ProoFit: CONFIDENTIAL AUDITING OF MODEL OVERFITTING

Anonymous authors
Paper under double-blind review

ABSTRACT

As machine learning models are increasingly deployed in high-stakes settings, auditing their quality becomes essential. A critical property of interest is whether the model is overfitted. Yet direct audits are often impossible: model owners cannot reveal model weights without risking exposure of proprietary IP. While zero-knowledge proofs, a cryptographic primitive, offer a compelling theoretical solution, naive approaches require proving the correctness of entire training trajectories–an operation that is prohibitively expensive for modern models. In this paper, we introduce *ProoFit*, a practical system for auditing overfitting without violating model confidentiality. Rather than proving properties of the training process, *ProoFit* employs novel tests that depend only on trained model weights, leveraging theoretical connections between generalization and model complexity. Our experiments demonstrate that *ProoFit* significantly outperforms baseline cryptographic approaches, achieving a 10^3 to $10^7\times$ improvement in runtime.

1 INTRODUCTION

Machine learning (ML) models now regularly make high-stake, consumer-facing decisions: approving loans, diagnosing diseases, hiring, and pricing insurance policies. As the societal impact of these systems has grown, so has the scrutiny they face. Governments worldwide European Data Protection Supervisor (2025); Innovation, Science and Economic Development Canada (2023); Executive Office of the President (2023); of the Auditor General of Norway (2023), are increasingly mandating *audits* of ML models, requiring organizations to attest model quality and demonstrate compliance with regulatory standards. A central challenge in this setting is the tension between *transparency* and *confidentiality*. Effective auditing requires visibility into model behavior, yet both model weights and training data are often proprietary and sensitive. Revealing them exposes organizations to serious risks: private training data can be leaked Thakkar et al. (2021); Henderson et al. (2018); Carlini et al. (2021); Lee et al. (2021), and competitors may steal the model for their own products without incurring the associated R&D costs Oliynyk et al. (2023); Carlini et al. (2024). Full disclosure therefore cannot be assumed in practice.

Within this constrained audit setting, a fundamental property of interest is whether a model is overfitted. Overfitting is a classical but critical failure mode: instead of learning generalizable patterns, an overfitted model memorizes its training data. Such models generalize poorly, achieving high accuracy on training data while failing on new, real-world inputs – precisely where they must succeed. In high-stakes domains such as healthcare or finance, these failures can directly cause harm. For these reasons alone, auditing for overfitting is an important objective in its own right.

Beyond generalization, overfitting also serves as practical proxy for a second auditing concern: privacy leakage. In particular, the auditor wants to determine whether a model reveals information about its training data. The state-of-the-art approach for measuring privacy leakage relies on membership inference (MI) tests, where, given a data point x, the goal is to determine whether it was part of the model's training set DeAlcala et al. (2025a); Mancera et al. (2026); DeAlcala et al. (2025b); Mancera et al. (2025); DeAlcala et al. (2024). MI tests face two significant limitations in our setting. First, conducting an MI test requires knowing the specific data point x to test – but access to individual training points is precisely what our confidentiality constraints prohibit. Second, a growing body of work has shown that MI tests can be unreliable in practice Zhang et al. (2024); Duan et al. (2024a); Das et al. (2025); Kong et al. (2023); Mangaokar et al. (2025); Aubinais et al. (2025), undermining their utility as a direct auditing tool. Overfitting provides a compelling alternative. A large body of work establishes a strong connection between memorization and privacy leakage Duan et al. (2024b); Ramakrishnan & Balaji (2025); Sander et al. (2025); Borkar (2023); Song & Mittal (2021); Li et al. (2024); Revisiting; Carlini et al. (2021); Zhang (2026); Zhang et al. (2025a); El Mestari et al. (2024);

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

Leino & Fredrikson (2020); Carlini et al. (2019); Thakkar et al. (2020); Carlini et al. (2022); Yeom et al. (2017). Thus, certifying that a model does *not* overfit provides indirect but meaningful evidence that it is unlikely to leak training data.*

Designing such an audit, however, is non-trivial. As noted above, the audit must operate *without* revealing either the model or its training data. In principle, this can be resolved using zero-knowledge proofs (ZKPs), which allows the model owner to prove properties of a model without revealing the model itself. In practice, however, ZKPs are notoriously computationally expensive. The challenge is compounded in our setting: standard overfitting audits require repeated evaluations over both training and test datasets-essentially cross-validation. Worse still, a technically sound audit must establish that the model was actually trained on the claimed dataset, which requires reasoning about the *entire* training procedure. For modern models, this entails massive datasets, long optimization trajectories, and hundreds of iterations. Generating ZKPs of such training processes is extremely costly.

This gap between cryptographic feasibility and practical auditing is the core problem we address. To this end, we propose ***ProoFit***†, practical audit for overfitting without violating model confidentiality. Rather than proving properties of the training process itself, *ProoFit* introduces novel tests that depend solely on the trained model's weights. Our key insight is to leverage the theoretical connections between generalization and model complexity. While these theoretical bounds offer valuable intuition, they are too loose to be directly useful, yielding vacuous guarantees. We operationalize these insights into a practical solution by casting them as relative tests: comparisons between the candidate model and a reference model, which allow for meaningful thresholds. We further introduce optimizations that enable efficient ZKPs for these tests. The crucial observation is that we do not need to replicate the entire computation inside the ZKP circuit; we only need to *verify* its correctness. Accordingly, we design ZKP-friendly verification protocols that replace expensive operations (such as computing spectral norms and performing large matrix multiplications) with lightweight randomized checks.

In summary, we make the following contributions:

- We design a new overfitting tests that avoid reasoning about the full training procedure and depend only on the final model. We provide formal guarantees characterizing the correctness and soundness of the test.
- We design and implement *ProoFit*, a specialized ZK proof protocol, to efficiently verify the above test without violating model confidentiality.
- We evaluate *ProoFit* on a suite of datasets, demonstrating its practicality. *ProoFit* audits ResNet-18 on CIFAR-10 in under 3 minutes. In contrast, the baseline approach generating ZK proof of training takes 5.64×10^6 seconds per epoch translating to > 17 years for 100 epochs. This represents $10^7\times$ improvement. Importantly, the baseline is completely impractical, whereas *ProoFit* runs well within realistic deployment constraints.

Remark 1.1. *Here we focus on auditing neural networks. Our work is along the line of a growing line of research in cryptographic auditing Huang et al. (2024); Lycklama et al. (2022); Waiwitlikhit et al. (2024); Lycklama et al. (2024); Franzese (2024); Park et al. (2022). Important point here is that prior work has been confined to simpler settings, such as linear models, decision trees, or very shallow neural networks (none of them address auditing for overfitting). Our results highlight, baseline techniques do not scale even to CNNs, let alone to large models such as LLMs. In contrast, we are the first to demonstrate that confidential auditing is feasible for realistic CNNs, taking a concrete and significant step in this direction. We leave the extension to LLMs as future work.*

2 BACKGROUND

Notations. Consider a k-class classifier $\mathcal{M} : \mathcal{X} \to \mathbb{R}^k$, where $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\|_2 \le B\}$. We model $\mathcal{M}$ as an ℓ-layer ReLU network with weights $\mathcal{W} = W_i{}_{i=1}^{\ell}$: $\mathcal{M}(x) = \varphi(W_\ell \varphi(\cdots \varphi(W_1 x)))$. The predicted label is $\arg\max_{i\in[k]} \mathcal{M}(x)_i$, and the margin of (x, y) is $\mathcal{M}(x)_y - \max_{i\neq y} \mathcal{M}(x)_i$. We use $|\cdot|_\sigma$ for the spectral norm. For $W \in \mathbb{R}^{n\times m}$, define $\|W\|_{p,q} = \left\| \left(\|W_{:,1}\|_p, \ldots, \|W_{:,m}\|_p \right) \right\|_q$.

Model Overfitting. Overfitting occurs when a model performs well on its training data but poorly on unseen data, yielding a large *generalization gap*. We use the following margin-loss definition.

Definition 1 (Margin Loss)**.** *For distribution $\mathcal{D}$ and $\gamma > 0$,*
$L_\gamma(\mathcal{M}) \Pr_{(x,y)\sim\mathcal{D}} [\mathcal{M}(x)[y] \le \gamma + \max_{j\neq y} \mathcal{M}(x)[j]]$.

*While this provides a lower bound on privacy leakage Yeom et al. (2017), such lower-bound guarantees are standard in empirical privacy auditing, including audits of differential privacy González et al. (2025); Ganev et al. (2026); Nasr et al. (2023).

†**Proo**f of (not) over**Fit**

Let $\hat{L}_\gamma(\mathcal{M})$ denote its empirical estimate. For $\gamma = 0$, $L_0(\mathcal{M})$ is the population classification risk and $\hat{L}_{0,\mathcal{D}_{\text{train}}}(\mathcal{M})$ is the training error. Thus, $\textit{GenGap}(\mathcal{M}) = L_0(\mathcal{M}) - \hat{L}_{0,\mathcal{D}_{\text{train}}}(\mathcal{M})$.

2.1 Cryptographic Primitives

ProoFit uses commitment schemes and zero-knowledge proofs to provide confidentiality and verifiability.

Commitment Schemes. A commitment scheme Blum (1981) allows a party to commit to a private value x through com_x. It is *hiding*, revealing nothing about x, and *binding*, preventing the committer from later opening it to $x' \neq x$.

Zero-Knowledge Proofs (ZKPs). A ZKP allows a prover to convince a verifier that it knows x satisfying a public predicate $P(x) = 1$, without revealing anything beyond this fact Goldwasser et al. (1985). A ZKP satisfies *completeness* for honest proofs, *soundness* against false claims, and *zero knowledge* against malicious verifiers. Although any NP predicate can be verified in zero knowledge Goldreich et al. (1991), complex predicates can incur prohibitive overhead. *ProoFit* therefore develops ZKP-friendly tests for efficiently auditing overfitting.

3 Problem Setting

Problem Statement. We consider two parties: a model owner and a third-party auditor. The owner trains and deploys a model $\mathcal{M}$ (e.g., ML-as-a-Service), while the auditor verifies regulatory compliance (e.g., government agencies or firms such as Deloitte, EY, or PwC). The owner (prover) must convince the auditor (verifier) that $\mathcal{M}$ is not overfitted while keeping the model weights and training data confidential. The auditor specifies a maximum acceptable generalization gap τ and trains a reference model $\mathcal{M}_{\mathsf{ref}}$ on data drawn from the same distribution $\mathcal{X}$, calibrated such that $\textit{GenGap}(\mathcal{M}_{\mathsf{ref}}) \approx \tau$, e.g., by selecting a checkpoint near the onset of overfitting Bartlett et al. (2020); Cao et al. (2022); Li et al. (2021). Since the auditor evaluates $\mathcal{M}_{\mathsf{ref}}$ in the clear, this calibration is straightforward. The audit then reduces to the relative test $\mathcal{M}$ is $\iff \textit{GenGap}(\mathcal{M}) \leq \textit{GenGap}(\mathcal{M}_{\mathsf{ref}})$. Thus, $\mathcal{M}_{\mathsf{ref}}$ defines the auditor-specified boundary for acceptable overfitting, consistent with standard auditing practice (Remark 4.1).

Technical Challenges. A sound audit must ensure that the reported training performance is computed on the *actual* training data; otherwise, a malicious owner could manipulate the dataset or training-test split. A standard approach would compare training and held-out performance, but doing so confidentially requires binding $\mathcal{M}$ to its training data. Cryptographically, this amounts to a *proof-of-training*: a ZK proof that $\mathcal{M}$ was produced by the specified training algorithm on the claimed dataset. Unfortunately, proving training is prohibitively expensive. ZK computation must be represented as an arithmetic circuit, and training combines costly non-linear operations with hundreds of sequential iterations over large datasets. The resulting circuits make proving orders of magnitude more expensive than training itself. For example, proving one epoch of ResNet18 on CIFAR10 takes 5.64×10^6 seconds, corresponding to > 17 years for 100 epochs (Table 1).

Solution Desiderata. We therefore design our audit around three properties:

- *Training-Process Independence.* The audit should depend only on the final model parameters, avoiding verification of iterative training.
- *Parallelization-Friendly.* ZK proof generation should be efficiently parallelizable across processors.
- *Theoretical Soundness.* The audit should rely on established learning-theoretic guarantees so that passing provides meaningful evidence against overfitting.

4 *ProoFit*: System Description

ProoFit operates in three stages:

- **1. Model Commitment.** The model owner commits to $\mathcal{M}$ using a binding and hiding commitment, preventing subsequent changes while revealing nothing about the model.
- **2. Certificate Generation.** The owner generates an audit certificate $\mathscr{C} = \langle \rho, \pi_\rho \rangle$, where ρ is a computation over $\mathcal{M}$ and π_ρ is a ZKP proving that ρ was correctly computed using the committed model without revealing $\mathcal{M}$.
- **3. Auditing.** The auditor verifies π_ρ and evaluates $f(\rho, \mathcal{M}_{\mathsf{ref}}) \in 0, 1$, comparing ρ against the reference model as described in Sec. 3. The output determines whether $\mathcal{M}$ passes the audit.

Stage 1 can be instantiated using standard commitment schemes Kate et al. (2010); Boneh et al. (2021), so the primary technical contributions of *ProoFit* lie in Stages 2 and 3. First, we describe our

audit test, namely, ρ, the decision function f, and the auditor parameters γ. Next, we detail how to efficiently compute the ZKP π_ρ.

4.1 Efficient Test of Model Overfitting

4.1.1 Technical Overview

Our construction is built on two key ideas.

Key Idea 1: Overfitting is Connected to Model Complexity.

Our first idea leverages a central insight from statistical learning theory: a model overfits when its effective complexity is too high relative to the data. Specifically, we draw upon a rich body of work on formalizing generalization bounds in terms of the model's complexity Bartlett et al. (2017); Neyshabur et al. (2017b); Langford & Shawe-Taylor (2002); Dziugaite & Roy (2017). At a high level, such bounds take the form of a sum of two terms: an empirical estimate of performance and a penalty term that grows with model complexity. For our setting, the complexity measure used in the bound must satisfy two properties: (1) *ZKP-friendly*: It must be efficiently computable inside a ZKP. (2) *Scale-invariant*: Neural networks admit trivial rescalings of the weights that leave the underlying function unchanged. A non-scale-invariant measure would assign different complexities to functionally identical models, rendering it meaningless for auditing.

Unfortunately, classical measures such as the Vapnik-Chervonenkis (VC) dimension Vapnik & Chervonenkis (1971) or Rademacher complexity Bartlett & Mendelson (2002) do not satisfy these requirements. To address this challenge, we adopt the spectrally normalized, margin-based framework introduced by Bartlett et al. Bartlett et al. (2017) and extended by Neyshabur et al. Neyshabur et al. (2017b). In this framework, empirical performance is captured using the margin loss (Def. 1), while model capacity is quantified via a spectral-norm-based complexity measure, $R_{\mathcal{W}}$. For a model with weight matrices $\mathcal{W} = \{W_1, \cdots, W_l\}$, its spectral complexity measure is defined as: $R_{\mathcal{W}} = \left(\prod_{i=1}^{L} \|W_i\|_\sigma\right) \left(\sum_{i=1}^{L} \frac{\|W_i^T\|_{2,1}^{2/3}}{\|W_i\|_\sigma^{2/3}}\right)^{3/2}$. The first term-the product of spectral norms-is the Lipschitz constant of the model, while the second term is a structural correction factor. Using this complexity measure, Bartlett et al. prove the following generalization guarantee.

