Shuffling is Not Enough: Breaking Permutation-Based Model Confidentiality in Hybrid FHE Inference

Jiseung Kim[1] and Hyung Tae Lee[2(✉)]

[1] Jeonbuk National University, jiseungkim@jbnu.ac.kr
[2] Chung-Ang University, hyungtaelee@cau.ac.kr

Abstract. Hybrid fully homomorphic encryption (FHE) inference improves the practicality of private inference by letting the server evaluate linear layers homomorphically while the client decrypts and applies nonlinearities. Recent schemes attempt to protect model confidentiality by returning noisy, output-permuted responses and appealing to shuffle-model differential privacy (DP). We show that this protection fails in the correctness regime required by hybrid FHE systems. For a d-input linear layer, $d + 1$ admissible queries suffice for exact recovery of a permutation-invariant layer summary, hence for perfect model distinguishability. We further show that input DP is orthogonal to model confidentiality and that the local-DP premise required for shuffle amplification cannot hold under correctness-bounded noise. We recover all linear layers of a SAFHIRE-style ResNet-20 end-to-end from TFHE transcripts with zero error, using $d + 1$ queries per layer for a total of 5,712 direct queries. Under the same query model, we also confirm exact per-layer recovery on pretrained ImageNet-scale CNNs and ViT-B/16. The leaked spectra enable fingerprinting, lineage attribution, and improved logit-based extraction, while suppressing them destroys inference utility.

Keywords: Homomorphic encryption · Private inference · Model confidentiality · Structural cryptanalysis · Differential privacy.

1 Introduction

Private inference is increasingly expected to protect both sides of the interaction: the client's input should remain hidden from the server, and the server's model should remain hidden from the client. Fully homomorphic encryption (FHE) offers a principled route to input privacy, but evaluating modern neural networks entirely under FHE remains expensive in latency and bandwidth. Hybrid FHE inference sidesteps this bottleneck by partitioning the computation: the server evaluates each linear layer homomorphically, while the client decrypts, applies the nonlinearity in the clear, and re-encrypts for the next layer. Recent systems such as SAFHIRE [2] demonstrate this architecture on real convolution and transformer backbones.

However, the hybrid spirit reintroduces a threat that pure FHE had eliminated: because the client observes intermediate layer outputs in the clear, a curious client can in principle read off the server's weight directly. To close this gap without reverting to the cost of pure FHE, a line of recent work, such as Safhire [2], STIP [28], Centaur [16], and PermLLM [30], proposes permutation-based model confidentiality. For example, in Safhire, the server evaluates a linear layer, applies a secret permutation to the output coordinates, and returns the resulting ciphertext to the client; since ReLU commutes with coordinate permutation, the shuffled representation can be maintained across rounds. The resulting confidentiality intuition is appealing: FHE decryption noise acts as a local randomizer, the secret output permutation acts as a shuffler, and shuffle-model differential privacy (DP) [1,9] is invoked to argue that only limited model information is exposed.

This paper shows that the intuition is false in the very regime required for correct hybrid FHE inference. We study a broad class of permutation-based protection schemes in which the client may issue chosen inputs and observe a noisy permuted response to a linear layer on a known quantization lattice. Our starting point is the structural observation: if the quantization step is γ, correct decryption requires the additive noise to remain below $\gamma/2$ in magnitude. Under exactly this bound, response-side randomness can be removed by rounding, and permutation no longer protects the layer beyond relabeling.

We formalize this observation and show that, for a d-input linear layer, $d + 1$ admissible queries suffice for exact recovery of a permutation-invariant layer summary, hence for perfect model distinguishability. In the first round of Safhire [2], where no hidden input permutation is applied, the same attack recovers the ordered family itself. Since Safhire's threat model permits arbitrary chosen inputs without plaintext validation or well-formness proofs, these queries are admissible in practice. This immediately implies that any two challenge models with different sorted spectrum multisets are perfectly distinguishable in the model confidentiality game. This result is distribution-agnostic: it depends only on correctness-bounded noise and known quantization, not on the choice of permutation strategy or on any DP parameter attached to the scheme.

Our analysis also clarifies why the shuffle DP justification does not establish model confidentiality. Input DP and model confidentiality quantify different objects: the former compares neighboring inputs under a fixed model, whereas the latter compares different models under fixed queries. More fundamentally, the local-DP premise required for shuffle amplification cannot hold under correctness-bounded noise. Whenever two admissible inputs induce outputs that differ by at least one lattice step in some coordinate, the corresponding per-coordinate supports are disjoint once the noise remains below $\gamma/2$, forcing the local-DP slack parameter to become vacuous. Thus, the privacy amplification argument collapses before it can be applied.

We validate our analysis on TFHE transcripts [6] and ImageNet-scale models. For a Safhire-style ResNet-20, exact recovery holds for all 21 linear layers using $d + 1$ queries per layer for a total of 5,712. The same zero-error behav-

ior persists across eight pre-trained ImageNet-scale CNNs and ViT-B/16 under practical precisions and bounded-noise laws. The resulting leakage enables fingerprinting, lineage attribution, and improved logit-based extraction, while no viable privacy/utility operating points remain once the noise is increased enough to suppress the attack.

In summary, our results make three contributions. First, we provide an exact and query-tight recovery theorem for permutation-based protection under correctness-bounded noise. Second, we show that input DP is orthogonal to model confidentiality in this setting and that the local-DP premise required by shuffle amplification is itself unattainable. Finally, we demonstrate concrete security impact and a sharp privacy/utility dilemma on deployed TFHE transcripts and ImageNet-scale vision models.

1.1 Related Work

Model extraction. Classical extraction attacks [23,19,12] typically require many oracle queries because nonlinearities such as softmax obscure the linear layers. For example, the cryptanalytic attack of Carlini et al. [3] uses $2^{18.5}$–$2^{21.5}$ queries on 4K–100K-parameter MLPs. By contrast, our attack needs only $d+1$ queries per layer, totaling 5,712 on a 0.27M-parameter ResNet-20. The gap arises because permutation-based protection directly exposes a sorted linear-layer snapshot, a leakage channel already known to be severe in encrypted databases [18]. The link between extraction and fingerprinting [15], together with membership inference [21], further supports treating identification-level leakage as a model-confidentiality threat.

