the privacy risks of tail classes persistently exceed those of head classes. This highlights the need for further advancements in privacy-preserving fine-tuning techniques, making this an important direction for future research.

Privacy enhancement via hybrid strategy integration. Given the diverse nature of real-world applications, a one-size-fits-all mitigation strategy may be insufficient. In practice, combining the aforementioned strategies, such as rebalancing dataset distributions, expanding class taxonomies to diffuse risks, and employing differentially private fine-tuning training strategies, can yield a more robust privacy-preserving solution. Tailoring these hybrid methods to specific application scenarios allows for flexible trade-offs between privacy, utility, and resource constraints.

VIII. Conclusion and Future Work

We have made a significant advancement by conducting a comprehensive privacy analysis of long-tailed scenarios through membership inference attacks. (1) Prior studies [18], [19] make a foundational claim that memorization is necessary to achieve close-to-optimal generalization error in standard ML models via influence estimation methods. However, our work leverages black-box membership inference attacks, i.e., access to the logit outputs, to advance the understanding of membership inference vulnerabilities in long-tail data from different perspectives. Three privacy effects related to tail data are revealed for standard ML models trained on long-tailed datasets, showing that tail classes are extremely vulnerable to membership inference attacks compared to head classes. (2) Moving beyond standard ML models, we additionally demonstrate that even when state-of-the-art long-tailed learning techniques are employed — yielding substantial performance improvements for tail classes to match those of head classes — tail classes remain markedly more vulnerable to membership leakage. (3) Furthermore, our work also reveals that, even with DPSGD protection in place, tail classes exhibit over 100 times higher privacy risks than head classes under a privacy budget of 3.

Our work has primarily focused on unveiling privacy risks in long-tailed data via membership inference attacks. Building on our work, it is interesting to investigate other types of privacy attacks, such as model inversion and embedding inversion attacks. Given the observed variability in accuracy across classes in foundation model-based long-tailed learning with DPSGD, developing more stable gradient clipping and noise addition mechanisms within DPSGD presents a valuable avenue. In addition, designing a novel DPSGD training mechanism specifically for tail classes can help balance privacy risks across all classes. Finally, beyond DPSGD, exploring new privacy protection mechanisms to specifically address the amplified privacy risks in tail classes offers an intriguing direction for future research.

References

[1] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2016, pp. 308–318.

[2] M. Aerni, J. Zhang, and F. Tramèr, "Evaluations of machine learning privacy defenses are misleading," in *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2024, p. 1271—1284.

[3] E. Bagdasaryan, O. Poursaeed, and V. Shmatikov, "Differential privacy has disparate impact on model accuracy," in *Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 2019.

[4] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," in *Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 2019.

[5] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramer, "Membership inference attacks from first principles," in *IEEE Symposium on Security and Privacy (SP)*. IEEE, 2022, pp. 1519—1519.

[6] N. Carlini, M. Jagielski, C. Zhang, N. Papernot, A. Terzis, and F. Tramer, "The privacy onion effect: Memorization is relative," in *Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 2022, pp. 13 263–13 276.

[7] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[8] D. Chen, N. Yu, and M. Fritz, "Relaxloss: Defending membership inference attacks without losing utility," in *International Conference on Learning Representations (ICLR)*, 2022.

[9] S. Chen, C. Ge, Z. Tong, J. Wang, Y. Song, J. Wang, and P. Luo, "Adaptformer: Adapting vision transformers for scalable visual recognition," in *Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 2022, pp. 16 664–16 678.

[10] Z. Chen and K. Pattabiraman, "Overconfidence is a dangerous thing: Mitigating membership inference attacks by enforcing less confident prediction," in *Network and Distributed System Security Symposium (NDSS)*. Internet Society, 2023.

[11] C. A. Choquette-Choo, F. Tramer, N. Carlini, and N. Papernot, "Label-only membership inference attacks," in *International Conference on Machine Learning (ICML)*. PMLR, 2021, pp. 1964–1974.

[12] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2019, pp. 9260—9269.