Theorem 1. *Bartlett et al. (2017) For a dataset $\{(x_i, y_i)\}_{i=1}^n$ drawn iid from any probability distribution over $\mathbb{R}^d \times \{1, \ldots, k\}$, with probability at least $1 - \delta$ over $\{(x_i, y_i)\}_{i=1}^n$, every margin $\gamma > 0$ and model $\mathcal{M} : \mathbb{R}^d \to \mathbb{R}^k$ with weight matrices $\mathcal{W} = (W_1, \ldots, W_l)$ satisfy $L_0(\mathcal{M}) \leq \hat{L}_\gamma(\mathcal{M}) + \tilde{O}\left(\frac{B \cdot R_{\mathcal{W}}}{\gamma} \ln(w) + \sqrt{\frac{\ln(1/\delta)}{n}}\right)$ where $R_{\mathcal{W}}$ is the spectral complexity measure given by Eq. 4.1.1, w is the maximum width of the model layers, and each input satisfies $||x_i||_2 < B$.*

Thus, the above result provides a generalization bound directly in terms of the model weights, thereby providing the backbone for our test for overfitting that is independent of the training process.

Key Idea 2: Relative Test. Unfortunately, the above result cannot be used directly in practice. While it provides rigorous theoretical insights, the bound is loose and vacuous (greater than 1) for modern neural networks. Obtaining non-vacuous bounds requires strong structural assumptions about the model that rarely hold in practice Neyshabur et al. (2017b).

To address this challenge, we shift from using raw values of the bound to a *relative test*. Recall that the spectral complexity measure serves as a reliable *proxy* for a model's degree of overfitting: a higher spectral complexity typically corresponds to a larger generalization gap and thus a more overfitted model. However, the raw magnitude of this quantity is difficult to interpret in isolation – there is no natural threshold separating "acceptable" from "overfitted". In other words, what is missing is a reference point. We introduce such a reference point in two steps.

- **Reference model $\mathcal{M}_{\text{ref}}$.** The auditor uses $\mathcal{M}_{\text{ref}}$ to represent the maximum acceptable level of overfitting, as defined in Sec. 3.
- **Normalized complexity statistic.** We consider $\frac{\gamma}{R_{\mathcal{W}} B}$ to be our desired complexity measure. Two points are important. First, this is essentially the inverse of the complexity term in the theoretical bound in Eq. 1, which means now higher values indicate better generalization. We adopt this inversion because raw margins can be zero, making the direct complexity term numerically unstable to be estimated empirically (as discussed in the following section). Second, $\frac{\gamma}{B}$ is actually a normalization factor which makes the statistics *scale-invariant*. Thus, it can be compared meaningfully across different models with different parameters.

Our relative test compares the above statistic of the candidate model with that of the reference model. If the candidate's statistic is significantly lower than the reference model's, it is more complex (and more overfitted) than what the auditor allows. If it is at least as high, the model meets the acceptance criterion.

4.2 PUTTING IT ALL TOGETHER

We now describe the complete audit. The auditor samples an audit dataset $D_{\mathsf{audit}} = \{(x_i, y_i)\}_{i=1}^m$ from the public data distribution and evaluates its calibrated reference model $\mathcal{M}_{\mathsf{ref}}$ on the same examples. The model owner evaluates the committed candidate model $\mathcal{M}$ on each $(x_i, y_i) \in D_{\mathsf{audit}}$ and computes:

- the prediction margin $p_i = \mathcal{M}(x_i)_{y_i} - \max_{j \neq y_i} \mathcal{M}(x_i)_j$,
- the model's spectral complexity $R_{\mathcal{W}}$.

The owner returns $\rho = (\{p_i\}_{i=1}^m, R_{\mathcal{W}})$ along with a ZKP π_ρ certifying that these quantities were computed correctly from the committed model (elaborated in Sec. 2.1). After verifying π_ρ, the auditor forms the normalized margins $p_i^N = \frac{p_i}{B R_{\mathcal{W}}}$, where B is the public bound on the input norm. Let $\hat{\mu}_{\mathsf{cndt}}$ and $\hat{\sigma}_{\mathsf{cndt}}$ denote the empirical mean and standard deviation of the candidate's normalized margins, and define $\hat{\mu}_{\mathsf{ref}}$ and $\hat{\sigma}_{\mathsf{ref}}$ analogously for the reference model.

The audit test requires the candidate to satisfy two complementary conditions: **Condition 1:** $\hat{\mu}_{\mathsf{cndt}} > \hat{\mu}_{\mathsf{ref}}$, and **Condition 2:** $\frac{\hat{\mu}_{\mathsf{cndt}}}{\hat{\sigma}_{\mathsf{cndt}}} > \frac{\hat{\mu}_{\mathsf{ref}}}{\hat{\sigma}_{\mathsf{ref}}}$.

Condition 1 captures the central idea of our test: the link between a model's spectral complexity and its generalization gap. The mean margin alone, however, does not describe how the margin distribution behaves near the decision boundary. A candidate model may have a larger normalized mean but a much larger variance, placing more probability mass near or below zero. Because a non-positive margin corresponds to a misclassification, such a model need not have lower population error. Condition 2 accounts for this by comparing $\hat{\mu}/\hat{\sigma}$, which measures the location of the margin distribution relative to its spread. Under the distributional assumptions of our formal analysis, a larger $\hat{\mu}/\hat{\sigma}$ implies less probability mass below zero Boucheron et al. (2013); Casella & Berger (2002). We next describe how these population-level conditions are certified statistically from empirical quantities computed on D_{audit}.

Condition 1. We statistically validate Condition 1 using a one-sided paired permutatation test Phipson & Smyth (2016); Ernst (2004); Efron & Tibshirani (1994); Fisher & Fisher (1966); Anderson & Robinson (2001); Nyblom (2015).

Lemma 1 (Paired permutation test)**.** *Given an audit dataset D_{audit} of size m, let $d_i = p_i^{N,\mathsf{cndt}} - p_i^{N,\mathsf{ref}}$, $\bar{d}_{\mathsf{obs}} = \frac{1}{m}\sum_{i=1}^m d_i = \hat{\mu}_{\mathsf{cndt}} - \hat{\mu}_{\mathsf{ref}}$. We test $H_0 : \hat{\mu}_{\mathsf{cndt}} = \hat{\mu}_{\mathsf{ref}}$ vs. $H_1 : \hat{\mu}_{\mathsf{cndt}} > \hat{\mu}_{\mathsf{ref}}$, where under H_0 the paired margins are exchangeable and d_i is symmetric around zero. For $K \geq 10^5$, sample $S^{(j)} \xleftarrow{\$} \{-1, 1\}^m$ and compute $\bar{d}_{S^{(j)}} = \frac{1}{m}\sum_{i=1}^m S_i^{(j)} |d_i|$. The empirical p-value is $\hat{p} = \frac{1}{K}\sum_{j=1}^K \mathbb{I}\big(\bar{d}_{S^{(j)}} \leq \bar{d}_{\mathsf{obs}}\big)$. At significance level α, if $\hat{p} \leq \alpha$, we reject H_0 and conclude with confidence $1 - \alpha$ that $\hat{\mu}_{\mathsf{cndt}} > \hat{\mu}_{\mathsf{ref}}$, identifying the candidate as well-fitted.*

The proof is given in App. A.1. The paired construction is natural because both models are evaluated on the same audit examples, so example difficulty induces correlated margins. Working with per-example differences preserves this dependence and removes variation shared by the two models. Under the symmetry assumption on the paired differences, the test provides finite-sample control of the Type-I error Phipson & Smyth (2016); Ernst (2004); Fisher & Fisher (1966); Anderson & Robinson (2001).

Condition 2. We statistically validate condition 2 using a paired non-parametric bootstrap test Efron et al. (2000).

Lemma 2. *Define $\hat{\theta}_{\mathsf{cndt}} = \frac{\hat{\mu}_{\mathsf{cndt}}}{\hat{\sigma}_{\mathsf{cndt}}}, \hat{\theta}_{\mathsf{ref}} = \frac{\hat{\mu}_{\mathsf{ref}}}{\hat{\sigma}_{\mathsf{ref}}}, \hat{\Delta}_{\mathsf{obs}} = \hat{\theta}_{\mathsf{cndt}} - \hat{\theta}_{\mathsf{ref}}$. We test $H_0 : \frac{\mu_{\mathsf{ref}}}{\sigma_{\mathsf{ref}}} \geq \frac{\mu_{\mathsf{cndt}}}{\sigma_{\mathsf{cndt}}}$ vs. $H_1 : \frac{\mu_{\mathsf{ref}}}{\sigma_{\mathsf{ref}}} < \frac{\mu_{\mathsf{cndt}}}{\sigma_{\mathsf{cndt}}}$, where H_1 identifies the candidate as well-fitted. We perform a paired non-parametric bootstrap: for each $j \in [K]$, sample $I^{(j)} \xleftarrow{\$} [m]^m$ with replacement and compute $\hat{\Delta}^{(j)} = \frac{\hat{\mu}_{\mathsf{cndt}}^{(j)}}{\hat{\sigma}_{\mathsf{cndt}}^{(j)}} - \frac{\hat{\mu}_{\mathsf{ref}}^{(j)}}{\hat{\sigma}_{\mathsf{ref}}^{(j)}}$ from the corresponding paired samples. Let $\hat{L}_{1-\alpha}$ be the empirical α-quantile of $\hat{\Delta}^{(j)}$, i.e., the one-sided $(1-\alpha)$ lower confidence bound. For sufficiently large K (e.g., $K \geq 10^5$), if $\hat{L}_{1-\alpha} > 0$, we reject H_0 and conclude with confidence $1 - \alpha$ that $\frac{\mu_{\mathsf{ref}}}{\sigma_{\mathsf{ref}}} < \frac{\mu_{\mathsf{cndt}}}{\sigma_{\mathsf{cndt}}}$, thus identifying the candidate as well-fitted.*

The proof is given in App. A.2. As in Condition 1, resampling the example indices jointly preserves the dependence between candidate and reference margins. This is particularly important because $\hat{\mu}/\hat{\sigma}$ is nonlinear and its sampling distribution can be poorly approximated in the small-sample regime. The paired bootstrap directly estimates the variability of the statistic while preserving the candidate-reference pairing. Unlike the permutation test used for Condition 1, the bootstrap confidence bound is not an exact finite-sample procedure, but is consistent under standard bootstrap regularity conditions Efron et al. (2000).

Final Audit Decision. *ProoFit* accepts the candidate only if both conditions are statistically certified.

In particular, $f(\mathcal{M}_{\mathsf{cndt}}, \mathcal{M}_{\mathsf{ref}}) = 1$ iff $\hat{\mu}_{\mathsf{cndt}} > \hat{\mu}_{\mathsf{ref}}$ and $\frac{\hat{\mu}_{\mathsf{cndt}}}{\hat{\sigma}_{\mathsf{cndt}}} > \frac{\hat{\mu}_{\mathsf{ref}}}{\hat{\sigma}_{\mathsf{ref}}}$, where both inequalities are required to hold with the prescribed statistical confidence.

Remark 4.1. *The use of a reference model aligns with standard auditing practice, where independently trained baselines are routinely used to define acceptable performance thresholds. A prominent example is the champion-challenger framework, widely adopted in financial services and MLOps, where a production model (the "champion") is continuously evaluated against an independently trained alternative (the "challenger") to internal audits Digalakis Jr et al. (2025); Nath (2007); Amed et al. (2025); Eken et al. (2025). Training such a reference model is feasible in practice: auditors are typically third-party entities (government agencies or professional services firms, such as, Deloitte, PwC) for whom auditing is part of their professional service and who therefore possess sufficient computational resources to train a single reference model as part of the audit process. An additional advantage is that our complexity measure is scale-invariant and architecture-agnostic, allowing meaningful comparisons across different model. This means the auditor need only train a single reference model, which can then be reused to audit any number of models trained on the same distribution. This scenario directly reflects real-world settings where the same service is offered by multiple competing companies (e.g., multiple lenders providing consumer credit score). Consequently, the cost of training the reference model is a one-time investment that is amortized across multiple audits, making the overall auditing process both principled and practical.*

4.3 Formal Analysis

Correctness. We show that passing both audit conditions, together with a mild separation assumption, is sufficient to establish the audit goal, i.e., $\mathit{GenGap}(\mathcal{M}_{\mathsf{cndt}}) < \mathit{GenGap}(\mathcal{M}_{\mathsf{ref}})$.

Theorem 2. *Let $p^N_{\mathsf{cndt}} \sim \mathcal{N}(\mu_{\mathsf{cndt}}, \sigma^2_{\mathsf{cndt}})$ and $p^N_{\mathsf{ref}} \sim \mathcal{N}(\mu_{\mathsf{ref}}, \sigma^2_{\mathsf{ref}})$. Suppose Conditions 1 and 2 are certified, with $\mu_{\mathsf{cndt}} - \mu_{\mathsf{ref}} \geq d$, and $\widehat{L}_{0,T_{\mathsf{ref}}}(\mathcal{M}_{\mathsf{ref}}) - \widehat{L}_{0,T_{\mathsf{cndt}}}(\mathcal{M}_{\mathsf{cndt}}) \leq \epsilon$. For any $0 < \tau < d$, let $C_1(d,\tau) = \psi_{d,\tau} - h_m - \kappa_{\tau,\delta}$, a lower bound on $L_0(\mathcal{M}_{\mathsf{ref}}) - L_0(\mathcal{M}_{\mathsf{cndt}})$, where $\psi_{d,\tau}$ is the Gaussian tail difference, h_m the empirical-CDF error, and $\kappa_{\tau,\delta}$ the residual term from Eq. 1 with $\gamma = \tau B R_{\mathcal{W}}$ (App. A.10). Also define $C_2 = \Phi\big(-\frac{\mu_{\mathsf{ref}}}{\sigma_{\mathsf{ref}}}\big) - \Phi\big(-\frac{\mu_{\mathsf{cndt}}}{\sigma_{\mathsf{cndt}}}\big)$. Whenever the statistical tests and concentration bounds hold, $\mathit{GenGap}(\mathcal{M}_{\mathsf{ref}}) - \mathit{GenGap}(\mathcal{M}_{\mathsf{cndt}}) \geq \max\{C_1(d,\tau), C_2\} - \epsilon$. Hence, if $\max C_1(d,\tau), C_2 > \epsilon$, then $\mathit{GenGap}(\mathcal{M}_{\mathsf{cndt}}) < \mathit{GenGap}(\mathcal{M}_{\mathsf{ref}})$.*

First we give the prof of condition 1 in App. A.10 followed by the proof of condition 2 in App. A.7. We finally of the give the full proof of the theorem in App. A.11. Intuitively, the theorem shows how the two audit conditions translate into lower bounds on the candidate's improvement in population error relative to the reference model. If either bound exceeds the maximum possible advantage of the candidate in training error, then the candidate must have a smaller generalization gap than the reference. The theorem relies on two assumptions. First, we model the p^N distributions as Gaussian. Under this assumption, Condition 2 has a direct interpretation: $L_0(\mathcal{Q}) = \Phi(-\mu_{\mathcal{Q}}/\sigma_{\mathcal{Q}})$, and hence C_2 exactly captures the difference $L_0(\mathcal{M}_{\mathsf{ref}}) - L_0(\mathcal{M}_{\mathsf{cndt}})$. Second, ϵ upper bounds the difference in empirical training errors. In practice, the auditor need not know the candidate's private training error to obtain such a bound. Since $\widehat{L}_{0,T_{\mathsf{cndt}}}(\mathcal{M}_{\mathsf{cndt}}) \geq 0$, the auditor can conservatively set $\epsilon = \widehat{L}_{0,T_{\mathsf{ref}}}(\mathcal{M}_{\mathsf{ref}})$, which is known because the reference model is trained and evaluated in the clear. Thus, choosing a reference model with small training error keeps this conservative bound small. We empirically evaluate the Gaussian approximation and verify that the resulting certified population-error advantage exceeds the training-error difference across our evaluated models.

Soundness. We next consider whether a malicious model owner can game the audit. The natural attack is to train an overfitted model while artificially reducing its spectral complexity $R_{\mathcal{W}}$, thereby inflating the normalized-margin statistic and making the model appear better than it is. We show that this attack is fundamentally limited: if a model has a sufficiently large generalization gap, then its spectral complexity is necessarily lower bounded. Consequently, an overfitted model cannot arbitrarily reduce $R_{\mathcal{W}}$ while preserving its overfitting behavior.