Impossibility and attacks on ML defenses. Conditional impossibility results are known for black-box observational defenses of polynomial-size decision trees under subexponential learning parity with noise (LPN) [14] and for learnability-preserving data encoding [26]. Our result is incomparable, giving an *unconditional* guarantee of exact sorted-spectrum recovery on the narrower channel of Definition 1. ArrowMatch [25] empirically shows that TEE-shielded weight obfuscation alone fails against adaptive clients.

Attacks on cryptographic inference stacks. Prior attacks are empirical. SEEK [5] extracts approximate weights from an 11.7M-parameter ResNet-18 over homomorphic encryption/multiparty computation (MPC) using fewer than 50 queries per parameter, about 5×10^8 in total, with relative error about 3×10^{-4}. Its leverage is arbitrary intermediate-feature shifts rather than a permuted linear snapshot, so it does not target Definition 1. Works such as [4,10] attack pure-FHE and split-model FHE LLM inference outside the permutation class.

Attacks on permutation defenses and DP limits. Thomas et al. [22] use sequential-token markers to recover client *input* tokens in PermLLM, STIP, and Centaur, which is disjoint from our chosen-input attack on *model* parameters. The limitations of DP as a *model*-protection, rather than training-data, primitive are discussed in [7,11,27]; Proposition 3 quantifies this limitation in the hybrid-FHE regime.

2 Background

We briefly review the SAFHIRE setting and the shuffle-model DP argument used by permutation-based protection schemes. We need only the quantized plaintext lattice, the correctness condition, the per-round permutation interface, and the chosen-input threat model, not the full algebraic machinery of TFHE [6].

2.1 Hybrid FHE Inference via Safhire

We illustrate hybrid FHE inference through SAFHIRE [2], the primary instantiation of Definition 1.

RLWE encryption (torus framing). SAFHIRE follows the TFHE torus convention [6]: plaintexts lie on $\mathcal{T}_p := \frac{1}{p}\mathbb{Z}/\mathbb{Z}$, where typically $p = 2^b$ for $b \in \{8, 12, 16\}$. Thus the relevant plaintext values lie on a lattice of spacing $\gamma = 1/p$. A ring learning with error (RLWE) ciphertext with additive noise e decrypts correctly when $|e| < 1/(2p)$. We therefore write $\gamma = 1/p$ for the quantization step and $\Delta < \gamma/2$ for the correctness-bounded noise magnitude. This bound is the key operational condition in our analysis: it implies that decoded layer outputs are observed on a known lattice and perturbed by noise smaller than half a lattice step.

Batched protocol. Inference proceeds in L rounds, one per linear layer. In round r, the server first removes the previous round permutation by applying σ_{r-1}^{-1} to the incoming activation, homomorphically evaluates $y = W_r x + b_r$, applies the round-r output permutation σ_r, and returns the resulting ciphertext to the client. The client decrypts and applies coordinate-wise ReLU before re-encrypting the activation for the next round. Because ReLU commutes with coordinate permutation, the shuffled representation can be maintained across rounds. As in SAFHIRE, the final layer is left unshuffled so that the client can read the argmax directly. The permutations are derived from a secret session seed, fixed within a session and re-randomized across sessions [2, Section 4.4].

Threat model. We adopt SAFHIRE's threat model [2, Section 3]. The server is honest-but-curious, whereas the client may send arbitrary inputs and may attempt to reverse-engineer the model. No distributional restriction, plaintext validation, or zero-knowledge proof of well-formedness is imposed on client queries. In particular, the input encoding packs a flattened feature chunk into RLWE plaintext coefficients [2, Section 4.3], so the basis inputs $\{0, e_1, \ldots, e_d\}$ used in our attack are admissible. These properties are precisely the ingredients exploited later: the client observes a shuffled, noisy linear-layer response on a known quantization lattice and may query it adaptively.

2.2 Shuffle Model of Differential Privacy

A randomized mechanism M satisfies (ε, δ)-differential privacy [8] if $\Pr[M(D) \in S] \leq e^{\varepsilon} \Pr[M(D') \in S] + \delta$ for every neighboring pair D, D' and every measurable set S. In the shuffle model [1,9], n local randomizers first privatize their

inputs, after which a trusted shuffler applies a uniform random permutation to the resulting messages. If each local randomizer satisfies $(\varepsilon_0, \delta_0)$-local DP and the parameters lie in the appropriate regime, the shuffled mechanism enjoys an amplified central-DP guarantee, with privacy improving roughly as $O(\varepsilon_0/\sqrt{n})$.

Permutation-based defenses in our class instantiate this template by treating FHE decryption noise as the local randomization step and the secret output-channel permutation as the shuffler. The claim is that, because the client observes only a noisy permuted response rather than the ordered vector $Wx + b$, shuffling amplifies the privacy contributed by the noise. Our analysis in Section 3 examines this claim in the hybrid-FHE setting. The central question is whether the correctness-bounded noise required for reliable decryption can also satisfy the local-DP premise needed for shuffle amplification.

3 Breaking Permutation-Based Model Protection

In this section, we show that permutation-based protection is incompatible with hybrid FHE inference under the correctness-bound noise regime. We first formalize the scheme class and prove exact sorted-spectrum recovery from $d+1$ adaptive chosen queries. We then show that shuffle-DP does not ensure model confidentiality, and that increasing the noise enough to suppress the attack quickly destroys utility. We begin with notation used throughout the paper.

Notation. For $v \in \mathbb{R}^m$, let $\mathrm{sort}(v)$ denote the vector obtained by arranging its entries in non-decreasing order. For a linear layer $W \in \mathbb{R}^{m\times d}$ and $b \in \mathbb{R}^m$, define $s_i(W, b) := \mathrm{sort}(W_{:,i} + b)$ as the *sorted column spectrum* of the i-th column $W_{:,i}$, and let $L(W, b) := (s_1(W, b), \ldots, s_d(W, b))$ denote the resulting ordered family. By construction, $L(W, b)$ is precisely the information preserved under an output-channel permutation.

3.1 Permutation-Based Protection Schemes

Definition 1 (Permutation-Based Protection). *Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a plaintext query space containing $\{0, e_1, \ldots, e_d\}$. A* permutation-based protection scheme *for a linear layer $W \in \mathbb{R}^{m\times d}$ and $b \in \mathbb{R}^m$ is a pair of algorithms $\Pi = (\mathsf{Query}, \mathsf{Respond})$ such that:*

- $\mathsf{Query}(x)$*: on input $x \in \mathcal{X}$, the client produces a well-formed query object q.*
- $\mathsf{Respond}(W, b, q)$*: on input q, the server determines an effective input $x^\star \in \mathbb{R}^d$ and returns a response from which the client obtains $\widetilde{y} = \pi(Wx^\star + b) + \eta$, where $\pi \in S_m$ is a secret output permutation chosen according to the scheme's permutation strategy and $\|\eta\|_\infty < \Delta$.*
  If the scheme uses a hidden input permutation, then $x^\star = \rho(x)$ for some secret $\rho \in S_d$; otherwise $x^\star = x$ (i.e., $\rho = \mathrm{id}$).