[13] S. De, L. Berrada, J. Hayes, S. L. Smith, and B. Balle, "Unlocking high-accuracy differentially private image classification through scale," *arXiv preprint arXiv:2204.13650*, 2022.

[14] B. Dong, P. Zhou, S. Yan, and W. Zuo, "Lpt: Long-tailed prompt tuning for image classification," *arXiv preprint arXiv:2210.01033*, 2022.

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2021.

[16] C. Dwork, "Differential privacy," in *International colloquium on automata, languages, and programming*. Springer, 2006, pp. 1–12.

[17] A. Estabrooks, T. Jo, and N. Japkowicz, "A multiple resampling method for learning from imbalanced data sets," *Computational Intelligence*, vol. 20, no. 1, pp. 18–36, 2004.

[18] V. Feldman, "Does learning require memorization? A short tale about a long tail," in *ACM SIGACT Symposium on Theory of Computing*. ACM, 2020, p. 954—959.

[19] V. Feldman and C. Zhang, "What neural networks memorize and why: Discovering the long tail via influence estimation," in *Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 2020, pp. 2881–2891.

[20] Y. Hong, S. Han, K. Choi, S. Seo, B. Kim, and B. Chang, "Disentangling label distribution for long-tailed visual recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2021, pp. 6622—-6632.

[21] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," in *International Conference on Machine Learning (ICML)*. PMLR, 2019, pp. 2790–2799.

[22] E. J. Hu, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models," in *International Conference on Learning Representations (ICLR)*, 2022.

[23] J. Jia, A. Salem, M. Backes, Y. Zhang, and N. Z. Gong, "Memguard: Defending against black-box membership inference attacks via adversarial examples," in *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2019, pp. 259–274.

[24] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim, "Visual prompt tuning," in *European Conference on Computer Vision (ECCV)*. Springer, 2022, pp. 709–727.

[25] R. Jia, F. Wu, X. Sun, J. Xu, D. Dao, B. Kailkhura, C. Zhang, B. Li, and D. Song, "Scalability vs. utility: Do we have to sacrifice one for the other in data importance quantification?" in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2021, pp. 8235—8243.

[26] Kaggle, "Newsgroups20," Kaggle, https://www.kaggle.com/datasets/crawford/20-newsgroups.

[27] B. Kang, S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng, and Y. Kalantidis, "Decoupling representation and classifier for long-tailed recognition," in *International Conference on Learning Representations (ICLR)*, 2020.

[28] X. Li, Q. Li, Z. Hu, and X. Hu, "On the privacy effect of data enhancement via the lens of memorization," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 4686—-4699, 2024.

[29] Z. Li, Y. Liu, X. He, N. Yu, M. Backes, and Y. Zhang, "Auditing membership leakages of multi-exit networks," in *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2022, pp. 1917—-1931.

[30] Z. Li and Y. Zhang, "Membership leakage in label-only exposures," in *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2021, pp. 880–895.

[31] X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory undersampling for class-imbalance learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 39, no. 2, pp. 539–550, 2008.

[32] Y. Liu, Z. Zhao, M. Backes, and Y. Zhang, "Membership inference attacks by exploiting loss trajectory," in *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2022, p. 2085—2098.

[33] A. K. Menon, S. Jayasumana, A. S. Rawat, H. Jain, A. Veit, and S. Kumar, "Long-tail learning via logit adjustment," in *International Conference on Learning Representations (ICLR)*, 2021.

[34] M. Nasr, R. Shokri, and A. Houmansadr, "Machine learning with membership privacy using adversarial regularization," in *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2018, pp. 634–646.

[35] M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning," in *IEEE Symposium on Security and Privacy (SP)*. IEEE, 2019, pp. 739–753.

[36] M. Nasr, S. Songi, A. Thakurta, N. Papernot, and N. Carlin, "Adversary instantiation: Lower bounds for differentially private machine learning," in *IEEE Symposium on Security and Privacy (SP)*. IEEE, 2021, pp. 866–882.

[37] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning (ICML)*. PMLR, 2021, pp. 8748–8763.

[38] W. J. Reed, "The pareto, zipf and other power laws," *Economics Letters*, vol. 74, no. 1, pp. 15–19, 2001.