Lemma 3. *Suppose $\mathrm{GenGap}(\mathcal{M}) \geq \tau$ and $\widehat{L}_{\gamma,\mathcal{D}_{train}}(\mathcal{M}) \leq \widehat{L}_{0,\mathcal{D}_{train}}(\mathcal{M}) + \beta_\gamma$. Then, we will have $R_{\mathcal{W}} \geq \frac{\gamma\sqrt{n}}{CB\ln(W)}\left[\tau - \beta_\gamma - c_\delta\sqrt{\frac{\log(1/\delta)}{n}}\right]$.*

We provide the proof in App. A.8. For a large enough dataset (i.e., n is large) the term $c_\delta\sqrt{\frac{\log(1/\delta)}{n}}$ will be negligible. Also since a overfitted model is really confident over it's training dataset, there exists a large enough value of γ such that β_γ is negligible. We empirically validate the above lemma using Fig. 2 and Table 4.

4.4 Efficient ZKP for Verification

In this section, we describe how to efficiently generate the ZKP π_ρ that certifies the correctness of the computation $\rho = (\{p_i\}, R_{\mathcal{W}})$.

Proof of p_i. Verifying the margins reduces to a *proof-of-inference*: a ZKP that the committed

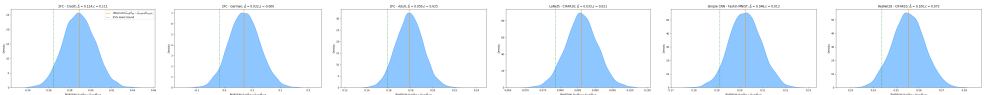

Figure 1: Paired-bootstrap validation of Condition 2 for well-fit candidates ($\alpha = 0.05$). For all models, $\hat{\Delta} \geq \epsilon$, where $\hat{\Delta} := \Phi(-\hat{\mu}_{\text{ref}}/\hat{\sigma}_{\text{ref}}) - \Phi(-\hat{\mu}_{\text{wellfit}}/\hat{\sigma}_{\text{wellfit}})$ and $\epsilon := \widehat{L}_{0,\mathcal{D}^{\text{wellfit}}_{\text{train}}}(\mathcal{M}_{\text{wellfit}}) - \widehat{L}_{0,\mathcal{D}^{\text{ref}}_{\text{train}}}(\mathcal{M}_{\text{ref}})$.

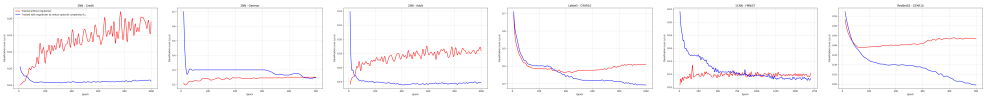

Figure 2: Classification error of overfitted model vs. regularized model as spectral-complexity regularization reduces $R_{\mathcal{W}}$. The regularized model ceases to be overfitted, supporting Lemma 3.

model was correctly evaluated on D_{audit} and that the reported margins are faithful. This is orders of magnitude more efficient than proof-of-training. The prover generates a ZKP for *one* forward pass on m data points–not the entire training trajectory over thousands of examples and hundreds of epochs. Better yet, proof-of-inference is embarrassingly parallel: each evaluation is independent, so all m proofs run concurrently.

Proof of $R_{\mathcal{W}}$. The spectral complexity $R_{\mathcal{W}}$ involves two main operations: (1) computing spectral norms $\|W_i\|_\sigma$ for each weight matrix, and (2) computing various matrix-matrix multiplications. We introduce two key optimizations to make this efficient within a ZKP.

Optimization 1: Upper Bound of the Spectral Norm. The naive approach for proving correctness using ZKP is to replicate the entire computation inside a ZKP circuit. However, computing the spectral norm directly would require power iteration or computing the singular value decomposition, involving multiple loops and resulting in a very large circuit. Our key observation is that for the audit an upper bound of $R_{\mathcal{W}}$ suffices, which can be verified more efficiently. We obtain such an upper bound as follows.

Lemma 4. *For $W \in \mathbb{R}^{m\times n}$ and $\lambda \geq 0$, $\lambda = \|W\|_\sigma$ if and only if $A = \lambda^2 I - W^T W \succeq 0$ and A is singular. Moreover, if $A \succ 0$, then $\lambda > \|W\|_\sigma$.*

Corollary 1. *If, for every layer i, the prover supplies $\bar{\lambda}_i$ such that $\bar{\lambda}_i^2 I - W_i^T W_i \succ 0$, then $\bar{\lambda}_i > \|W_i\|_\sigma$. Substituting these values into Eq. 4.1.1 yields a certified upper bound $\bar{R}_{\mathcal{W}} \geq R_{\mathcal{W}}$.*

The proofs are given in App. A.3 and App. A.9. This result lets us avoid computing the spectral norm exactly. Instead, we certify an upper bound $\lambda > \|W\|_\sigma$ by proving that $A = \lambda^2 I - W^\top W$ is positive definite. By Sylvester's criterion, this reduces to a cheap, ZKP-friendly check that all leading principal minors of A are strictly positive, which small arithmetic circuits can express. Moreover, using $\bar{R}_{\mathcal{W}}$ is *conservative* for our audit. Since the complexity term in Eq. 1 is monotone in $R_{\mathcal{W}}$, replacing $R_{\mathcal{W}}$ with an upper bound can only worsen the resulting generalization gap. The relaxation therefore never favors the prover: an overfitted model *cannot* appear to generalize better than it does.

Optimization 2: Freivalds' Check for Matrix Multiplication. A key operation in the construction is repeated matrix-matrix multiplication. Given three matrices $A, B, C \in \mathbb{R}^{n\times n}$, the goal is to check whether $AB = C$. Explicitly computing the matrix product AB would take $O(n^3)$ time using the naive approach. Instead, we use a randomized verification technique due to Freivalds (1977):

Lemma 5 (Freivald's Check)**.** *Let A, B and C be $n \times n$ matrices over a field $\mathbb{F}$ and let s be a uniformly random vector in S^n, for $S \subseteq \mathbb{F}$. Then,* $\Pr\left[\frac{Cs = A(Bs)}{C \neq AB}\right] = \Pr\left[\frac{(C-AB)s=0}{(C-AB)\neq 0}\right] \leq \frac{1}{|S|}$.

Proof of the above lemma is in App. A.4. Freivalds' check requires only $3n^2$ multiplications-a substantial reduction from $O(n^3)$-and introduces no false negatives. For higher soundness, repeating the check k times lowers the soundness error to $1/|S|^k$. The method also generalizes naturally to rectangular matrices. Next, we present our security guarantee.

Theorem 3. *(Informal) Given a model $\mathcal{M}$ and an audit dataset D_{audit}, ProoFit reveals nothing about $\mathcal{M}$ beyond the margins $\{p_i \mid (x_i, y_i) \in D_{audit}\}$ and the model's spectral complexity $R_{\mathcal{W}}$.*

Proof Sketch. Proof of the above theorem follows directly from the properties of ZKP. The formal guarantee and detailed proof is presented in App. A.13.

5 Implementation and Evaluation

We empirically evaluate *ProoFit* to answer the following key questions:

- **Q1.** How effective is our proposed overfitting test in practice?
- **Q2.** How does the performance of *ProoFit* compare to baseline approaches?

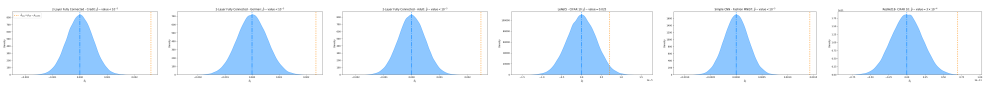

Figure 3: Evaluating the *randomized paired permutation* test over overfit candidate and reference models. The results show this test consistently validates overfitting with high confidence.

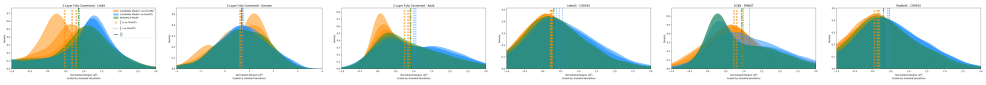

Figure 4: Comparing the scaled normalized margin plots from different instances of overfitting training to find where exactly does the reference model starts testing positive for overfittedness.

Datasets and Models. We evaluate *ProoFit* across five datasets using a diverse set of models. For *Adult* Becker & Kohavi (1996), *Credit* Yeh (2009), and *German* Hofmann (1994), which are widely used tabular datasets, we use a 2-layer fully connected network with hidden layers $(128, 64)$. For *FMNIST* Xiao et al. (2017), we use a 2-layer CNN, SimpleCNN, consisting of a single 3×3 convolutional layer with 32 channels, followed by max-pooling and a fully connected layer. For *CIFAR-10* Krizhevsky et al. (2009), we use LeNet5 LeCun et al. (2002) and ResNet18 He et al. (2016).

Configurations. For *ProoFit*, we use the proof-of-inference protocol from *ZKML* Chen et al. (2024) and implement the proof of spectral complexity using gnark Botrel et al. (2025). We compare against a baseline that generates a ZK proof-of-training. As discussed earlier, the naive baseline mirrors a cross-validation-style audit: it evaluates the model on both the training and test sets and would require ZK proofs for all such evaluations. For a conservative comparison, we only include the cost of generating the ZK proof-of-training itself and exclude the additional proof-of-inference cost (which would make the baseline even more expensive). For CNNs, we use the protocol *KAIZEN* Abbaszadeh et al. (2024). For the fully connected network we implement them using gnark. Because these implementations report only per-iteration runtime, we project it to entire training as: Num Training Epochs $\times \left(\frac{\text{Dataset Size}}{\text{Batch Size}}\right) \times$ Time per Iteration. All experiments were ran on an AWS `u7i-6tb.112xlarge` instance (448 vCPUs, 6 TB RAM). For each dataset, we trained a reference model to satisfy a generalization threshold of $\tau = 0.2$ (Training error - Test error). We used D_{audit} of size set to be 0.1% of the training dataset size, sampled from the same underlying distribution. Empirically, we found this size to be sufficient for reliably evaluating the validity of our overfitting test. Additional details about the setup are in App. A.12.

5.1 Results

Q1: Efficacy of the Overfitting Test. We first evaluate the practical efficacy of our overfitting test and validate the assumptions underlying our analysis. Fig. 3 shows that the paired permutation test consistently distinguishes candidate and reference models for Condition 1, while Fig. 1 shows the same for Condition 2. Fig. 1 further shows that $C_2 > \epsilon$ across all benchmark models and datasets, satisfying $\max C_1(d, \tau), C_2 > \epsilon$ from Theorem 2 and enabling correct identification of well-fitted models. We next test whether these statistics reflect the relative degree of overfitting. For each reference model, we compare candidates with larger generalization gaps (overfitted) and smaller gaps (well-fitted). As shown in Fig. 4, the mean-to-standard-deviation statistic consistently satisfies $\frac{\mu_{\text{overfit}}}{\sigma_{\text{overfit}}} < \frac{\mu_{\text{ref}}}{\sigma_{\text{ref}}} < \frac{\mu_{\text{well-fit}}}{\sigma_{\text{well-fit}}}$ across models and datasets, while the normalized-margin distributions are approximately Gaussian. Repeating the experiment over 100 independently trained candidate models, we report true-positive and true-negative rates in Table 3, measuring how often *ProoFit* correctly distinguishes models on either side of the auditor-defined threshold. Finally, we empirically examine the soundness intuition behind Lemma 3. We emulate a malicious owner attempting to reduce the spectral complexity $R_{\mathcal{W}}$ while preserving overfitting by increasingly regularizing $R_{\mathcal{W}}$ during training. Fig. 2 shows that as $R_{\mathcal{W}}$ decreases, classification error improves until the model is no longer overfitted, supporting the claim that an overfitted model cannot arbitrarily reduce $R_{\mathcal{W}}$ while preserving its behavior. For each overfitted model, we also identify a γ for which β_γ is negligibly small, as required by Lemma 3; these values are reported in Table 4.

Q2: Performance Analysis. Table 1 reports the performance of *ProoFit* and the baseline. The dominant cost in any ZKP pipeline is proof generation, and we observe that *ProoFit* is extremely efficient, requiring 3.69 seconds for smaller models (2-layer fully connected network), and only 161.52 seconds (<3 minutes) for the largest model (ResNet-18). Overall, *ProoFit* achieves a staggering 10^3–$10^7\times$ speedup over the baseline. To put this in perspective, the baseline would require over 17 years to train ResNet-18 for 100 epochs, rendering it entirely infeasible in practice. This gap only

Table 1: Comparison of *ProoFit* with ZK proof-of-training baselines. We use *KAIZEN* Abbaszadeh et al. (2024) for CNNs and a gnark Botrel et al. (2025) implementation for fully connected models. Parentheses show *ProoFit*'s performance improvement.

Model	Dataset	*ProoFit*			Baseline		
		Prover (s)	Verifier (ms)	Size (kB)	Prover per 100 epoch (s)	Verifier (ms)	Size (kB)
2-layer NN	Adult	3.69 ($8.48 \times 10^3 \downarrow$)	17.15	658.32	3.13×10^4	4.75×10^4	1025.00
2-layer NN	Credit	2.28 ($8.03 \times 10^3 \downarrow$)	15.06	371.32	1.83×10^4	4.50×10^4	944.92
2-layer NN	German	2.97 ($0.37 \times 10^3 \downarrow$)	16.01	514.82	0.11×10^4	0.10×10^4	32.03
Simple CNN	FMNIST	37.66 ($0.93 \times 10^7 \downarrow$)	16.81	205.07	3.51×10^8	28.51	416.03
LeNet5	CIFAR10	25.93 ($1.14 \times 10^7 \downarrow$)	43.05	432.57	2.96×10^8	32.65	482.03
ResNet18	CIFAR10	161.52 ($0.35 \times 10^7 \downarrow$)	44.39	861.32	5.64×10^8	86.21	1269.66

Table 2: Breakdown of cost for *ProoFit*. Here π_{p_i} is the Proof Inference for single datapoint and $\pi_{R_{\mathcal{W}}}$ is the Proof of Model's Spectral Complexity

Model	Dataset	π_{p_i} Prover (s)	$\pi_{R_{\mathcal{W}}}$ Prover (s)	Total Prover (s)
2-layer NN	Adult	3.40	0.29	3.69
2-layer NN	Credit	2.00	0.28	2.28
2-layer NN	German	2.69	0.28	2.97
Simple CNN	FMNIST	36.35	1.31	37.66
LeNet5	CIFAR10	21.38	4.55	25.93
ResNet18	CIFAR10	154.10	7.42	161.52

widens with model scale, as the cost of proof generation grows rapidly with training complexity. In sharp contrast, *ProoFit* remains firmly in the realm of practical deployment and can support realistic CNNs. This marks a clear departure from prior cryptographic auditing approaches, which have been unable to scale beyond toy models or simplified settings.

Verification time, i.e., the time taken by the verifier (auditor) to check the validity of the proof, remains extremely small, at most 45 ms across all models. As is standard in ZKPs, verification is significantly faster than proof generation.

We also report communication cost, measured as the size of the proof. *ProoFit* has very low communication cost, requiring < 861.32 KB for all models. Moreover, our approach reduces communication cost by up to $2\times$ compared to the baseline. The only exception is the German dataset, which has a very small training dataset (1K) resulting in a smaller baseline proof size (App. A.12.2).

Next, we present the runtime breakdown of *ProoFit* in Table 2. Our optimized implementation for generating proofs of spectral complexity is highly efficient, requiring only 7.42 seconds even for the largest model. Hence, the overall runtime is dominated by the cost of proof generation for inference.