We say that Π has quantization step γ *if, for every $x \in \{0, e_1, \ldots, e_d\}$ and the corresponding effective input $x^\star$, the vector $Wx^\star + b$ lies on a lattice of minimum spacing γ.*

Remark 1. The following remarks clarify the operational meaning of Definition 1 and its instantiation in FHE-based protocols.

- Definition 1 subsumes SAFHIRE round-r interface by taking $x^\star = \sigma_{r-1}^{-1}(x)$ and $\pi = \sigma_r$, with $\sigma_0 = \mathrm{id}$ in the first round. The special case $x^\star = x$ recovers the output-shuffling-only interface.
- The client may issue adaptive chosen queries $x^{(1)}, x^{(2)}, \ldots \in \mathcal{X}$, obtain $q^{(t)} \leftarrow \mathsf{Query}(x^{(t)})$, and observe the corresponding decoded responses $\widetilde{y}^{(t)}$.
- The zero query ensures that b lies on the same lattice as the basis responses. If the relevant values lie in $\frac{1}{p}\mathbb{Z}$, then one may take $\gamma = 1/p$.
- In FHE instantiations, $\mathsf{Query}(x)$ is the client-side encryption of x, and $\mathsf{Respond}$ returns a ciphertext whose decryption yields $\widetilde{y}$. The basis queries are admissible whenever arbitrary client activations are permitted and no validation mechanism rules them out.

Definition 1 abstracts a common pattern in which the client observes a permuted, noisy linear-layer response and may issue chosen queries against it. While motivated here by hybrid FHE inference, the definition itself is not specific to FHE. SAFHIRE's first layer is the clearest instantiation: the server evaluates a linear layer, applies a secret output permutation, and returns the result under correctness-bounded noise. Its later layers additionally apply a server-side input unshuffle, which we analyze separately in Section 4.1. Section 6 examine Fission [24], STIP [28], Centaur [16], and PermLLM [30] with respect to the hypotheses of Definition 1. Among them, Fission is MPC-based rather than FHE-based, and is included only for comparison at the level of hypotheses rather than as a direct instantiation of the definition. A scheme that augments client queries with zero-knowledge well-formedness proofs would violate hypothesis (C) and therefore fall outside the definition; none of the schemes considered here does so.

Theorem 1 (Exact Sorted-Spectrum Recovery). *Let Π be a permutation-based protection scheme under Definition 1, with noise bound Δ and quantization step $\gamma \geq 2\Delta$. Conditioned on $\|\eta\|_\infty < \Delta$, there exists an efficient adversary $\mathcal{A}$ that, using $d{+}1$ adaptive queries, recovers* $\mathrm{sort}(b)$ *from the zero query and the sorted-spectrum multiset* $\{\{\mathrm{sort}(W_{:,j}+b)\}\}_{j=1}^{d}$ *from the d basis queries $e_1, \ldots, e_d$. The recovery is exact on every correctly recovered noise realization.*

Proof. Let Q_γ denote rounding to the nearest γ-lattice point, and define the round-then-sort operator $\mathsf{RTS}(\tilde{y}) := \mathrm{sort}(Q_\gamma(\tilde{y}))$.

Bias recovery. For the zero query $x = 0$, the response satisfies $\tilde{y} = \pi(b) + \eta$. Since b lies on the γ-lattice and $\gamma \geq 2\Delta$, we have $|\eta_j| < \Delta \leq \gamma/2$ for every coordinate j. Hence rounding recovers $\pi(b)$ exactly, and sorting yields $\mathrm{sort}(b)$.

Spectrum recovery. For each $i \in [d]$, query $x = e_i$. By Definition 1, the effective input is $x^\star = \rho(e_i) = e_{\rho(i)}$, so the server computes $Wx^\star + b = W_{:,\rho(i)} + b$, and the observed response is $\tilde{y} = \pi(W_{:,\rho(i)} + b) + \eta$. The same round-then-sort argument recovers $\mathrm{sort}(W_{:,\rho(i)} + b)$ exactly. As i ranges over $[d]$ and $\rho \in S_d$ is a bijection, the adversary recovers the multiset $\{\{\mathrm{sort}(W_{:,j} + b) : j \in [d]\}\}$. The output is therefore fully determined by W and b on every noise realization satisfying $\|\eta\|_\infty < \Delta$. □

Theorem 1 gives the general recovery guarantee under a hidden input permutation ρ. When $\rho = \mathrm{id}$, as in the first round of SAFHIRE, the recovered spectra are additionally labeled by the queried column indices, so the adversary obtains the ordered family $L(W, b)$ rather than merely its underlying multiset. More generally, the guarantee is distribution-agnostic. The recovered output is a deterministic function of W and b, since all response-side randomness is eliminated by rounding. Accordingly, the guarantee depends only on the bound $\|\eta\|_\infty < \Delta$ and is therefore independent of the output-permutation strategy, the noise law, and any DP parameter attached to the scheme. In particular, a DP parameter ε constrains the ratio of output distributions across neighboring inputs for a fixed model, and thus does not affect any quantity used by the attack.

The hypothesis $\gamma \geq 2\Delta$ in Theorem 1 is also tight. Indeed, if $\Delta > \gamma/2$, then no deterministic algorithm can guarantee exact recovery over all noise realizations; see Proposition 6 in Appendix B.