[39] J. Ren, C. Yu, X. Ma, H. Zhao, S. Yi *et al.*, "Balanced meta-softmax for long-tailed visual recognition," in *Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 2020, pp. 4175–4186.

[40] T.-Y. Ross and G. Dollár, "Focal loss for dense object detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 2980–2988.

[41] A. Salem, Y. Zhang, M. Humbert, P. Berrang, M. Fritz, and M. Backes, "Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models," in *Network and Distributed System Security Symposium (NDSS)*. Internet Society, 2019.

[42] T. Sander, P. Stock, and A. Sablayrolles, "Tan without a burn: Scaling laws of dp-sgd," in *International Conference on Machine Learning (ICML)*. PMLR, 2023, pp. 29 937–29 949.

[43] J.-X. Shi, T. Wei, Z. Zhou, J.-J. Shao, X.-Y. Han, and Y.-F. Li, "Long-tail learning with foundation model: Heavy fine-tuning hurts," in *International Conference on Machine Learning (ICML)*. PMLR, 2024, pp. 45 014–45 039.

[44] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *IEEE Symposium on Security and Privacy (SP)*. IEEE, 2017, pp. 3–18.

[45] T. Steinke, M. Nasr, and M. Jagielski, "Privacy auditing with one (1) training run," in *Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 2023, pp. 49 268–49 280.

[46] S. Truex, L. Liu, M. E. Gursoy, W. Wei, and L. Yu, "Effects of differential privacy and data skewness on membership inference vulnerability," in *IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications*. IEEE, 2019, pp. 82–91.

[47] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 2017, p. 6000—6010.

[48] B. Wang, P. Wang, W. Xu, X. Wang, Y. Zhang, K. Wang, and Y. Wang, "Kill two birds with one stone: Rethinking data augmentation for deep long-tailed learning," in *International Conference on Learning Representations (ICLR)*, 2024.

[49] X. Wang, J. Zhao, E. Marostica, W. Yuan, J. Jin, J. Zhang, R. Li, H. Tang, K. Wang, Y. Li *et al.*, "A pathology foundation model for cancer diagnosis and prognosis prediction," *Nature*, pp. 1–9, 2024.

[50] R. Wen, M. Backes, and Y. Zhang, "Understanding data importance in machine learning attacks: Does valuable data pose greater harm?" in *Network and Distributed System Security Symposium (NDSS)*. Internet Society, 2025.

[51] R. Wen, Z. Li, M. Backes, and Y. Zhang, "Membership inference attacks against in-context learning," in *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2024, p. 3481—3495.

[52] D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization," *IEEE Transactions on Evolutionary Computation*, vol. 1, no. 1, pp. 67–82, 1997.

[53] A. Yousefpour, I. Shilov, A. Sablayrolles, D. Testuggine, K. Prasad, M. Malek, J. Nguyen, S. Ghosh, A. Bharadwaj, J. Zhao, G. Cormode, and I. Mironov, "Opacus: User-friendly differential privacy library in PyTorch," *arXiv preprint arXiv:2109.12298*, 2021.

[54] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *Proceedings of the British Machine Vision Conference*, 2016.

[55] E. B. Zaken, Y. Goldberg, and S. Ravfogel, "Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models," in *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022, pp. 1–9.

[56] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations (ICLR)*, 2018.

[57] M. Zhang, Z. Ren, Z. Wang, P. Ren, Z. Chen, P. Hu, and Y. Zhang, "Membership inference attacks against recommender systems," in *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2021, pp. 864–879.

[58] Y. Zhang, B. Kang, B. Hooi, S. Yan, and J. Feng, "Deep long-tailed learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10 795–10 816, 2023.

[59] D. Zhong, X. Wang, Z. Xu, J. Xu, and W. H. Wang, "Interaction-level membership inference attack against recommender systems with long-tailed distribution," in *ACM International Conference on Information and Knowledge Management (CIKM)*. ACM, 2024, p. 3433—3442.

[60] G. K. Zipf, *Human Behavior and the Principle of Least Effort*. Cambridge, MA: Addison-Wesley, 1949.