6 Related Work

ML Generalization Bounds. Generalization theory has progressed from classical complexity measures Bartlett et al. (2019) to margin-based Boser et al. (1992); Bartlett et al. (1998) and norm-based bounds Bartlett (1998); Neyshabur et al. (2015), showing that weight magnitude can be more informative than parameter count, with further refinements via Rademacher complexity Bartlett & Mendelson (2002). PAC-Bayesian approaches Langford & Caruana (2001); Langford & Shawe-Taylor (2002); Dziugaite & Roy (2017) instead relate generalization to the "flatness" of minima. Other measures, such as landscape sharpness Keskar et al. (2016) and Jacobian sensitivity Novak et al. (2018), correlate empirically with generalization but are expensive to compute inside ZKPs. We therefore build on spectrally-normalized bounds Bartlett et al. (2017); Neyshabur et al. (2017b), where generalization depends on products of spectral norms, enabling efficient ZKP constructions.

Confidential Audit. Verifiable inference Ghodsi et al. (2017) has become practical for production-scale models through systems such as *ZKML* Chen et al. (2024) and Halo2 Kang et al. (2022). Since verifying full training Zhao et al. (2021); Abbaszadeh et al. (2024) remains prohibitively expensive, recent work instead verifies post-hoc properties including fairness Zhang et al. (2025b); Franzese et al. (2025); Yadav et al. (2024), explanations Yadav et al. (2025), and calibration Rabanser et al. (2025). To our knowledge, *ProoFit* is the first work to confidentially audit model overfitting.

7 Conclusion

Here we have presented *ProoFit*, a system for auditing model overfitting without compromising model confidentiality. Naive solutions require a computationally prohibitive approach of proving correctness of entire training trajectory using a zero-knowledge proof. *ProoFit* makes cryptographic auditing practical through a novel, weight-only overfitting test grounded in model complexity. This approach yields orders-of-magnitude speedups over baselines across diverse architectures and datasets.

AI USE STATEMENT

In this work, we used generative AI tools for simple polishing the writing of the paper.

ETHICS STATEMENT

This work does not introduce significant new ethical concerns. Instead, *ProoFit* is designed to improve the privacy and accountability of ML auditing by enabling third parties to verify model overfitting without requiring disclosure of proprietary model weights or sensitive training data. This directly addresses the tension between transparency and confidentiality in model auditing, where revealing models or training data can create privacy and intellectual-property risks. By using zero-knowledge proofs to reveal only the information required for the audit, *ProoFit* provides a privacy-preserving mechanism for assessing model generalization while limiting unnecessary exposure of sensitive information.

REPRODUCIBILITY STATEMENT

Visit this link for the anonymized repository of our code https://anonymous.4open.science/r/Privacy-Preserving-Overfitting-Detection-5137/

REFERENCES

Kasra Abbaszadeh, Christodoulos Pappas, Jonathan Katz, and Dimitrios Papadopoulos. Zero-knowledge proofs of training for deep neural networks. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pp. 4316–4330, 2024.

Sultan Amed, Chan Yu Hang, and Sayantan Banerjee. Pdx–adaptive credit risk forecasting model in digital lending using machine learning operations. *arXiv preprint arXiv:2512.22305*, 2025.

Marti J Anderson and John Robinson. Permutation tests for linear models. *Australian & New Zealand Journal of Statistics*, 43(1):75–88, 2001.

Eric Aubinais, Elisabeth Gassiat, and Pablo Piantanida. Fundamental limits of membership inference attacks on machine learning models. *Journal of Machine Learning Research*, 26(263):1–54, 2025.

David Balbás, Dario Fiore, María Isabel González Vasco, Damien Robissout, and Claudio Soriente. Modular sumcheck proofs with applications to machine learning and image processing. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1437–1451, 2023.

Peter Bartlett. For valid generalization the size of the weights is more important than the size of the network. *Advances in neural information processing systems*, 9, 1996.

Peter Bartlett, Yoav Freund, Wee Sun Lee, and Robert E Schapire. Boosting the margin: A new explanation for the effectiveness of voting methods. *The annals of statistics*, 26(5):1651–1686, 1998.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of machine learning research*, 3(Nov):463–482, 2002.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30, 2017.

Peter L Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17, 2019.

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

P.L. Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory*, 44(2):525–536, 1998. doi: 10.1109/18.661502.

Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

Manuel Blum. Coin flipping by telephone a protocol for solving impossible problems. In *Proceedings of the 24th IEEE Computer Society International Conference (COMPCON)*, pp. 133–137. IEEE, 1981.

Dan Boneh, Wilson Nguyen, and Alex Ozdemir. Efficient functional commitments: How to commit to a private function. Cryptology ePrint Archive, Paper 2021/1342, 2021. URL https://eprint.iacr.org/2021/1342.

Jaydeep Borkar. What can we learn from data leakage and unlearning for law? *arXiv preprint arXiv:2307.10476*, 2023.

Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pp. 144–152, 1992.

Gautam Botrel, Thomas Piellard, Youssef El Housni, Ivo Kubjas, and Arya Tabaie. Consensys/gnark: v0.14.0, June 2025. URL https://doi.org/10.5281/zenodo.5819104.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Yuan Cao, Zixiang Chen, Misha Belkin, and Quanquan Gu. Benign overfitting in two-layer convolutional neural networks. *Advances in neural information processing systems*, 35:25237–25250, 2022.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX security symposium (USENIX security 19)*, pp. 267–284, 2019.

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

Nicholas Carlini, Matthew Jagielski, Chiyuan Zhang, Nicolas Papernot, Andreas Terzis, and Florian Tramer. The privacy onion effect: Memorization is relative. *Advances in Neural Information Processing Systems*, 35:13263–13276, 2022.

Nicholas Carlini, Daniel Paleka, Krishnamurthy Dj Dvijotham, Thomas Steinke, Jonathan Hayase, A. Feder Cooper, Katherine Lee, Matthew Jagielski, Milad Nasr, Arthur Conmy, Itay Yona, Eric Wallace, David Rolnick, and Florian Tramèr. Stealing part of a production language model, 2024. URL https://arxiv.org/abs/2403.06634.

George Casella and Roger L. Berger. *Statistical Inference*. Duxbury, 2 edition, 2002.

Bing-Jyue Chen, Suppakit Waiwitlikhit, Ion Stoica, and Daniel Kang. Zkml: An optimizing system for ml inference in zero-knowledge proofs. In *Proceedings of the Nineteenth European Conference on Computer Systems*, pp. 560–574, 2024.

Bing-Jyue Chen, Lilia Tang, and Daniel Kang. Zktorch: Compiling ml inference to zero-knowledge proofs via parallel proof accumulation. *arXiv preprint arXiv:2507.07031*, 2025.

Debeshee Das, Jie Zhang, and Florian Trantèr. Blind baselines beat membership inference attacks for foundation models. In *2025 IEEE Security and Privacy Workshops (SPW)*, pp. 118–125. IEEE, 2025.

Daniel DeAlcala, Aythami Morales, Julian Fierrez, Gonzalo Mancera, Ruben Tolosana, and Javier Ortega-Garcia. Is my data in your ai model? membership inference test with application to face images. *arXiv e-prints*, pp. arXiv–2402, 2024.

Daniel DeAlcala, Aythami Morales, Julian Fierrez, Gonzalo Mancera, and Ruben Tolosana. gmint: Gradiant-based membership inference test applied to image models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2781–2790, 2025a.

Daniel DeAlcala, Aythami Morales, Julian Fierrez, Gonzalo Mancera, Ruben Tolosana, and Javier Ortega-Garcia. Active membership inference test (amint): Enhancing model auditability with multi-task learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 647–656, 2025b.

Vassilis Digalakis Jr, Christophe Pérignon, Sébastien Saurin, and Flore Sentenac. The challenger: When do new data sources justify switching machine learning models? *arXiv preprint arXiv:2512.18390*, 2025.

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

Sunny Duan, Mikail Khona, Abhiram Iyer, Rylan Schaeffer, and Ila R Fiete. Uncovering latent memories: Assessing data leakage and memorization patterns in frontier ai models. *arXiv preprint arXiv:2406.14549*, 2024b.

Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.

Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. Chapman and Hall/CRC, 1994.

Bradley Efron, Robert J Tibshirani, et al. *An introduction to the bootstrap*. Boca Raton, Florida, 2000.

Beyza Eken, Samodha Pallewatta, Nguyen Tran, Ayse Tosun, and Muhammad Ali Babar. A multivocal review of mlops practices, challenges and open issues. *ACM Computing Surveys*, 58(2):1–35, 2025.

Soumia Zohra El Mestari, Gabriele Lenzini, and Huseyin Demirci. Preserving data privacy in machine learning systems. *Computers & Security*, 137:103605, 2024.

Michael D Ernst. Permutation methods: a basis for exact inference. *Statistical Science*, pp. 676–685, 2004.

European Data Protection Supervisor. *AI Act Regulation (EU) 2024/1689 – Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance)*. Publications Office of the European Union, 2025. doi: doi/10.2804/4225375.

Executive Office of the President. Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. 88 FR 75191. `https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence`, November 2023. Accessed: 2026-09-26.

Sir Ronald Aylmer Fisher and Ronald A Fisher. *The design of experiments*, volume 21. Springer, 1966.

Olive Franzese. *Cryptographic Verification of Trustworthy Machine Learning*. PhD thesis, Northwestern University, 2024.

Olive Franzese, Ali Shahin Shamsabadi, Carter Luck, and Hamed Haddadi. Secure and confidential certificates of online fairness. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Rusins Freivalds. Probabilistic machines can use less running time. In *IFIP Congress*, 1977. URL https://api.semanticscholar.org/CorpusID:45405368.

Georgi Ganev, Meenatchi Sundaram Muthu Selva Annamalai, and Bogdan Kulynych. Tight auditing of differential privacy in mst and aim. *arXiv preprint arXiv:2604.18352*, 2026.

Sanjam Garg, Aarushi Goel, Somesh Jha, Saeed Mahloujifar, Mohammad Mahmoody, Guru-Vamsi Policharla, and Mingyuan Wang. Experimenting with zero-knowledge proofs of training. In *Proceedings of the 2023 ACM SIGSAC conference on computer and communications security*, pp. 1880–1894, 2023.

Zahra Ghodsi, Tianyu Gu, and Siddharth Garg. Safetynets: Verifiable execution of deep neural networks on an untrusted cloud. *Advances in Neural Information Processing Systems*, 30, 2017.

Oded Goldreich, Silvio Micali, and Avi Wigderson. Proofs that yield nothing but their validity or all languages in np have zero-knowledge proof systems. *J. ACM*, 38(3):690–728, jul 1991. ISSN 0004-5411. doi: 10.1145/116825.116852. URL https://doi.org/10.1145/116825.116852.

S Goldwasser, S Micali, and C Rackoff. The knowledge complexity of interactive proof-systems. In *Proceedings of the Seventeenth Annual ACM Symposium on Theory of Computing*, STOC '85, pp. 291–304, New York, NY, USA, 1985. Association for Computing Machinery. ISBN 0897911512. doi: 10.1145/22145.22178. URL https://doi.org/10.1145/22145.22178.

Tomás González, Mateo Dulce-Rubio, Aaditya Ramdas, and Mónica Ribero. Sequentially auditing differential privacy. *arXiv preprint arXiv:2509.07055*, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Peter Henderson, Koustuv Sinha, Nicolas Angelard-Gontier, Nan Rosemary Ke, Genevieve Fried, Ryan Lowe, and Joelle Pineau. Ethical challenges in data-driven dialogue systems. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, pp. 123–129, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360128. doi: 10.1145/3278721.3278777. URL https://doi.org/10.1145/3278721.3278777.

Hans Hofmann. Statlog (German Credit Data). UCI Machine Learning Repository, 1994. DOI: https://doi.org/10.24432/C5NC77.

R.A. Horn and C.R. Johnson. *Matrix Analysis*. Cambridge University Press, 2012. ISBN 9781139788885. URL https://books.google.com/books?id=O7sgAwAAQBAJ.

Zonghao Huang, Neil Zhenqiang Gong, and Michael K Reiter. A general framework for data-use auditing of ml models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pp. 1300–1314, 2024.

Innovation, Science and Economic Development Canada. The Artificial Intelligence and Data Act (AIDA) – Companion document. https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document, March 2023. Accessed: 2026-09-26.

Daniel Kang, Tatsunori Hashimoto, Ion Stoica, and Yi Sun. Scaling up trustless dnn inference with zero-knowledge proofs. *arXiv preprint arXiv:2210.08674*, 2022.

Aniket Kate, Gregory M. Zaverucha, and Ian Goldberg. Constant-size commitments to polynomials and their applications. In Masayuki Abe (ed.), *Advances in Cryptology - ASIACRYPT 2010*, pp. 177–194, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. ISBN 978-3-642-17373-8.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

Zhifeng Kong, Amrita Roy Chowdhury, and Kamalika Chaudhuri. Can membership inferencing be refuted? *arXiv preprint arXiv:2303.03648*, 2023.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

John Langford and Rich Caruana. (not) bounding the true error. *Advances in Neural Information Processing Systems*, 14, 2001.

John Langford and John Shawe-Taylor. Pac-bayes & margins. *Advances in neural information processing systems*, 15, 2002.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 2002.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. *arXiv preprint arXiv:2107.06499*, 2021.

Seunghwa Lee, Hankyung Ko, Jihye Kim, and Hyunok Oh. vcnn: Verifiable convolutional neural network based on zk-snarks. *IEEE Transactions on Dependable and Secure Computing*, 21(4): 4254–4270, 2024.

Klas Leino and Matt Fredrikson. Stolen memories: Leveraging model memorization for calibrated {White-Box} membership inference. In *29th USENIX security symposium (USENIX Security 20)*, pp. 1605–1622, 2020.

Xiao Li, Qiongxiu Li, Zhanhao Hu, and Xiaolin Hu. On the privacy effect of data enhancement via the lens of memorization. *IEEE Transactions on Information Forensics and Security*, 19:4686–4699, 2024.

Zhu Li, Zhi-Hua Zhou, and Arthur Gretton. Towards an understanding of benign overfitting in neural networks. *arXiv preprint arXiv:2106.03212*, 2021.

Tianyi Liu, Xiang Xie, and Yupeng Zhang. Zkcnn: Zero knowledge proofs for convolutional neural network predictions and accuracy. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pp. 2968–2985, 2021.

Hidde Lycklama, Nicolas Küchler, Alexander Viand, Emanuel Opel, Lukas Burkhalter, and Anwar Hithnawi. Cryptographic auditing for collaborative learning. In *NeurIPS ML Safety Workshop*, 2022.

Hidde Lycklama, Alexander Viand, Nicolas Küchler, Christian Knabenhans, and Anwar Hithnawi. Holding secrets accountable: Auditing privacy-preserving machine learning. In *USENIX Security Symposium*, 2024.

Gonzalo Mancera, Daniel DeAlcala, Julian Fierrez, Ruben Tolosana, and Aythami Morales. Is my text in your ai model? gradient-based membership inference test applied to llms. *arXiv preprint arXiv:2503.07384*, 2025.

Gonzalo Mancera, Daniel DeAlcala, Aythami Morales, Ruben Tolosana, and Julian Fierrez. Membership inference test: Auditing training data in object classification models. *arXiv preprint arXiv:2601.12929*, 2026.

Neal Mangaokar, Ashish Hooda, Zhuohang Li, Bradley A Malin, Kassem Fawaz, Somesh Jha, Atul Prakash, and Amrita Roy Chowdhury. What really is a member? discrediting membership inference via poisoning. *arXiv preprint arXiv:2506.06003*, 2025.

David A McAllester. Some pac-bayesian theorems. In *Proceedings of the eleventh annual conference on Computational learning theory*, pp. 230–234, 1998.

David A McAllester. Pac-bayesian model averaging. In *Proceedings of the twelfth annual conference on Computational learning theory*, pp. 164–170, 1999.