Proposition 1 (Query Complexity Tightness). *Let Π be a permutation-based protection scheme under Definition 1 with $b = 0$. Then any adaptive adversary that recovers $\{\{\mathrm{sort}(W_{:,j})\}\}_{j=1}^{d}$ on every instance requires at least d queries. Together with Theorem 1, this gives exact minimax query complexity: d for bias-free layers, and $d+1$ when $\mathrm{sort}(b)$ must also be recovered.*

Proof. Fix any adaptive strategy using $k \leq d-1$ queries $x^{(1)}, \ldots, x^{(k)} \in \mathbb{R}^d$. The subspace $V = \mathrm{span}\{x^{(1)}, \ldots, x^{(k)}\}$ has dimension at most $k \leq d-1 < d$, so there exists a unit vector $v \in \mathbb{R}^d$ with $v \perp x^{(t)}$ for all $t \in [k]$. Define $W' := W + uv^T$ for any nonzero $u \in \mathbb{R}^m$. For every query $x^{(t)}$, $W'x^{(t)} = Wx^{(t)} + u\underbrace{(v^T x^{(t)})}_{=0} = Wx^{(t)}$, so $(W', 0)$ and $(W, 0)$ produce identical responses under any permutation strategy and noise realization. The adversary's view is identical, hence it outputs the same answer for both models. Since $v \neq 0$, there exists $j^* \in [d]$ with $v_{j^*} \neq 0$, so $W'_{:,j^*} = W_{:,j^*} + uv_{j^*}$. For generic W and u, this column is distinct from every column of W, so the sorted-spectrum multisets differ. The adversary therefore errs on at least one of $(W, 0)$ and $(W', 0)$. □

Corollary 1. *The first round of SAFHIRE, which is directly queried by the client, satisfies the hypotheses of Theorem 1. At precision p, it has $\gamma = 1/p$ and $\Delta = 1/(2p)$, so $\gamma = 2\Delta$, and the condition $\|\eta\|_\infty < \gamma/2$ is exactly the FHE correctness condition. Under standard lattice-based FHE parameters, this condition fails only with negligible probability, typically below 2^{-128} [2]. Hence, over at most $\mathrm{poly}(\lambda)$ attack queries, exact sorted-spectrum recovery holds with probability $1 - \mathsf{negl}(\lambda)$.*

3.2 Input DP and Model Confidentiality Are Orthogonal

We begin by emphasizing that input DP and model confidentiality are distinct notions. For a fixed model W, input DP bounds the ratio of output distributions induced by two neighboring inputs x and x'. Model confidentiality, by

contrast, concerns the distinguishability of two different models under a fixed input. The two quantifications are therefore logically independent. Nevertheless, prior permutation-based proposals invoke input DP for model confidentiality.

Definition 2. *The adversary $\mathcal{A}$ selects challenge models (W_0, b_0) and (W_1, b_1) whose sorted-spectrum multisets differ, i.e., $\{\{s_i(W_0, b_0) : i \in [d]\}\} \neq \{\{s_i(W_1, b_1) : i \in [d]\}\}$, and submits them to the challenger. The challenger samples $\beta \xleftarrow{\$} \{0, 1\}$ and grants $\mathcal{A}$ oracle access to $\Pi.\mathsf{Respond}(W_\beta, b_\beta, \cdot)$. $\mathcal{A}$ then issues adaptive queries and outputs a bit $\hat{\beta}$. We define the model confidentiality advantage of $\mathcal{A}$ by $\mathsf{Adv}^{\mathrm{mc}}_{\Pi,\mathcal{A}} := |2\Pr[\hat{\beta} = \beta] - 1|$.*

Proposition 2. *Let Π be a permutation-based protection scheme under Definition 1 with $\gamma \geq 2\Delta$. Then, there exists an adversary $\mathcal{A}$ issuing d chosen queries such that $\mathsf{Adv}^{\mathrm{mc}}_{\Pi,\mathcal{A}} = 1$, regardless of any input DP guarantee satisfied by $\Pi.\mathsf{Respond}(\cdot\,; W)$.*

Proof. The adversary $\mathcal{A}$ issues the d basis queries $e_1, \ldots, e_d$. By Theorem 1, the round-then-sort procedure recovers $\mathrm{sort}(W_{\beta,:,\rho(i)} + b_\beta)$ for each $i \in [d]$, and hence the d responses reveal the multiset $\mathcal{M}_\beta := \{\{\mathrm{sort}(W_{\beta,:,j} + b_\beta)\}\}_{j=1}^{d}$. Since ρ is a bijection on $[d]$, this multiset is recovered regardless of the hidden permutation ρ. By Definition 2, $\mathcal{M}_0 \neq \mathcal{M}_1$ as multisets, so $\mathcal{A}$ outputs the unique bit β whose multiset matches the observation and $\mathsf{Adv}^{\mathrm{mc}}_{\Pi,\mathcal{A}} = 1$. By contrast, input DP constrains single-model ratios of the form $\Pr[f(x; W) \in S] / \Pr[f(x'; W) \in S]$, which are orthogonal to the two-model comparison required in Definition 2. □

3.3 The Shuffle DP Premise Cannot Hold

Shuffle amplification is conditional on an $(\varepsilon_0, \delta_0)$-local DP input. Under the FHE correctness bound, each local mechanism moves its output by less than $\gamma/2$, so neighboring inputs whose outputs differ by one lattice step produce disjoint per-coordinate supports. The non-overlapping region captures the *entire* support of one side, driving δ_0 to 1 and voiding any downstream amplification.

Proposition 3. *Let $M_k(x) = (Wx + b)_k + \eta_k$ denote the per-coordinate output of a permutation-based scheme in Definition 1, and assume that each noise coefficient η_k lies almost surely in the interval $(-B, B)$, i.e., in the correctness-conditioned bounded-noise regime with $B < \gamma/2$. Let x, x' be inputs such that, for some coordinate k, the sensitivity satisfies $\Delta f := |(Wx + b)_k - (Wx' + b)_k| \geq \gamma$, where Δf denotes DP sensitivity and is unrelated to the noise bound Δ. Then any $(\varepsilon_0, \delta_0)$-DP guarantee for M_k on the pair (x, x') with finite ε_0 necessarily satisfies $\delta_0 \geq 1$. Equivalently, M_k cannot satisfy $(\varepsilon_0, \delta_0)$-DP for any $\delta_0 < 1$.*

Proof. Let $I_x := \mathrm{supp}(M_k(x))$ and $I_{x'} := \mathrm{supp}(M_k(x'))$. Write $c_x := (Wx + b)_k$ and $c_{x'} := (Wx' + b)_k$. Since $\eta_k \in (-B, B)$ almost surely, we have $I_x \subseteq [c_x - B,\, c_x + B]$ and $I_{x'} \subseteq [c_{x'} - B,\, c_{x'} + B]$, where we emphasize that these are containments, not necessarily equalities. Now consider the set $S := I_x \setminus I_{x'}$. By construction, S contains exactly those outputs that are attainable under

x but impossible under x'. Applying the $(\varepsilon_0, \delta_0)$-DP condition to the event S yields $\Pr[M_k(x) \in S] \leq e^{\varepsilon_0} \Pr[M_k(x') \in S] + \delta_0$. Since $S \cap I_{x'} = \emptyset$, we have $\Pr[M_k(x') \in S] = 0$, and therefore $\Pr[M_k(x) \in S] \leq \delta_0$. Under the bound $B < \gamma/2$, any coordinate gap $\Delta f \geq \gamma$ implies $\Delta f > 2B$. Hence the closed intervals $[c_x - B,\, c_x + B]$ and $[c_{x'} - B,\, c_{x'} + B]$ are disjoint and separated by a positive gap. It follows that $I_x \cap I_{x'} = \emptyset$, in fact $S = I_x$. So, $\Pr[M_k(x) \in S] = \Pr[M_k(x) \in I_x] = 1$. Combining with the previous inequality gives $1 \leq \delta_0$. □

The above proof uses only the almost-sure support bound $\eta_k \in (-B, B)$. In particular, it does not require a density or absolute continuity, and therefore applies equally to discretely supported RLWE noise and to almost-surely bounded continuous surrogates.

Remark 2 (Beyond $(\varepsilon_0, \delta_0)$-DP). Proposition 3 assumes almost-surely bounded noise $\eta_k \in (-B, B)$ with $B < \gamma/2$, i.e., the correctness-conditioned regime. This covers practical discrete-Gaussian RLWE noise as well as clipped continuous surrogates. By contrast, an unbounded Gaussian falls outside the hypothesis, since then $I_x = \mathbb{R}$ and the disjoint-support argument fails; moreover, any Gaussian large enough to overlap adjacent lattice points already violates $|\eta| < \gamma/2$ with non-negligible probability. Within the bounded-noise regime, disjoint supports imply total variation $\mathrm{TV}(M_k(x), M_k(x')) = 1$, so any divergence-based privacy notion that upper-bounds total variation, including standard relaxations of $(\varepsilon_0, \delta_0)$-DP, becomes infinite or vacuous on such neighbors. The same obstruction extends to group privacy, since a single lattice-step gap $\Delta f \geq \gamma$ already gives $\Delta f > 2B$.

In the Safhire setting, a per-coordinate gap of γ is unavoidable. For the basis queries used in Theorem 1, moving from $x = 0$ to $x = e_i$ changes the k-th output coordinate by the entry of $W_{:,i}$, which lies on the quantization lattice of spacing γ. Hence, any scheme with at least one nonzero weight entry admits neighboring inputs to which Proposition 3 applies. Since $\delta_0 \geq 1$ renders the local-DP premise of shuffle amplification trivially impossible to satisfy, no choice of shuffle size n, target δ, or Gaussian-surrogate noise can restore the guarantee.

One could try to restrict the shuffle-amplification analysis to neighbors with $\Delta f < \gamma$, thereby placing them outside the scope of Proposition 3. In Safhire, however, this restriction is unavailable: its threat model [2, Section 3] allows the client to issue arbitrary inputs, with no bound on $\|x\|$, so the basis queries of Theorem 1 are admissible and the vacuity result applies directly. Even a modified variant that artificially restricted sensitivity would recover only an input-privacy guarantee, which remains orthogonal to model confidentiality by Proposition 2.

3.4 Noise/Utility Tradeoff

Theorem 1 rules out confidentiality under the correctness bound $B < \gamma/2$. A natural remedy is to relax this bound and inject larger noise, trading decryption accuracy for defense. We establish a distribution-free lower bound on the noise required to defeat a Lipschitz-fingerprint attack, instantiated in Section 5.

Proposition 4. *Let Π be a permutation-based protection scheme under Definition 1 with quantization step γ, and let Q_γ denote rounding to the nearest γ-lattice point. Assume a relaxed noise bound $B > 0$, not necessarily with $B \le \gamma/2$, and define $\tilde{s}_i := \mathrm{sort}(Q_\gamma(\pi(W_{:,i} + b) + \eta))$. Then, for every column i and every noise realization with $\|\eta\|_\infty < B$,*

$$\|\tilde{s}_i - s_i(W, b)\|_\infty \le e_B := \gamma \left\lfloor \tfrac{B}{\gamma} + \tfrac{1}{2} \right\rfloor \le B + \tfrac{\gamma}{2}.$$

Proof. Fix a column i and write $v := \pi(W_{:,i} + b) \in \gamma\mathbb{Z}^m$. Since $\|\eta\|_\infty < B$, each coordinate satisfies $|\eta_j| < B$, and thus $|Q_\gamma(v_j + \eta_j) - v_j| \le \gamma \left\lfloor \frac{B}{\gamma} + \frac{1}{2} \right\rfloor = e_B$. Therefore, $\|Q_\gamma(v + \eta) - v\|_\infty \le e_B$. Since sort is invariant under the permutation π, and order statistics are non-expansive in ℓ_∞,

$$\|\tilde{s}_i - s_i(W, b)\|_\infty = \|\mathrm{sort}(Q_\gamma(v + \eta)) - \mathrm{sort}(v)\|_\infty \le \|Q_\gamma(v + \eta) - v\|_\infty \le e_B.$$

Moreover, since $\lfloor \frac{B}{\gamma} + \frac{1}{2} \rfloor \le \frac{B}{\gamma} + \frac{1}{2}$, we have $e_B \le \gamma\left(\frac{B}{\gamma} + \frac{1}{2}\right) = B + \frac{\gamma}{2}$. □

Parameterizing the bound by the Lipschitz constant of the fingerprint map makes it directly applicable to the attacks in Section 5.

Theorem 2 (Noise/Utility Tradeoff). *Let Π be a permutation-based protection scheme under Definition 1 with relaxed noise bound $B > 0$, let $\mathcal{F}$ be a finite model family, and let $\phi\colon L \mapsto \phi(L) \in (\mathbb{R}^k, \|\cdot\|_q)$ be a fingerprint map satisfying*

$$\|\phi(L) - \phi(L')\|_q \le L_\phi \cdot \max_i \|s_i - s'_i\|_\infty \qquad \text{for some } L_\phi > 0. \tag{1}$$

Define $\Delta^\phi_{\mathcal{F}} := \min_{M \ne M' \in \mathcal{F}} \|\phi(L(M)) - \phi(L(M'))\|_q$, and let $\mathcal{A}_\phi$ denote the round-then-sort nearest-neighbor classifier under ϕ given the $d+1$ chosen queries $x = 0, e_1, \ldots, e_d$. Then the followings hold, independently of the noise distribution: (i) if $L_\phi e_B < \Delta^\phi_{\mathcal{F}}/2$, then $\mathcal{A}_\phi$ correctly identifies every $M \in \mathcal{F}$ on every noise realization satisfying $\|\eta\|_\infty < B$. (ii) Conversely, defeating $\mathcal{A}_\phi$ on any $M \in \mathcal{F}$ requires $B \ge \max\{0,\ \Delta^\phi_{\mathcal{F}}/(2L_\phi) - \gamma/2\}$.