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

Shyam Varan Nath. Champion-challenger based predictive model selection. In *Proceedings 2007 IEEE SoutheastCon*, pp. 254–254. IEEE, 2007.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Conference on learning theory*, pp. 1376–1401. PMLR, 2015.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. *Advances in neural information processing systems*, 30, 2017a.

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. *arXiv preprint arXiv:1707.09564*, 2017b.

Roman Novak, Yasaman Bahri, Daniel A Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Sensitivity and generalization in neural networks: an empirical study. *arXiv preprint arXiv:1802.08760*, 2018.

Jukka Nyblom. Permutation tests in linear regression. In *Modern Nonparametric, Robust and Multivariate Methods: Festschrift in Honour of Hannu Oja*, pp. 69–90. Springer, 2015.

Office of the Auditor General of Norway. Auditing machine learning algorithms. `https://www.auditingalgorithms.net/`, 2023.

Daryna Oliynyk, Rudolf Mayer, and Andreas Rauber. I know what you trained last summer: A survey on stealing machine learning models and defences. *ACM Computing Surveys*, 55(14s):1–41, 2023.

Saerom Park, Seongmin Kim, and Yeon-sup Lim. Fairness audit of machine learning models with confidential computing. In *Proceedings of the ACM Web Conference 2022*, pp. 3488–3499, 2022.

Belinda Phipson and Gordon K Smyth. Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *arXiv preprint arXiv:1603.05766*, 2016.

Wenjie Qu, Yanpei Guo, Yue Ying, and Jiaheng Zhang. Verfcnn, optimal complexity zksnark for convolutional neural networks. *Cryptology ePrint Archive*, 2025.

Stephan Rabanser, Ali Shahin Shamsabadi, Olive Franzese, Xiao Wang, Adrian Weller, and Nicolas Papernot. Confidential guardian: Cryptographically prohibiting the abuse of model abstention. *arXiv preprint arXiv:2505.23968*, 2025.

Badrinath Ramakrishnan and Akshaya Balaji. Assessing and mitigating data memorization risks in fine-tuned large language models. *arXiv preprint arXiv:2508.14062*, 2025.

PII Revisiting. Do llms really memorize personally identifiable information? revisiting pii leakage with a cue-controlled memorization framework.

Tom Sander, Bargav Jayaraman, Mark Ibrahim, Kamalika Chaudhuri, and Chuan Guo. Rethinking the role of verbatim memorization in llm privacy. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Ali Shahin Shamsabadi, Sierra Calanda Wyllie, Nicholas Franzese, Natalie Dullerud, Sébastien Gambs, Nicolas Papernot, Xiao Wang, and Adrian Weller. Confidential-profitt: Confidential proof of fair training of trees. In *The Eleventh International Conference on Learning Representations*, 2022.

Liwei Song and Prateek Mittal. Systematic evaluation of privacy risks of machine learning models. In *30th USENIX security symposium (USENIX security 21)*, pp. 2615–2632, 2021.

Ambuj Tewari and Peter L Bartlett. On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8(5), 2007.

Om Thakkar, Swaroop Ramaswamy, Rajiv Mathews, and Françoise Beaufays. Understanding unintended memorization in federated learning. *arXiv preprint arXiv:2006.07490*, 2020.

Om Dipakbhai Thakkar, Swaroop Ramaswamy, Rajiv Mathews, and Francoise Beaufays. Understanding unintended memorization in language models under federated learning. In Oluwaseyi Feyisetan, Sepideh Ghanavati, Shervin Malmasi, and Patricia Thaine (eds.), *Proceedings of the Third Workshop on Privacy in Natural Language Processing*, pp. 1–10, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.privatenlp-1.1. URL https://aclanthology.org/2021.privatenlp-1.1.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

Vladimir N Vapnik and Alexey Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971.

Suppakit Waiwitlikhit, Ion Stoica, Yi Sun, Tatsunori Hashimoto, and Daniel Kang. Trustless audits without revealing data or models. *arXiv preprint arXiv:2404.04500*, 2024.

Chenkai Weng, Kang Yang, Jonathan Katz, and Xiao Wang. Wolverine: Fast, scalable, and communication-efficient zero-knowledge proofs for boolean and arithmetic circuits. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 1074–1091. IEEE, 2021a.

Chenkai Weng, Kang Yang, Xiang Xie, Jonathan Katz, and Xiao Wang. Mystique: Efficient conversions for {Zero-Knowledge} proofs with applications to machine learning. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 501–518, 2021b.

Jiasi Weng, Jian Weng, Gui Tang, Anjia Yang, Ming Li, and Jia-Nan Liu. pvcnn: Privacy-preserving and verifiable convolutional neural network testing. *IEEE Transactions on Information Forensics and Security*, 18:2218–2233, 2023.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

Chhavi Yadav, Amrita Roy Chowdhury, Dan Boneh, and Kamalika Chaudhuri. Fairproof: Confidential and certifiable fairness for neural networks. *arXiv preprint arXiv:2402.12572*, 2024.

Chhavi Yadav, Evan Monroe Laufer, Dan Boneh, and Kamalika Chaudhuri. Expproof: Operationalizing explanations for confidential models with zkps. *arXiv preprint arXiv:2502.03773*, 2025.

I-Cheng Yeh. Default of Credit Card Clients. UCI Machine Learning Repository, 2009. DOI: https://doi.org/10.24432/C55S3H.

Samuel Yeom, Matt Fredrikson, and Somesh Jha. The unintended consequences of overfitting: Training data inference attacks. *CoRR*, abs/1709.01604, 2017. URL http://arxiv.org/abs/1709.01604.

Chenxiang Zhang, Jun Pang, and Sjouke Mauw. Spurious privacy leakage in neural networks. *arXiv preprint arXiv:2505.20095*, 2025a.

Jie Zhang, Debeshee Das, Gautam Kamath, and Florian Tramèr. Membership inference attacks cannot prove that a model was trained on your data. *arXiv preprint arXiv:2409.19798*, 2024.

Kaiyuan Zhang. *Memorization in Private and Secure Machine Learning*. PhD thesis, Purdue University Graduate School, 2026.

Tianyu Zhang, Shen Dong, O Deniz Kose, Yanning Shen, and Yupeng Zhang. Fairzk: A scalable system to prove machine learning fairness in zero-knowledge. In *2025 IEEE Symposium on Security and Privacy (SP)*, pp. 3460–3478. IEEE, 2025b.

Tong Zhang. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2(Mar):527–550, 2002.

Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–85, 2004.

Lingchen Zhao, Qian Wang, Cong Wang, Qi Li, Chao Shen, and Bo Feng. Veriml: Enabling integrity assurances and fair payments for machine learning as a service. *IEEE Transactions on Parallel and Distributed Systems*, 32(10):2524–2540, 2021.

Pan Zhou and Jiashi Feng. Understanding generalization and optimization performance of deep cnns. In *International Conference on Machine Learning*, pp. 5960–5969. PMLR, 2018.

A Appendix

A.1 Proof of Lemma 1

Proof. We aim to show that the paired permutation test is a statistically valid procedure that exactly controls the Type I error rate at level α for paired distributions, independent of any underlying parametric assumptions (such as normality or sub-Gaussianity).

Let (X_i, Y_i) be paired observations drawn from an unknown joint distribution, and define the difference $D_i = X_i - Y_i$. We wish to test the null hypothesis H_0 that the marginal distributions of X and Y are identical, against the alternative that the mean of X is systematically shifted relative to Y.

Under H_0, the random variables X_i and Y_i are exchangeable. This implies that the paired difference $D_i = X_i - Y_i$ and its negation $-D_i = Y_i - X_i$ have the exact same distribution. Therefore, the distribution of D_i is perfectly symmetric around zero.

Because of this symmetry, if we condition on the absolute values of the observed differences, $|D| = (|D_1|, \ldots, |D_m|)$, the signs of the original differences D_i are independent and equally likely to be positive or negative. There are exactly 2^m possible sign assignments, denoted by the set $\mathcal{S} = \{-1, 1\}^m$. Under H_0, the conditional probability of observing any specific sequence of signs $S \in \mathcal{S}$ is:

$$P\left(S \middle| |D|, H_0\right) = \frac{1}{2^m}$$

Let $T(D)$ be our chosen test statistic (e.g., the sample mean difference). By applying all 2^m sign assignments to $|D|$, we construct the exact conditional null distribution of T. The conditional p-value is calculated as the fraction of this finite set of statistics that are as extreme as the observed statistic T_{obs}:

$$P\left(T \leq T_{\text{obs}} \middle| |D|, H_0\right) = \frac{1}{2^m} \sum_{S \in \mathcal{S}} \mathbb{I}(T_S \leq T_{\text{obs}})$$

Because this construction evaluates all possible configurations under the assumption of exchangeability, the test guarantees that the probability of incorrectly rejecting H_0 (Type I error) is exactly α when conditioned on $|D|$:

$$P\left(\text{p-value} \leq \alpha \middle| |D|, H_0\right) \leq \alpha$$

By taking the expectation over the marginal distribution of the absolute differences $|D|$, the unconditional Type I error rate is also bounded by α:

$$\mathbb{E}_{|D|}\left[P\left(\text{p-value} \leq \alpha \middle| |D|, H_0\right)\right] \leq \alpha$$

Thus, the paired permutation test is exactly valid for comparing paired distributions without requiring asymptotic approximations or assumptions about the shape of the underlying distributions.

Given $|\mathcal{S}|$ grows exponentially with m it becomes infeasible to construct the exact conditional null distribution T. For a large K sampling $\{S_i\}_{i \in [K]}$ drawn iid uniformly from $\mathcal{S}$ we can get a good estimate of the conditional null distribution $\hat{T}$. Thus the estimated p̂-value will be:

$$P\left(\hat{T} \leq T_{\text{obs}} \middle| |D|, H_0\right) = \frac{1}{K} \sum_{i \in [K]} \mathbb{I}(T_{S_i} \leq T_{\text{obs}})$$

□

A.2 Proof of Lemma 2

Proof. We aim to show that the paired non-parametric bootstrap provides a statistically valid procedure for testing whether the mean-to-standard deviation ratio of the reference model is larger than that of the candidate model, while preserving the dependence induced by evaluating both models on the same audit samples. Importantly, the procedure does not require the normalized-margin distributions to be Gaussian or sub-Gaussian. Let

$$Z_i = (X_i, Y_i), \qquad i \in [m],$$

be i.i.d. paired observations drawn from an unknown joint distribution P, where X_i and Y_i denote the normalized margins of the reference and candidate models, respectively, on the same audit sample. Let

$$\mu_X = \mathbb{E}[X], \qquad \mu_Y = \mathbb{E}[Y],$$

and

$$\sigma_X^2 = \mathrm{Var}(X), \qquad \sigma_Y^2 = \mathrm{Var}(Y).$$

We assume $\sigma_X, \sigma_Y > 0$ and that the required moments are finite so that the empirical means and variances consistently estimate their population counterparts. In particular, finite fourth moments are sufficient for the argument below. Define the population statistic

$$\Delta(P) = \frac{\mu_X}{\sigma_X} - \frac{\mu_Y}{\sigma_Y}.$$

We test

$$H_0 : \Delta(P) \le 0 \qquad \text{against} \qquad H_1 : \Delta(P) > 0.$$

Given the empirical distribution

$$\hat{P}_m = \frac{1}{m} \sum_{i=1}^{m} \delta_{Z_i},$$

the corresponding empirical statistic is

$$\hat{\Delta} = \Delta(\hat{P}_m) = \frac{\hat{\mu}_X}{\hat{\sigma}_X} - \frac{\hat{\mu}_Y}{\hat{\sigma}_Y}.$$

The important observation is that $\Delta(P)$ is a functional of the *joint* distribution P of the pair (X, Y) rather than of the two marginal distributions independently. Indeed, the sampling variability of

$$\hat{\Delta} = \frac{\hat{\mu}_X}{\hat{\sigma}_X} - \frac{\hat{\mu}_Y}{\hat{\sigma}_Y}$$

depends not only on the marginal variability of X and Y, but also on their dependence. Since the two models are evaluated on the same audit example, X_i and Y_i may be strongly correlated. Consequently, independently resampling the two collections $\{X_i\}_{i=1}^m$ and $\{Y_i\}_{i=1}^m$ would in general fail to reproduce the correct sampling distribution of $\hat{\Delta}$. The paired bootstrap instead draws

$$Z_1^*, \ldots, Z_m^* \overset{\text{i.i.d.}}{\sim} \hat{P}_m,$$

where each

$$Z_\ell^* = (X_{I_\ell}, Y_{I_\ell})$$

for an index I_ℓ drawn uniformly with replacement from $[m]$. Equivalently, the same resampled index is always used for both models. Hence, every bootstrap sample preserves the empirical joint distribution, and therefore also preserves the empirical dependence between the reference and candidate margins. For each bootstrap sample, define

$$\hat{\Delta}^* = \frac{\hat{\mu}_X^*}{\hat{\sigma}_X^*} - \frac{\hat{\mu}_Y^*}{\hat{\sigma}_Y^*}.$$

Conditioned on the observed audit data, the distribution of $\hat{\Delta}^* - \hat{\Delta}$ is the bootstrap approximation to the sampling distribution of $\hat{\Delta} - \Delta(P)$. To see why this approximation is valid, consider the vector of population moments

$$\theta = \left(\mu_X, \mu_Y, \mathbb{E}[X^2], \mathbb{E}[Y^2]\right),$$

with empirical counterpart

$$\hat{\theta} = \left(\hat{\mu}_X, \hat{\mu}_Y, \widehat{\mathbb{E}[X^2]}, \widehat{\mathbb{E}[Y^2]}\right).$$

The statistic Δ can be written as the smooth function

$$g(\theta) = \frac{\mu_X}{\sqrt{\mathbb{E}[X^2] - \mu_X^2}} - \frac{\mu_Y}{\sqrt{\mathbb{E}[Y^2] - \mu_Y^2}}.$$

Since $\sigma_X, \sigma_Y > 0$, the function g is continuously differentiable in a neighborhood of the true parameter θ. Therefore, by the multivariate central limit theorem and the delta method,

$$\sqrt{m}\left(\hat{\Delta} - \Delta(P)\right) \xrightarrow{d} G,$$

for some mean-zero limiting distribution G determined by the joint distribution of (X, Y). Applying the same argument conditionally to bootstrap samples drawn from $\hat{P}_m$ gives

$$\sqrt{m}\left(\hat{\Delta}^* - \hat{\Delta}\right) \xrightarrow{d} G$$

in probability. Equivalently, if F_m denotes the sampling distribution of $\sqrt{m}(\hat{\Delta} - \Delta(P))$ and F_m^* denotes the conditional bootstrap distribution of $\sqrt{m}(\hat{\Delta}^* - \hat{\Delta})$, then

$$\sup_{t \in \mathbb{R}} |F_m^*(t) - F_m(t)| \xrightarrow{p} 0.$$

Thus, the paired bootstrap consistently estimates the sampling distribution of the statistic of interest. Let q_α^* denote the conditional empirical α-quantile of the bootstrap distribution of $\hat{\Delta}^*$. The corresponding one-sided bootstrap lower confidence bound is

$$\hat{L}_{1-\alpha} = q_\alpha^*.$$

By bootstrap consistency, this bound satisfies asymptotically

$$\Pr\left(\Delta(P) \geq \hat{L}_{1-\alpha}\right) = 1 - \alpha + o(1).$$

Therefore, if

$$\hat{L}_{1-\alpha} > 0,$$

then zero is excluded from the one-sided $(1-\alpha)$ confidence region for $\Delta(P)$, and we reject

$$H_0 : \Delta(P) \leq 0$$

in favor of

$$H_1 : \Delta(P) > 0.$$

Equivalently, under the boundary null hypothesis $\Delta(P) = 0$, the probability of rejection satisfies

$$\Pr_{H_0}\left(\hat{L}_{1-\alpha} > 0\right) \leq \alpha + o(1).$$

Hence, the test asymptotically controls the Type I error rate at level α. We emphasize that this argument does not assume that X or Y follows a Gaussian or sub-Gaussian distribution. The bootstrap distribution is constructed directly from the empirical paired observations and therefore retains the skewness, heavy-tailed behavior, and dependence present in the observed data. The required assumptions are instead regularity conditions ensuring that the statistic is well defined and that the empirical first and second moments converge appropriately. Thus, moderately heavy-tailed and skewed normalized-margin distributions are permitted, provided the relevant moments are finite and σ_X, σ_Y are non-zero. This is particularly useful in our audit setting because $\hat{\mu}/\hat{\sigma}$ is a nonlinear statistic involving both the empirical mean and empirical standard deviation. A direct parametric approximation would require specifying or approximating the joint sampling distribution of these quantities. The paired non-parametric bootstrap instead approximates the sampling distribution of the entire statistic $\hat{\Delta}$ directly, while automatically retaining both the correlation between the estimated mean and variance within each model and the correlation between the two models induced by their evaluation on the same audit samples. Finally, the ideal bootstrap distribution contains all possible samples of size m drawn with replacement from the empirical distribution $\hat{P}_m$. Since enumerating all

such resamples is computationally infeasible, we approximate it by drawing K independent paired bootstrap samples. Let

$$\hat{\Delta}^{(1)}, \ldots, \hat{\Delta}^{(K)}$$

denote the resulting bootstrap statistics. The empirical bootstrap distribution is

$$\hat{F}_K(t) = \frac{1}{K} \sum_{j=1}^{K} \mathbb{I}\left(\hat{\Delta}^{(j)} \leq t\right).$$