Proof. Fix the true model $M = (W, b) \in \mathcal{F}$ with sorted spectra (s_i), and let $(\tilde{s}_i)$ be the recovered vectors. By Proposition 4, $\max_i \|\tilde{s}_i - s_i\|_\infty \le e_B$. Write $\tilde{L} := (\tilde{s}_i)$ and $L := L(M)$. Applying (1) to $\tilde{L}$ and L gives $\|\phi(\tilde{L}) - \phi(L)\|_q \le L_\phi e_B$.

(i) Assume $L_\phi e_B < \Delta^\phi_{\mathcal{F}}/2$. For any $M' \ne M$ in $\mathcal{F}$, the reverse triangle inequality and the definition of $\Delta^\phi_{\mathcal{F}}$ yield

$$\begin{aligned}\|\phi(\tilde{L}) - \phi(L(M'))\|_q &\ge \|\phi(L) - \phi(L(M'))\|_q - \|\phi(\tilde{L}) - \phi(L)\|_q \\ &\ge \Delta^\phi_{\mathcal{F}} - L_\phi e_B > L_\phi e_B \ge \|\phi(\tilde{L}) - \phi(L)\|_q.\end{aligned}$$

Hence $\mathcal{A}_\phi$ uniquely returns M. *(ii)* This is immediate by contraposition. If $\mathcal{A}_\phi$ fails on some $M \in \mathcal{F}$, then $L_\phi e_B \ge \Delta^\phi_{\mathcal{F}}/2$, and therefore $e_B \ge \frac{\Delta^\phi_{\mathcal{F}}}{2L_\phi}$. By Proposition 4, $e_B \le B + \frac{\gamma}{2}$. Combining the two inequalities gives $B + \frac{\gamma}{2} \ge e_B \ge \frac{\Delta^\phi_{\mathcal{F}}}{2L_\phi}$, hence $B \ge \frac{\Delta^\phi_{\mathcal{F}}}{2L_\phi} - \frac{\gamma}{2}$. Taking the maximum with 0 removes the vacuous case in which the right-hand side is negative. □

Remark 3. Section 5 uses the column-norm fingerprint $\phi(L) := \mathrm{sort}(\|s_1\|_2, \ldots, \|s_d\|_2)$ with ℓ_2 nearest-neighbor. This map satisfies Eq. (1) with $L_\phi = \sqrt{md}$. Indeed, if $\max_i \|\tilde{s}_i - s_i\|_\infty \le e$, then the reverse triangle inequality and $\|\cdot\|_2 \le \sqrt{m}\|\cdot\|_\infty$ give $|\|\tilde{s}_i\|_2 - \|s_i\|_2| \le \sqrt{m}\, e$ for each i. Since sort is non-expansive in ℓ_∞ and $\|\cdot\|_2 \le \sqrt{d}\|\cdot\|_\infty$ on $\mathbb{R}^d$, it follows that $\|\phi(\tilde{L}) - \phi(L)\|_2 \le \sqrt{md}\, e$. Thus Eq. (1) holds with $L_\phi = \sqrt{md}$. For ResNet-20 `conv1`, where $m = 16$ and $d = 27$, at 8-bit precision with $\Delta_{\mathcal{F}}^{\phi} \approx 0.298$, Theorem 2 yields $B \ge 2.7(\gamma/2)$. This bound is loose and remains well below the empirical transition in Table 3.

Theorem 2 certifies attack success below the threshold, not defender safety above it. The certified bound $2.7(\gamma/2)$ is two to three orders of magnitude below the empirical transition at 10^2–$10^3(\gamma/2)$ in Table 3, because the worst-case bound $L_\phi = \sqrt{md}$ assumes adversarial alignment across all m coordinates, whereas symmetric noise gives a tighter $\sqrt{m}$ scaling. Together, Theorem 1 covers $B < \gamma/2$, $\delta_0 \ge 1$ rules out the shuffle route, and Theorem 2 ties brute-force noise to intrinsic fingerprint separation.

4 Chosen-Input Attack

In this section, we instantiate the recovery attack against SAFHIRE [2], the primary hybrid-FHE instantiation of Definition 1.

4.1 Multi-Round Layers

In SAFHIRE's multi-round protocol, the server applies σ_{r-1}^{-1} before evaluating layer r and σ_r afterward. Accordingly, a basis query e_i issued in round r probes column $\sigma_{r-1}^{-1}(i)$ of W_r. Since σ_{r-1} remains fixed within a session, the d basis queries in round r collectively probe every column of W_r exactly once. Moreover, SAFHIRE's threat model explicitly permits chosen-input queries [2, Section 3], and the basis vectors are admissible plaintext polynomials under its coefficient encoding [2, Section 4.3].

Proposition 5. *Under the hypotheses of Theorem 1, if σ_{r-1} is sampled independently for each inference request, then all distinct sorted spectra of W_r can be recovered using $O(d \log d)$ queries with probability at least $1 - e^{-c}$. More precisely, recovery succeeds if the number T of queries satisfies $T \ge d(\ln d^* + c)$, where $d^* \le d$ denotes the number of distinct spectra.*

Proof. Each query probes a uniformly random column. If a distinct spectrum s appears with multiplicity m_s, then a single query recovers s with probability $m_s/d \ge 1/d$. By Theorem 1, the queried spectrum is recovered exactly from each response, and duplicate recoveries can be removed by equality testing on the sorted vectors. Fix a distinct spectrum s. The probability that T independent queries all miss s is at most $(1 - 1/d)^T \le e^{-T/d}$. Applying a union bound over the d^* distinct spectra, the probability that some spectrum is missed is at most $d^* e^{-T/d}$. Thus, whenever $T \ge d(\ln d^* + c)$, the failure probability is at most $d^* e^{-T/d} \le d^* e^{-(\ln d^* + c)} = e^{-c}$. Therefore, recovery succeeds with probability at least $1 - e^{-c}$. □

In the within-session regime (described in Section 2.1), the d basis queries recover $L(W_r, b_r)$ within the overall d+1 query budget of Theorem 1. By contrast, when the permutation is resampled independently for each query, the query complexity becomes $\Theta(d \log d)$. On trained ResNet-20, we observe $d^* = d$ in every layer, i.e., no duplicate spectra arise. Consequently, the bound $T \geq d(\ln d + c)$ yields approximately 5,400 queries for the widest layer (d=576, c=3). In simulation, the independently-resampled-permutation attack converges with $T_{\text{emp}}/T_{\text{theory}} \in [0.58, 0.87]$ over 5 trials per layer.