By the law of large numbers, for every continuity point t of the conditional bootstrap distribution,

$$\hat{F}_K(t) \xrightarrow[K\to\infty]{a.s.} F_m^*(t).$$

Consequently, the empirical α-quantile obtained from a sufficiently large number of bootstrap samples converges to the corresponding bootstrap quantile. Thus, for large K, the computed lower confidence bound $\hat{L}_{1-\alpha}$ provides an arbitrarily accurate Monte Carlo approximation to the paired-bootstrap confidence bound. Therefore, under the stated regularity conditions, the paired non-parametric bootstrap consistently estimates the sampling distribution of

$$\frac{\hat{\mu}_{\text{cndt}}}{\hat{\sigma}_{\text{cndt}}} - \frac{\hat{\mu}_{\text{ref}}}{\hat{\sigma}_{\text{ref}}},$$

preserves the dependence between the two models' margins, requires no Gaussian or sub-Gaussian assumption on their underlying distributions, and yields an asymptotically level-α test for determining whether

$$\frac{\mu_{\text{ref}}}{\sigma_{\text{ref}}} < \frac{\mu_{\text{cndt}}}{\sigma_{\text{cndt}}}.$$

□

A.3 Proof of Lemma 4

Proof. Recall that the spectral norm of W satisfies $\|W\|_\sigma^2 = \lambda_{\max}(W^T W)$, where $W^T W$ is symmetric and positive semidefinite. By definition, we know that $\lambda = \|W\|_\sigma$ if and only if $\lambda^2 \geq \lambda_{\max}(W^T W)$ and λ^2 is an eigenvalue of $W^T W$.

The first condition is equivalent to say that $\lambda^2 I - W^T W \succeq 0$ and the second condition is equivalent to $\det(\lambda^2 I - W^T W) = 0$. Thus, $\lambda = \|W\|_\sigma$ if and only if $A := \lambda^2 I - W^T W$ is positive semidefinite and singular. Moreover, since we know that A is symmetric, a classical identification in linear algebra gives us that $A \succeq 0$ is equivalent to say that $\det(A_I) \geq 0$ for all principal $i \times i$ submatrices A_I where $i = 1, \ldots, n$ and it is a well-known fact in linear algebra that this is equivalent to say that all eigenvalues of A is nonnegative Horn & Johnson (2012). □

A.4 Proof of Lemma 5

Proof. For Completeness, if $AB = C$, then for every r, we have $A(Br) = (AB)r = Cr$, hence the check always passes.

For Soundness, we assume that $AB \neq C$ and we denote $D := AB - C \neq 0$ The test accepts if and only if $A(Br) = Cr$, which is equivalent to $Dr = 0$. Since we know $D \neq 0$, there exists at least one row d^T of D that is nonzero. If $Dr = 0$, then we have $d^T r = 0$. Choose an index t such that $d^t \neq 0$. Conditioning on all coordinates of r except r^t, and expanding the equation $d^T r = \sum_{i=1}^n d_i r_i = 0$ becomes a nontrivial linear equation in the single variable r_t, hence it can be satisfied by at most one value of $r_t \in S$. Since we know r_t is uniform in S, then

$$\Pr[d^T r = 0 \mid r_1, \ldots, r_{t-1}, r_{t+1}, \ldots, r_n] \leq \frac{1}{|S|}$$

Taking expectation over the conditioning gives us $\Pr[d^T r = 0] \leq 1/|S|$. Since the event $\{Dr = 0\} \subset \{d^T r = 0\}$, we have $\Pr[Dr = 0] \leq \Pr[d^T r = 0] \leq 1/|S|$. Repeating the check independently k times and we accept only if all checks pass, then by independence the error probability is at most $(1/|S|)^k = |S|^{-k}$.

□

A.5 LEMMA ABOUT σ-SUB-GAUSSIAN VALIDATION TEST

Lemma 6. *Let $\mathcal{P}^N$ denote the normalized margin distribution of the candidate model, and assume $\mathcal{P}^N$ is σ-sub-Gaussian with mean μ_{cndt}. Let $\hat{\mu}_{cndt} = \frac{1}{m}\sum_{i=1}^{m} p_i^N$ be the empirical mean over sample size m, and let $\hat{\mu}_{ref}$ denote the mean normalized margin of the reference model. If*

$$\hat{\mu}_{cndt} \leq \hat{\mu}_{ref} + (-)\sigma\sqrt{\frac{2\log(1/\beta)}{m}},$$

then with probability at least $1-\beta$, the true mean satisfies $\hat{\mu}_{cndt} \leq \hat{\mu}_{ref}$, implying that the candidate model fails the audit.

A.6 PROOF OF LEMMA 6

Remove:

Proof. The theorem is an application of Chernoff Bound for σ-sub-Gaussian distributions. From Chernoff Bound we know that for all $t > 0$, $P(\hat{\mu}_{\text{ref}} - \hat{\mu}_{\text{cndt}} \geq t) \leq \exp(-\frac{mt^2}{2\sigma^2})$. We set $t = \sigma\sqrt{\frac{2\log(1/\beta)}{m}}$, and get

$$P(\hat{\mu}_{\text{ref}} - \hat{\mu}_{\text{cndt}} \geq \sigma\sqrt{\frac{2\log(1/\beta)}{m}}) \leq \beta$$

Rearranging and taking the complement event gives us

$$P(\hat{\mu}_{\text{ref}} \leq \hat{\mu}_{\text{cndt}} + \sigma\sqrt{\frac{2\log(1/\beta)}{m}}) \geq 1-\beta$$

Hence in our test, if we observe

$$\hat{\mu}_{\text{cndt}} + \sigma\sqrt{\frac{2\log(1/\beta)}{m}} \leq \hat{\mu}_{\text{ref}}$$

then we may conclude $\hat{\mu}_{\text{cndt}} \leq \hat{\mu}_{\text{ref}}$ with probability at least $1-\beta$. □

A.7 PROOF OF CONDITION 2

Proof. We first show that the ordering of the standardized normalized margins implies an ordering of the population classification losses.

Recall that the normalized margin is obtained from the ordinary margin by dividing by a strictly positive normalization factor. Therefore, normalization preserves its sign, and hence

$$L_0(\mathcal{M}_{\mathsf{cndt}}) = \Pr[p^N_{\mathsf{cndt}} \leq 0], \qquad L_0(\mathcal{M}_{\mathsf{ref}}) = \Pr[p^N_{\mathsf{ref}} \leq 0].$$

Since, $p^N_{\mathsf{cndt}} \sim \mathcal{N}(\mu_{\mathsf{cndt}}, \sigma^2_{\mathsf{cndt}})$ and $p^N_{\mathsf{ref}} \sim \mathcal{N}(\mu_{\mathsf{ref}}, \sigma^2_{\mathsf{ref}})$

$$L_0(\mathcal{M}_{\mathsf{cndt}}) = \Phi\left(-\frac{\mu_{\mathsf{cndt}}}{\sigma_{\mathsf{cndt}}}\right).$$

Similarly,

$$L_0(\mathcal{M}_{\mathsf{ref}}) = \Phi\left(-\frac{\mu_{\mathsf{ref}}}{\sigma_{\mathsf{ref}}}\right).$$

Consequently,

$$\begin{aligned} L_0(\mathcal{M}_{\mathsf{ref}}) - L_0(\mathcal{M}_{\mathsf{cndt}}) &\geq \Phi\left(-\frac{\mu_{\mathsf{ref}}}{\sigma_{\mathsf{ref}}}\right) - \Phi\left(-\frac{\mu_{\mathsf{cndt}}}{\sigma_{\mathsf{cndt}}}\right) \\ &= C_2. \end{aligned}$$

Given that it has been statistically verified with confidence at least $1-\alpha$

$$\frac{\mu_{\mathsf{cndt}}}{\sigma_{\mathsf{cndt}}} > \frac{\mu_{\mathsf{ref}}}{\sigma_{\mathsf{ref}}}.$$

Since $\Phi(\cdot)$ is strictly increasing,

$$\Phi\left(-\frac{\mu_{\mathsf{ref}}}{\sigma_{\mathsf{ref}}}\right) > \Phi\left(-\frac{\mu_{\mathsf{cndt}}}{\sigma_{\mathsf{cndt}}}\right),$$

and therefore

$$C_2 > 0.$$

Thus it shows that, up to the Gaussian-approximation errors, the population classification loss of $\mathcal{M}_{\mathsf{cndt}}$ is strictly smaller than that of $\mathcal{M}_{\mathsf{ref}}$.

We now compare the generalization gaps. By definition,

$$\begin{aligned}
&\mathrm{GenGap}(\mathcal{M}_{\mathsf{ref}}) - \mathrm{GenGap}(\mathcal{M}_{\mathsf{cndt}}) \\
&= \Big(L_0(\mathcal{M}_{\mathsf{ref}}) - \widehat{L}_{0,T_{\mathsf{ref}}}(\mathcal{M}_{\mathsf{ref}})\Big) - \Big(L_0(\mathcal{M}_{\mathsf{cndt}}) - \widehat{L}_{0,T_{\mathsf{cndt}}}(\mathcal{M}_{\mathsf{cndt}})\Big) \\
&= L_0(\mathcal{M}_{\mathsf{ref}}) - L_0(\mathcal{M}_{\mathsf{cndt}}) - (\widehat{L}_{0,T_{\mathsf{ref}}}(\mathcal{M}_{\mathsf{ref}}) - \widehat{L}_{0,T_{\mathsf{cndt}}}(\mathcal{M}_{\mathsf{cndt}})).
\end{aligned}$$

We know

$$\widehat{L}_{0,T_{\mathsf{ref}}}(\mathcal{M}_{\mathsf{ref}}) - \widehat{L}_{0,T_{\mathsf{cndt}}}(\mathcal{M}_{\mathsf{cndt}}) \le \epsilon.$$

Combining the inequalities, we obtain

$$\mathrm{GenGap}(\mathcal{M}_{\mathsf{ref}}) - \mathrm{GenGap}(\mathcal{M}_{\mathsf{cndt}}) \ge C_2 - \epsilon.$$

with confidence atleast $1-\alpha$ □

A.8 Proof of Lemma 3

Proof. If $\mathrm{GenGap}(\mathcal{M}) \ge \tau$, Then $L_0(\mathcal{M}) \ge \widehat{L}_{0,\mathcal{D}_{\text{train}}}(\mathcal{M}) + \tau$.

From Bartlett et al. (2017) we get $L_0(\mathcal{M}) \le \widehat{L}_{\gamma,\mathcal{D}_{\text{train}}}(\mathcal{M}) + C\frac{\|X\|_2 R_{\mathcal{W}}}{\gamma n}\ln(W) + c_\delta\sqrt{\frac{\log(1/\delta)}{n}}$. where $\|X\|_2 = \sqrt{\sum_{i=1}^n \|x_i\|_2^2}$

Given, $\widehat{L}_{\gamma,\mathcal{D}_{\text{train}}}(\mathcal{M}) \le \widehat{L}_{0,\mathcal{D}_{\text{train}}}(\mathcal{M}) + \beta_\gamma$, we get $L_0(\mathcal{M}) \le \widehat{L}_{0,\mathcal{D}_{\text{train}}}(\mathcal{M}) + \beta_\gamma + C\frac{\|X\|_2 R_{\mathcal{W}}}{\gamma n}\ln(W) + c_\delta\sqrt{\frac{\log(1/\delta)}{n}}$.

Combining the equations together, yields $\widehat{L}_{0,\mathcal{D}_{\text{train}}}(\mathcal{M}) + \tau \le \widehat{L}_{0,\mathcal{D}_{\text{train}}}(\mathcal{M}) + \beta_\gamma + C\frac{\|X\|_2 R_{\mathcal{W}}}{\gamma n}\ln(W) + c_\delta\sqrt{\frac{\log(1/\delta)}{n}}$. Canceling $\widehat{L}_{0,\mathcal{D}_{\text{train}}}(\mathcal{M})$ and rearranging gives

$$C\frac{\|X\|_2 R_{\mathcal{W}}}{\gamma n}\ln(W) \ge \tau - \beta_\gamma - c_\delta\sqrt{\frac{\log(1/\delta)}{n}}.$$

Therefore,

$$R_{\mathcal{W}} \ge \frac{\gamma n}{C\|X\|_2 \ln(W)}\left[\tau - \beta_\gamma - c_\delta\sqrt{\frac{\log(1/\delta)}{n}}\right]_+.$$

Finally, if $\|x_i\|_2 \le B$ for every i, then

$$\|X\|_2 = \sqrt{\sum_{i=1}^n \|x_i\|_2^2} \le B\sqrt{n}.$$

Substituting this into the preceding inequality gives

$$R_{\mathcal{W}} \ge \frac{\gamma\sqrt{n}}{CB\ln(W)}\left[\tau - \beta_\gamma - c_\delta\sqrt{\frac{\log(1/\delta)}{n}}\right]_+.$$

This proves the claim. □

A.9 Proof of Positive Definiteness test is enough in our setting

Lemma 7 (Efficient spectral-complexity verification). *Let $W \in \mathbb{R}^{m\times n}$ and define*

$$A(\lambda) := \lambda^2 I - W^\top W.$$

Then

$$A(\lambda) \succ 0 \quad \Longleftrightarrow \quad \lambda > \|W\|_\sigma.$$

Hence, by Sylvester's criterion, checking positivity of the n leading principal minors of $A(\lambda)$ is sufficient to certify that λ is a strict upper bound on $\|W\|_\sigma$. For an L-layer network, let

$$s_i := \|W_i\|_\sigma, \qquad b_i := \|W_i^\top\|_{2,1},$$

and

$$R_W = \left(\prod_{i=1}^{L} s_i\right)\left(\sum_{i=1}^{L} \frac{b_i^{2/3}}{s_i^{2/3}}\right)^{3/2}.$$

If $\bar{s}_i > s_i$ is certified as above for every layer and $\bar{R}_W$ is defined by replacing s_i with $\bar{s}_i$, then

$$\bar{R}_W \geq R_W.$$

Moreover, replacing R_W with $\bar{R}_W$ preserves the validity of the generalization bound and leaves the mean-to-standard-deviation audit statistic unchanged.