What is recovered? For each layer the adversary recovers $L(W, b)$ and hence every permutation-invariant statistic derived from it, including the affine column norms $\|W_{:,i} + b\|_2$ and the empirical value distribution of each column. For bias-free layers, where $b = 0$ as is standard under batch normalization, this reduces to the exact weight multisets $\text{sort}(W_{:,i})$, their column norms, and the Frobenius norm $\|W\|_F$. Since sort does not commute with subtraction, the bias cannot be disentangled from the weights in general; however, this residual ambiguity does not affect any of the applications in Section 5.

Safhire [2, Section 5, scenario (iii)] treats leakage of only the weight histogram as benign residual leakage, purportedly bounded by a shuffle-model DP guarantee. Section 3 shows that this guarantee is vacuous: Proposition 2 establishes that the DP argument is formulated in the wrong direction, while Proposition 3 shows that its local-DP premise fails under the correctness bound. Section 5 further demonstrates that the recovered spectra suffice for reliable fingerprinting and lineage detection, and that suppressing these attacks requires noise at scales for which inference utility itself collapses (see Table 3 in Section 5).

4.2 Experimental Validation

We validate the attack at two levels of concreteness: a deployment-faithful TFHE transcript replay on Safhire's backend in Table 1, and simulated per-layer queries on eight ImageNet-pretrained architectures in Table 2. Our CIFAR-10 evaluation uses ResNet-20 with 21 bias-free convolutional layers after BN folding, 0.27M parameters, mean test accuracy 82.8%, and a per-model query budget of $\sum_{l=1}^{21}(d_l + 1) = 5{,}712$. We also confirm zero-error recovery for randomly initialized models, six bounded noise distributions, and MLPs with nonzero bias; see Appendix C. The codes for all experiments in this paper are available at `https://github.com/JiseungKim90/permutation-confidentiality-experiments`.

TFHE transcript replay. To rule out simulation artifacts, we replay the attack end-to-end through Concrete v2.11 [29], the TFHE backend of Safhire [6,2]. Table 1 reports the results across all 21 convolutional layers under both a fixed permutation and a fresh-per-query permutation.

Architecture-agnostic ImageNet validation. We apply the simulated attack to eight torchvision-pretrained ImageNet models, namely ResNet-50, ResNet-101, ResNet-152, VGG-16, DenseNet-121, MobileNet-V2, EfficientNet-B0, and ConvNeXt-Tiny. These models comprise 600 convolutional layers in total, with

Table 1. TFHE transcript on all 21 convolutional layers of a trained ResNet-20 (p=256, 11,424 encrypted queries total, 5,712 per regime). Max recovery error = 0 for every layer in both fixed-perm and fresh-perm regimes.

Shape	#layers	d+1	Total q	Shape	#layers	d+1	Total q
16×27	1	28	28	32×288	5	289	1,445
16×144	6	145	870	64×32	1	33	33
32×16	1	17	17	64×288	1	289	289
32×144	1	145	145	64×576	5	577	2,885
Total (21 layers)			5,712	Max error (all)			**0**

up to 58M parameters and a largest layer of size 2048×4608. Every layer exhibits zero recovery error at all three precision levels and under all six bounded noise distributions, as reported in Table 2. Applying the same pipeline to the 37 linear layers of torchvision ViT-B/16 with `IMAGENET1K_V1` weights, which contains 86M parameters, likewise yields max_err = 0 at $p = 256$, thereby extending the recovery to transformer architectures. These 37 layers are exactly the attention query/key/value and output projections and the MLP layers used in transformer blocks. Because the same linear primitives underlie language transformers, including the model targeted by PermLLM [30], our recovery applies without modification.

Table 2. Exact sorted-spectrum recovery on eight ImageNet-pretrained architectures in per-layer simulation. Max recovery error=0 for every convolutional layer at 4-, 8-, and 12-bit precision across all architectures. The largest layer has shape 2048×4608.

Architecture	Layers	Params	Architecture	Layers	Params
ResNet-50	53	23,454,912	ResNet-101	104	42,382,016
ResNet-152	155	57,992,384	VGG-16	13	14,710,464
DenseNet-121	120	6,870,208	MobileNet-V2	52	2,189,760
EfficientNet-B0	81	3,956,192	ConvNeXt-Tiny	22	1,877,472

5 Practical Impact of Sorted Spectrum Leakage

The defender's residual claim in Section 4 is that the loss of column labels renders the recovered multisets benign: even if $L(W, b)$ is exposed, the adversary still cannot reconstruct the weight matrix, even up to row and column permutations. This characterization, however, substantially understates the leakage. Many security-relevant quantities are permutation-invariant and hence are fully determined by $L(W, b)$. These include column norms, the per-column empirical distributions $s_i(W, b)$, and the weight-value histogram. This leakage has three concrete consequences: *fingerprinting*, *lineage detection* for fine-tuned descendants, and accelerated logit-based *extraction*. Moreover, suppressing any of these attacks requires noise at a scale that materially degrades inference utility, as shown in Section 5.3.

5.1 Attribution: Fingerprinting and Lineage

Even for a fixed architecture, independently trained models converge to different local minima under different random seeds. Their sorted column spectra are therefore deterministically distinct and can serve as reliable model identifiers. Using the 20 trained ResNet-20 checkpoints from Section 4, we define the *fingerprint* of a model as the sorted column-norm vector $\phi(W, b) := \mathrm{sort}(\|s_1\|_2, \ldots, \|s_d\|_2)$ of the `conv1` layer, which is the specialization described in Remark 3, and compare models by ℓ_2 distance in the fingerprint space. Since ϕ is invariant under column permutations, the multiset-level recovery guaranteed by Theorem 1, which is the general case under a secret $\rho \neq \mathrm{id}$, already suffices for both fingerprinting and lineage attribution; the stronger labeled recovery available when $\rho = \mathrm{id}$ is unnecessary. Since recovery is exact under the correctness bound, the within-model distance is zero across all noise realizations, whereas the minimum between-model distance is $\Delta^{\phi}_{\mathcal{F}} = 0.298$. Consequently, all $\binom{20}{2} = 190$ model pairs are perfectly separable under ϕ. On CIFAR-100, the same fingerprint yields $\Delta^{\phi}_{\mathcal{F}} = 0.565$ after 200 training epochs, and all 190 pairs remain perfectly separable. Each such pair also instantiates the model confidentiality game of Definition 2 empirically. In particular, the round-then-sort adversary of Proposition 2 achieves $\mathsf{Adv}^{\mathrm{mc}} = 1$ on every one of the 190 challenge pairs using the d basis queries of Theorem 1, or a single query in the first-round setting where $\rho = \mathrm{id}$.