Proof. Let $\sigma_1 \geq \cdots \geq \sigma_n \geq 0$ be the singular values of W. Since

$$W^\top W = Q\,\mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2)Q^\top,$$

we have

$$A(\lambda) = Q\,\mathrm{diag}(\lambda^2 - \sigma_1^2, \ldots, \lambda^2 - \sigma_n^2)Q^\top.$$

Therefore,

$$A(\lambda) \succ 0 \iff \lambda^2 > \sigma_1^2 \iff \lambda > \|W\|_\sigma.$$

As $A(\lambda)$ is symmetric, Sylvester's criterion implies that this can be verified using only its leading principal minors. It remains to show that upper bounding each layer norm upper bounds the full spectral complexity. Raising R_W to the power $2/3$ gives

$$R_W^{2/3} = \sum_{i=1}^{L} b_i^{2/3} \prod_{j\neq i} s_j^{2/3}.$$

Thus, if $\bar{s}_j \geq s_j$ for every j,

$$\bar{R}_W^{2/3} = \sum_{i=1}^{L} b_i^{2/3} \prod_{j\neq i} \bar{s}_j^{2/3} \geq R_W^{2/3},$$

and hence

$$\bar{R}_W \geq R_W.$$

Therefore, substituting $\bar{R}_W$ into the generalization bound only increases its complexity term and yields a conservative, valid bound. For the audit statistic, the normalized margins become

$$\bar{p}_i^N = \frac{p_i}{B\bar{R}_W} = c\,p_i^N, \qquad c := \frac{R_W}{\bar{R}_W} > 0.$$

Hence

$$\bar{\mu} = c\mu, \qquad \bar{\sigma} = c\sigma,$$

so

$$\frac{\bar{\mu}}{\bar{\sigma}} = \frac{\mu}{\sigma}.$$

Thus the final audit decision and the associated statistical test are identical to those obtained using the exact spectral complexity. Finally, any lower bound $R_W \geq R_{\min}$ for an overfitted model also implies

$$\bar{R}_W \geq R_W \geq R_{\min},$$

so the use of an upper-bound certificate does not weaken the anti-gaming guarantee. □

A.10 Proof of Condition 1

Proof. We prove the claim step by step.

Step 1: Condition on the statistically validated mean separation. By Condition 1, with the prescribed statistical confidence,

$$\mu_{\mathsf{cndt}} - \mu_{\mathsf{ref}} \geq d. \tag{1}$$

Equivalently,

$$\mu_{\mathsf{cndt}} \geq \mu_{\mathsf{ref}} + d. \tag{2}$$

We condition on this event throughout the remainder of the proof.

Choose any

$$0 < \tau < d$$

and define

$$\gamma := \tau B R_{\mathcal{W}} > 0. \tag{3}$$

Since $BR_{\mathcal{W}} > 0$,

$$p_{\mathsf{cndt}}(x, y) \leq \gamma \quad \Longleftrightarrow \quad p^{N}_{\mathsf{cndt}}(x, y) \leq \tau.$$

Step 2: Establish variance-independent Gaussian CDF bounds. Define Define

$$\sigma_- := \min\{\sigma_{\mathsf{cndt}}, \sigma_{\mathsf{ref}}\}, \qquad \sigma_+ := \max\{\sigma_{\mathsf{cndt}}, \sigma_{\mathsf{ref}}\}.$$

For $a \in \mathbb{R}$, define the lower and upper Gaussian CDF envelopes over the variance interval $[\sigma_-, \sigma_+]$ as

$$\underline{\Phi}(a) := \begin{cases} \Phi\left(\dfrac{a}{\sigma_+}\right), & a \geq 0, \\ \Phi\left(\dfrac{a}{\sigma_-}\right), & a < 0, \end{cases} \tag{4}$$

and

$$\overline{\Phi}(a) := \begin{cases} \Phi\left(\dfrac{a}{\sigma_-}\right), & a \geq 0, \\ \Phi\left(\dfrac{a}{\sigma_+}\right), & a < 0, \end{cases} \tag{5}$$

where Φ is the standard normal CDF.

Now for fixed $a \in \mathbb{R}$, consider

$$g_a(\sigma) := \Phi\left(\frac{a}{\sigma}\right), \qquad \sigma > 0.$$

If $a \geq 0$, then a/σ decreases as σ increases. Therefore

$$\Phi\left(\frac{a}{\sigma_+}\right) \leq \Phi\left(\frac{a}{\sigma}\right) \leq \Phi\left(\frac{a}{\sigma_-}\right).$$

If $a < 0$, then a/σ increases as σ increases. Therefore

$$\Phi\left(\frac{a}{\sigma_-}\right) \leq \Phi\left(\frac{a}{\sigma}\right) \leq \Phi\left(\frac{a}{\sigma_+}\right).$$

Thus, for every

$$\sigma \in [\sigma_-, \sigma_+],$$

we have

$$\underline{\Phi}(a) \leq \Phi\left(\frac{a}{\sigma}\right) \leq \overline{\Phi}(a). \tag{6}$$

This bound does not require knowing whether σ_{cndt} or σ_{ref} is larger.

Step 3: Lower bound the true classification error of the reference model. Since $BR_{\mathcal{W}_{\mathsf{ref}}} > 0$, normalization does not alter the sign of the margin. Therefore

$$L_0(\mathcal{M}_{\mathsf{ref}}) = \Pr[p_{\mathsf{ref}}(X,Y) \le 0] = \Pr[p^N_{\mathsf{ref}}(X,Y) \le 0].$$

By assumption,

$$p^N_{\mathsf{ref}} \sim \mathcal{N}(\mu_{\mathsf{ref}}, \sigma^2_{\mathsf{ref}}),$$

and hence

$$L_0(\mathcal{M}_{\mathsf{ref}}) = \Phi\left(-\frac{\mu_{\mathsf{ref}}}{\sigma_{\mathsf{ref}}}\right). \tag{7}$$

Since

$$\sigma_{\mathsf{ref}} \in [\sigma_-, \sigma_+],$$

the lower envelope in equation 6, applied with $a = -\mu_{\mathsf{ref}}$, gives

$$\boxed{L_0(\mathcal{M}_{\mathsf{ref}}) \ge \underline{\Phi}(-\mu_{\mathsf{ref}}).} \tag{8}$$

Step 4: Upper bound the positive-margin probability of the candidate model. Let

$$F_{\mathsf{cndt}}(t) := \Pr[p^N_{\mathsf{cndt}}(X,Y) \le t]$$

denote the population CDF of the normalized margins of $\mathcal{M}_{\mathsf{cndt}}$.

Since

$$p^N_{\mathsf{cndt}} \sim \mathcal{N}(\mu_{\mathsf{cndt}}, \sigma^2_{\mathsf{cndt}}),$$

we have

$$F_{\mathsf{cndt}}(\tau) = \Phi\left(\frac{\tau - \mu_{\mathsf{cndt}}}{\sigma_{\mathsf{cndt}}}\right). \tag{9}$$

By equation 2,

$$\mu_{\mathsf{cndt}} \ge \mu_{\mathsf{ref}} + d.$$

Therefore

$$\tau - \mu_{\mathsf{cndt}} \le \tau - \mu_{\mathsf{ref}} - d. \tag{10}$$

Since Φ is increasing,

$$F_{\mathsf{cndt}}(\tau) \le \Phi\left(\frac{\tau - \mu_{\mathsf{ref}} - d}{\sigma_{\mathsf{cndt}}}\right).$$

Using the upper envelope from equation 6, with

$$a = \tau - \mu_{\mathsf{ref}} - d,$$

we obtain

$$\boxed{F_{\mathsf{cndt}}(\tau) \le \overline{\Phi}(\tau - \mu_{\mathsf{ref}} - d).} \tag{11}$$

Combining equation 8 and equation 11 gives

$$\begin{aligned} L_0(\mathcal{M}_{\mathsf{ref}}) - F_{\mathsf{cndt}}(\tau) &\ge \underline{\Phi}(-\mu_{\mathsf{ref}}) - \overline{\Phi}(\tau - \mu_{\mathsf{ref}} - d) \\ &= \psi_{d,\tau}. \end{aligned} \tag{12}$$

Thus,

$$\boxed{L_0(\mathcal{M}_{\mathsf{ref}}) - F_{\mathsf{cndt}}(\tau) \ge \psi_{d,\tau}.} \tag{13}$$

Step 5: Relate the population margin probability to the empirical margin loss. Define

$$\widehat{F}_{\mathsf{cndt}}(t) := \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}\left[p^N_{\mathsf{cndt}}(x_i, y_i) \le t\right].$$

Further, let

$$h_m := \sqrt{\frac{\ln(2/\beta)}{2m}},$$

Then by the Dvoretzky–Kiefer–Wolfowitz inequality, with the corresponding probability greater than $1-\beta$,

$$\sup_t \left|\widehat{F}_{\mathsf{cndt}}(t) - F_{\mathsf{cndt}}(t)\right| \le h_m. \tag{14}$$

In particular,

$$\widehat{F}_{\mathsf{cndt}}(\tau) \le F_{\mathsf{cndt}}(\tau) + h_m. \tag{15}$$

Step 6: Apply Eq. 1 to the candidate model. We now invoke Eq. 1 using the margin parameter

$$\gamma = \tau B R_{\mathcal{W}}.$$

With the corresponding high probability,

$$L_0(\mathcal{M}_{\mathsf{cndt}}) \le \widehat{L}_\gamma(\mathcal{M}_{\mathsf{cndt}}) + \kappa_{\tau,\delta}, \tag{16}$$

where $\kappa_{\tau,\delta}$ denotes the residual complexity term after setting $\gamma = \tau B R_{\mathcal{W}}$.

Since

$$p_{\mathsf{cndt}}(x_i, y_i) \le \gamma \quad \Longleftrightarrow \quad p^N_{\mathsf{cndt}}(x_i, y_i) \le \tau,$$

we have

$$\widehat{L}_\gamma(\mathcal{M}_{\mathsf{cndt}}) = \widehat{F}_{\mathsf{cndt}}(\tau).$$

Hence,

$$L_0(\mathcal{M}_{\mathsf{cndt}}) \le \widehat{F}_{\mathsf{cndt}}(\tau) + \kappa_{\tau,\delta}. \tag{17}$$

Using equation 15,

$$L_0(\mathcal{M}_{\mathsf{cndt}}) \le F_{\mathsf{cndt}}(\tau) + h_m + \kappa_{\tau,\delta}.$$

Combining this with equation 13 gives

$$\begin{aligned} L_0(\mathcal{M}_{\mathsf{ref}}) - L_0(\mathcal{M}_{\mathsf{cndt}}) &\ge \psi_{d,\tau} - h_m - \kappa_{\tau,\delta} \\ &= C_1(d,\tau). \end{aligned} \tag{18}$$

Here, $\kappa_{\tau,\delta} := \widetilde{O}\left(\frac{\|X_V\|_2}{m\tau}\ln W_M + \sqrt{\frac{\ln(1/\delta)}{m}}\right)$.

Therefore,

$$\boxed{L_0(\mathcal{M}_{\mathsf{ref}}) - L_0(\mathcal{M}_{\mathsf{cndt}}) \ge C_1(d,\tau).} \tag{19}$$

Step 7: Compare the generalization gaps. By definition,

$$\begin{aligned} &\mathrm{GenGap}(\mathcal{M}_{\mathsf{ref}}) - \mathrm{GenGap}(\mathcal{M}_{\mathsf{cndt}}) \\ &= L_0(\mathcal{M}_{\mathsf{ref}}) - L_0(\mathcal{M}_{\mathsf{cndt}}) \\ &\quad - \left(\widehat{L}_{0,T_{\mathsf{ref}}}(\mathcal{M}_{\mathsf{ref}}) - \widehat{L}_{0,T_{\mathsf{cndt}}}(\mathcal{M}_{\mathsf{cndt}})\right). \end{aligned} \tag{20}$$

By assumption,

$$\widehat{L}_{0,T_{\mathsf{ref}}}(\mathcal{M}_{\mathsf{ref}}) - \widehat{L}_{0,T_{\mathsf{cndt}}}(\mathcal{M}_{\mathsf{cndt}}) \le \epsilon.$$

Consequently,

$$\boxed{\mathrm{GenGap}(\mathcal{M}_{\mathsf{ref}}) - \mathrm{GenGap}(\mathcal{M}_{\mathsf{cndt}}) \ge C_1(d,\tau) - \epsilon.} \tag{21}$$

□

A.11 Proof of Theorem 2

Proof. We first consider Condition 1. The statistically certified separation

$$\mu_{\mathsf{cndt}} - \mu_{\mathsf{ref}} \geq d$$

compares the models in complexity-normalized coordinates. For any $0 < \tau < d$, choose

$$\gamma = \tau B R_{\mathcal{W}}.$$

The positive-margin event appearing in Theorem 1 is expressed in exactly the same normalized coordinates used by Condition 1. Applying the normalized-mean analysis together with Theorem 1 yields (as seen App. A.10)

$$L_0(\mathcal{M}_{\mathsf{ref}}) - L_0(\mathcal{M}_{\mathsf{cndt}}) \geq C_1(d, \tau). \tag{22}$$

We next consider Condition 2. From App. A.7 we get,

$$L_0(\mathcal{M}_{\mathsf{ref}}) - L_0(\mathcal{M}_{\mathsf{cndt}}) = C_2. \tag{23}$$

Equations equation 22 and equation 23 give two valid certificates for the same population-error difference. Hence

$$L_0(\mathcal{M}_{\mathsf{ref}}) - L_0(\mathcal{M}_{\mathsf{cndt}}) \geq \max\left\{C_1(d, \tau), C_2\right\}.$$

Finally,

$$\begin{aligned} &\mathrm{GenGap}(\mathcal{M}_{\mathsf{ref}}) - \mathrm{GenGap}(\mathcal{M}_{\mathsf{cndt}}) \\ &= L_0(\mathcal{M}_{\mathsf{ref}}) - L_0(\mathcal{M}_{\mathsf{cndt}}) \\ &\quad - \left(\widehat{L}_{0,T_{\mathsf{ref}}}(\mathcal{M}_{\mathsf{ref}}) - \widehat{L}_{0,T_{\mathsf{cndt}}}(\mathcal{M}_{\mathsf{cndt}})\right). \end{aligned} \tag{24}$$

Applying the equations gives

$$\mathrm{GenGap}(\mathcal{M}_{\mathsf{ref}}) - \mathrm{GenGap}(\mathcal{M}_{\mathsf{cndt}}) \geq \max\left\{C_1(d, \tau), C_2\right\} - \epsilon,$$

which proves the claim. □

A.12 Evaluation Cntd.

A.12.1 Model Training and Margin Bound Evaluations

To rigorously evaluate the capability of our metric to detect overfitting, we trained two distinct variants for each architecture: a *Well-fit model* (trained for optimal generalization) and an *Overfit model*.

To induce overfitting, we performed an ablation of standard regularization techniques. Specifically, we disabled data augmentation (random flips, crops, and noise injection) and removed weight decay and dropout terms from the training schedulers. These models were trained until we observed a distinct divergence between training and validation loss, or until the training accuracy saturated at 100%.

A.12.2 Zero-Knowledge Proof System

Zero-Knowledge Proof of Inference
We utilized the Zero-Knowledge proof of inference protocol developed by Chen et al. (2024). To demonstrate the scalability of our approach, we leveraged the inherent parallelism of the inference task.

To obtain the performance data in Table 2, we used an AWS `u7i-6tb.112xlarge` instance (448 vCPUs, 6 TB RAM). This experiment demonstrate the parallelization capability of our scheme while also showing, utilizing multiple CPU cores further reduces latency.

Zero-Knowledge Proof of Spectral Complexity
To generate normalized margin bounds, our scheme requires a Zero-Knowledge proof of the model's

Table 3: Given the reference model and the statistical tests, we compute the True Positive Rate (TPR) i.e., out of all independently trained well-fitted models, ratio of them it correctly identifies and True Negative Rate (TNR) i.e., out of all independently trained overfitted models, ratio of them it correctly identifies

Model	Dataset	Well-fit models		Overfit models	
		Num. models	TPR (in %)	Num. models	TNR (in %)
2-layer NN	Adult	100	99.0	100	99.0
2-layer NN	Credit	100	95.0	100	98.0
2-layer NN	German	78	60.3	50	82.0
Simple CNN	FMNIST	167	60.0	200	62.5
LeNet5	CIFAR10	100	65.0	100	100.0
ResNet18	CIFAR10	140	94.3	260	60.0

spectral complexities. Importantly, this component is data-agnostic and dependent solely on the model weights, incurring only a one-time computational overhead.

We employed the GNARK library Botrel et al. (2025) to implement ZK proofs for the determinant and norm of the weight matrices. We utilized the full capacity of the machine to benchmark these operations.

The total proving and verification time for the *ProoFit* is calculated as the sum of the runtime for the ZK proofs of inference and the ZK proof of spectral complexity. Similarly, the total proof size is the sum of the spectral complexity proof and the aggregate inference proofs.

Baseline: Zero-Knowledge Proof of Training For our baseline comparison of CNNs, we utilized the implementation of *KAIZEN* Abbaszadeh et al. (2024). Unlike our approach, this method provides a proof of the training process rather than post-training inference.

The results reported in Table 1 represent the lower bound of the average projected runtime for a single training process of 100 epochs with a batch size of 4 and a SHA commitment level of 4. We utilized the same AWS `u7i-6tb.112xlarge` instance for these benchmarks. To ensure compatibility with the *KAIZEN* implementation, we reduced the kernel sizes of the neural network architectures to match the expected input formats.

Since batch iterations are inherently sequential, parallelization across batches is not possible. Consequently, the total runtime for the training process is calculated as:

$$\text{Total Time} = \#\,\text{Epochs} \times \left(\frac{\text{Dataset Size}}{\text{Batch Size}}\right) \times \text{Time per Iteration} \tag{25}$$

For full training, this cost scales linearly with the number of epochs. While the proof size and verification time remain constant regardless of the number of iterations (KAIZEN achieves this via incrementally verifiable composition, IVC, that keeps the proof succinct), the proving time becomes the bottleneck.

For the fully connected neural networks, we implemented the baseline ourselves using GNARK Botrel et al. (2025). In this case, we recorded the performance for a single iteration of 512 batch size and presented the projected numbers using equation 25. Since this implementation doesn't have the succintness of an IVC, here the verification time and proof size also scale linearly with the number of iterations and epochs following equation 25.

As indicated in Table 1, the proving time for ZK proof of training is prohibitive for the entire training process. Given that modern deep learning models typically require hundreds to thousands of epochs to converge, this approach remains impractical for our use case, highlighting the efficiency and superiority of our audit scheme.

A.13 Security Analysis

Theorem 4.

Completeness
$\forall D_{audit} \in \mathcal{X}^m, \mathcal{W}$ [‡]

[‡]Candidate model weights

Table 4: For the overfitted models of each architecture and dataset pairs, we compute $\tau = \mathrm{GenGap}(\mathcal{M})$ and empirically find the max. value of γ such that $\beta_\gamma \approx 0$

Model	Dataset	τ	γ	β_γ
2-layer NN	Adult	0.15	2.0	0.06
2-layer NN	Credit	0.20	1.0	0.06
2-layer NN	German	0.25	2.5	0.06
Simple CNN	FMNIST	0.12	20.0	0.02
LeNet5	CIFAR10	0.40	1.5	0.07
ResNet18	CIFAR10	0.12	6.3	0.01

$$\Pr\left[\begin{array}{l} \mathsf{pp} \leftarrow \mathit{ProoFit}.\mathsf{Setup}(1^\lambda) \\ (pk, vk) \leftarrow \mathit{ProoFit}.\mathsf{KeyGen}(\mathsf{pp}) \\ \mathsf{com}_{\mathcal{W}} \leftarrow \mathit{ProoFit}.\mathsf{Commit}(\mathcal{W}, \mathsf{pp}, r) \\ (\rho = (\{p_i\}, R_{\mathcal{W}}), \pi_\rho) \\ \quad \leftarrow \mathit{ProoFit}.\mathsf{Prove}(\mathcal{W}, \mathsf{com}_{\mathcal{W}}, r, D_{\mathit{audit}}, pk) \\ \mathit{ProoFit}.\mathsf{Verify}(\mathsf{com}_{\mathcal{W}}, D_{\mathit{audit}}, \rho, \pi_\rho, \mathsf{pp}) = 1 \end{array}\right] = 1 \tag{26}$$

Proof Sketch: The completeness proof mostly follows from the completeness of the underlying proof system, which in our case are ZKML Chen et al. (2024) and GNARK Botrel et al. (2025).

Theorem 5. *There exists an extractor $\mathcal{E}$ such that for all probabilistic polynomial time provers $\mathcal{P}$*

Knowledge Soundness
$\forall D_{\mathit{audit}}, \mathbf{W}$

$$\Pr\left[\begin{array}{l} \mathsf{pp} \leftarrow \mathit{ProoFit}.\mathsf{Setup}(1^\lambda) \\ (pk, vk) \leftarrow \mathit{ProoFit}.\mathsf{KeyGen}(\mathsf{pp}) \\ (\mathsf{com}_{\mathbf{W}^*}, \rho^*, \pi^*_\rho) \leftarrow \mathcal{P}(1^\lambda, pk) \\ (\mathcal{W}, \rho = (\{p_i\}, R_{\mathcal{W}})) \leftarrow \mathcal{E}^P(\cdots) \\ \mathit{ProoFit}.\mathsf{Verify}(\mathsf{com}_{\mathbf{W}}, vk, D_{\mathit{audit}}, \rho, \pi_\rho) = 1 \\ \exists x_i \in D_{\mathit{audit}}, \\ \mathsf{margin}(\mathcal{W}, x_i) \neq p_i \vee R_{\mathcal{W}} \neq \mathsf{SpectComp}(\mathcal{W}) \end{array}\right] < negl(\lambda) \tag{27}$$

Proof Sketch: Knowledge-soundness follows directly from the knowledge-soundness of the underlying proof system. The extractor runs the corresponding extractor and outputs the witness.

Theorem 6.

Zero-Knowledge *We say a protocol Π is zero-knowledge if there exists a polynomial time, randomized simulator S such that for all $(pk, vk) = \mathsf{Setup(pp)}$, for all $(x, w) \in \mathcal{R}$, for all verifiers V*

$$\{\mathcal{P}(pk, x, w)\} \sim \{S(pk, x)\}$$

Proof Sketch: The zero knowledge property also follows directly from that of the underlying proof system, which in our case are ZKML and GNARK. Specifically, we use the simulators of the respective proof systems and the relation $\mathcal{R}$ here is given by $\{x, p, R_{\mathcal{W}}; \mathcal{W}\}$ where $p = \mathsf{margin}(\mathcal{W}, x), R_{\mathcal{W}} = \mathsf{SpectComp}(\mathcal{W})$

A.14 Related Work

ML Generalization Bounds

The study of generalization in machine learning, particularly for over-parameterized deep neural networks, has evolved from classical complexity measures to more data-dependent and structure-dependent bounds.

A.14.1 Complexity Measures and Margin Bounds

Early learning theory relied heavily on the Vapnik-Chervonenkis (VC) dimension Vapnik & Chervonenkis (1971) to bound the generalization gap. While foundational, VC-dimension bounds often yield vacuous results for modern deep neural networks where the number of parameters significantly exceeds the number of training samples Bartlett et al. (2019). To address this, margin-based bounds were introduced, notably for Support Vector Machines, showing that classifiers with large margins could generalize well independent of the input dimension Boser et al. (1992); Bartlett et al. (1998).

In the context of neural networks, Bartlett Bartlett (1998; 1996) pioneered the idea that the magnitude of the weights is more critical for generalization than the sheer number of parameters. This shift towards norm-based capacity control has been further explored in various settings. Zhang Zhang (2002) provided covering number bounds for regularized linear function classes, and Neyshabur et al. Neyshabur et al. (2015) investigated norm-based capacity control, suggesting that standard norms (e.g., Euclidean) might not be sufficient to explain the generalization of deep networks without accounting for the network structure.

A.14.2 The PAC-Bayesian Framework

A powerful alternative to uniform convergence bounds is the PAC-Bayesian framework, originally introduced by McAllester McAllester (1998; 1999). Unlike VC bounds which hold for the worst-case predictor in a hypothesis class, PAC-Bayesian bounds provide guarantees for randomized predictors (interpretable as a posterior distribution) based on a chosen prior. Langford et al. Langford & Caruana (2001); Langford & Shawe-Taylor (2002) demonstrated that this framework could yield much tighter, non-vacuous bounds for neural networks and margin classifiers by optimizing the error bound directly.

More recently, Dziugaite et al. Dziugaite & Roy (2017) successfully computed non-vacuous numerical generalization bounds for deep networks by optimizing the PAC-Bayes bound with respect to the posterior variance, effectively linking generalization to the "flatness" of the minimum found by stochastic gradient descent (SGD).

A.14.3 Spectral Norms and Modern Generalization Bounds

Building on these foundations, recent work has focused on deriving bounds that explicitly account for the spectral properties of the network's weight matrices. Bartlett et al. Bartlett et al. (2017) derived a spectrally-normalized margin bound for deep neural networks, showing that the generalization gap scales with the product of the spectral norms of the weight matrices and the $\ell_{2,1}$ norm of the weight matrices, rather than the total number of parameters.

Contemporaneously, Neyshabur et al. Neyshabur et al. (2017b) applied the PAC-Bayesian framework to derive similar spectrally-normalized bounds, reinforcing the connection between the spectral complexity of the network and its generalization performance. Our work builds directly upon these two approaches Bartlett et al. (2017); Neyshabur et al. (2017b), leveraging their normalized bounds to construct Zero-Knowledge proofs of overfitting.

A.14.4 Sharpness, Sensitivity, and Optimization

The relationship between the geometry of the loss landscape and generalization has also been extensively studied. Keskar et al. Keskar et al. (2016) empirically observed that large-batch training tends to converge to "sharp" minima which generalize poorly, while small-batch methods find "flat" minima with better generalization. Neyshabur et al. Neyshabur et al. (2017a) further explored this, investigating how different complexity measures—including sharpness and norms—correlate with generalization.

Novak et al. Novak et al. (2018) introduced sensitivity-based metrics, showing that the Frobenius norm of the input-output Jacobian correlates strongly with generalization gap. For specific architectures like Convolutional Neural Networks (CNNs), Zhou and Feng Zhou & Feng (2018) provided theoretical guarantees linking optimization to generalization performance.

Additionally, technical tools from matrix concentration inequalities, such as those by Tropp Tropp (2012), and consistency results for multiclass classification Tewari & Bartlett (2007); Zhang (2004) have been instrumental in formalizing these high-dimensional probability bounds.

CONFIDENTIAL AUDITS

Our work builds upon a rapidly growing body of literature at the intersection of Zero-Knowledge Proofs (ZKPs) and Machine Learning (ML). We categorize related works into three primary streams: scalable architectures for verifiable inference, protocols for verifiable training, and the cryptographic verification of specific model properties such as fairness.

A.14.5 SCALABLE VERIFIABLE INFERENCE

The application of ZKPs to machine learning allows a Prover to convince a Verifier that a computation $y = M(x)$ was performed correctly without revealing the model M or input x. Early approaches like *SafetyNets* Ghodsi et al. (2017) utilized interactive proofs tailored for quadratic activations, while subsequent frameworks like *Mystique* Weng et al. (2021b) and *Wolverine* Weng et al. (2021a) introduced efficient conversions between arithmetic and Boolean shares to handle standard non-linearities (e.g., ReLU) efficiently.

Recent works have moved toward automated compilation and optimization to scale these proofs to production-grade models. Kang et al. Kang et al. (2022) achieved a significant milestone by presenting the first ZK-SNARK for ImageNet-scale inference (specifically MobileNet v2), leveraging the Halo2 proving system and lookup arguments to manage deep network operations.

Building on the trend of automated compilation, Chen et al. (2024) introduced *ZKML*, an end-to-end optimizing compiler that transforms high-level model specifications directly into optimized ZK circuits. Similarly, Chen et al. Chen et al. (2025) proposed *ZKTorch*, a compiler that converts PyTorch models into cryptographic “basic blocks". Unlike monolithic approaches, *ZKTorch* utilizes a parallel proof accumulation scheme (based on Mira) to efficiently compose these blocks, addressing the scalability bottlenecks of large constraint systems.

Parallel efforts have focused on modularity and specific architectural optimizations. Balbás et al. (2023) introduced a framework based on sumcheck protocols, proposing a *Verifiable Evaluation Scheme* to efficiently handle sequential ML pipelines. Specifically for Convolutional Neural Networks (CNNs), Liu et al. (2021) introduced *zkCNN*, which leverages the sumcheck protocol to verify convolution operations and polynomial activations with high efficiency. Building on this, Qu et al. (2025) proposed *VerfCNN*, a specialized framework that optimizes prover complexity for multi-channel convolutions. By aligning the proof generation cost with the input/output size rather than the computation size, *VerfCNN* significantly reduces overhead compared to prior sumcheck-based approaches.

Further in the domain of CNNs, Lee et al. (2024) proposed *vCNN*, which introduces a novel relation representation to reduce the proving complexity of convolutions from $O(l \cdot n)$ to $O(l + n)$. Complementing this, Weng et al. (2023) proposed *pvCNN*, which combines Homomorphic Encryption (HE) with ZK-SNARKs to outsource linear layers while verifying convolutions via optimized quadratic matrix programs.

A.14.6 VERIFIABLE TRAINING AND PROOF-OF-LEARNING

While verifiable inference assumes a fixed model, verifying the *training* process itself is considerably more challenging. Zhao et al. (2021) introduced *VeriML*, a framework enabling “ML as a Service" providers to prove they executed the training pipeline correctly. However, proving every step of Stochastic Gradient Descent (SGD) is often prohibitively expensive. This difficulty was formalized by Garg et al. (2023), who provided a comprehensive baseline for zero-knowledge proofs of training (zkPoT). They demonstrated that while cryptographically verifying the training of simple models like logistic regression is feasible, scaling to deep neural networks remains computationally impractical without significant optimizations to the underlying proof systems.

To address this, Abbaszadeh et al. (2024) proposed *KAIZEN*, a zero-knowledge proof of training (zkPoT) optimized for DNNs. KAIZEN utilizes a recursive GKR-style proof system to verify iterative gradient descent steps without the proof size growing linearly with the number of iterations.

A.14.7 VERIFIABLE MODEL PROPERTIES

Fairness: Several works have operationalized fairness checks using ZKPs. Shamsabadi et al. (2022) introduced *Confidential-PROFITT* to prove that decision trees satisfy fairness definitions like Demographic Parity by bounding information gain on sensitive attributes. Yadav et al. (2024) proposed *FairProof*, a system for certifying the fairness of neural networks post-training, while Zhang et al. (2025b) introduced *FairZK*. *FairZK* innovates by measuring fairness bounds using only model parameters and aggregated input statistics, avoiding the need to reveal the training dataset. For dynamic settings, Franzese et al. (2025) proposed *OATH* to provide certificates of "online fairness," ensuring that models remain fair regarding the stream of data received during deployment.

Explainability: Trust can also be established by proving that a model's explanations are faithful. Yadav et al. (2025) proposed *ExpProof*, which operationalizes explanations in adversarial settings. They utilize ZKPs to prove that an explanation (generated via methods like LIME) is indeed derived correctly from the concealed model, preventing service providers from manipulating explanations to hide bias.

Calibration and Generalization: Closest to our work is *Confidential Guardian* Rabanser et al. (2025), which uses ZK proofs to verify model calibration (e.g., Expected Calibration Error) to prevent the abuse of abstention mechanisms. Our work complements this landscape by focusing on *overfitting*. We utilize ZK proofs of inference — building on the efficient implementations provided by frameworks like *ZKML* Chen et al. (2024) — to verify normalized generalization bounds, providing a certificate that the model has not simply memorized its training data.