Lineage detection. Fine-tuning typically perturbs the weights incrementally, so the sorted column spectrum of a fine-tuned child remains closer to that of its parent than to that of an independently trained checkpoint. To quantify this effect, we select 5 of the 20 ResNet-20 checkpoints as base models, fine-tune three children from each base under mild, moderate, and aggressive regimes, and classify each child against all 20 bases using ℓ_2 nearest-neighbor search on the sorted column norms concatenated across three convolutional layers. The resulting separation ratios, defined as the minimum between-base distance divided by the maximum within-base distance, are 26.7, 6.0, and 1.59. All exceed 1, yielding 15/15 correct attributions and zero false positives over 285 non-lineage pairs.

At ImageNet scale, the same fingerprint map applied to torchvision-pretrained ResNet-50 V1 and V2, along with a lineage variant V_1^{ft} obtained by fine-tuning V1 for two epochs on Imagenette, gives $\|\phi(V_1) - \phi(V_2)\|_2 = 1106.43$ and $\|\phi(V_1) - \phi(V_1^{\mathrm{ft}})\|_2 = 1.38$. Since V_1^{ft} remains within the fine-tuning neighborhood of V_1, $\|\phi(V_2) - \phi(V_1^{\mathrm{ft}})\|_2$ agrees with $\|\phi(V_1) - \phi(V_2)\|_2$ to four significant digits. This yields an approximate 800:1 lineage-to-non-lineage distance ratio on a 25.5-million-parameter backbone.

5.2 Distillation Ablation

The leaked column norms further facilitate extraction: adding the recovered spectral prior to standard logit distillation raises student accuracy by up to 3.1 percentage points over the logit-only baseline, with the permutation-invariant *Col-norms* statistic capturing most of the gain at large query budgets. Full results across query budgets 1,000–5,000 and two architectures are in Appendix D.

Table 3. Confidentiality/utility tradeoff on ResNet-20 and ResNet-56 (CIFAR-10, 8-bit). *Atk err*: mean per-column ℓ_∞ error. *FP*: fingerprint top-1 among 20 checkpoints (ResNet-20 only). *Pred.agr*: prediction agreement. *Acc*: per-layer test accuracy under clipped-Gaussian noise (N=200 for ResNet-20, N=300 for ResNet-56).

	ResNet-20				ResNet-56		
Mult.	Atk err	FP	Pred.agr	Acc	Atk err	Pred.agr	Acc
1×	0	20/20	100.0%	82.2	0	100.0%	82.2
10×	1.4×10^{-2}	20/20	99.8%	82.2	1.3×10^{-2}	99.8%	82.2
100×	1.1×10^{-1}	20/20	98.3%	82.2	1.1×10^{-1}	98.2%	82.2
1,000×	1.02	1/20	63.7%	59.0	1.01	51.8%	48.9
5,000×	6.34	1/20	10.6%	10.1	6.34	9.6%	10.0

5.3 The Privacy/Utility Dilemma

SAFHIRE [2, Remark 4] suggests increasing the output noise level as a possible defensive countermeasure; the study below quantifies the utility cost. Since its shuffle-DP argument in Theorem 2 treats each round's convolution output as a local randomizer, this proposal is naturally interpreted as increasing the noise injected at every layer. Table 3 instantiates this sweep under two defender models. *Per-layer* injects clipped-Gaussian noise at the output of every convolutional layer, matching the defense suggested in Remark 2, whereas *Proxy*, which we retain only as a reference point, injects noise solely at the final classifier output and therefore upper-bounds the defender's attainable utility.[3]

Fingerprinting remains effective up to a 100× noise multiplier with essentially no utility loss: all 20 ResNet-20 models are still identified, and per-layer agreement remains 98.3%. It breaks down only at noise levels of at least 1,000×, where per-layer accuracy under clipped-Gaussian noise has already dropped to 59.0% on ResNet-20 and 48.9% on ResNet-56; see Figure 1. For ImageNet-scale ResNet-50 on Imagenette with $N = 20$, utility collapses between 30× and 70×, about an order of magnitude earlier than on CIFAR, since noise compounds across 53 layers, as shown by the green curve in Figure 1.

Theorems 1–2 are distribution-agnostic: across all six bounded noise distributions tested (zero, Gaussian, triangular, Laplace, uniform, and worst-case $\eta_k \in \{+B, -B\}$), no distribution reverses the ordering; adopting a smoother noise law can at best delay the utility collapse by a constant factor. Full per-distribution accuracy figures are in Appendix C.

In the per-layer noise-injection sweep with i.i.d. noise, no choice of noise scale, architecture, or bounded noise distribution defeats the attack while preserving usable accuracy. Thus, this defense family admits no viable operating point. Moreover, repeating each column query k times and averaging reduces the recovery error by a factor of $1/\sqrt{k}$: at 1,000× noise, $k = 5$ lowers the mean

[3] The closest prior noise-based defense study [22] injects Gaussian noise into input embeddings against input-privacy attacks on transformers, which differs both in attack target and injection point. To the best of our knowledge, no prior work quantifies per-layer noise as a defense against model-confidentiality attacks in this scheme class.

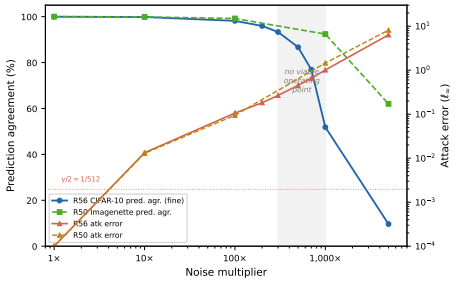


Fig. 1. Per-layer noise phase transition for ResNet-56 CIFAR-10 (blue, N=100, nine levels) and ResNet-50 Imagenette (green, N=20), clipped-Gaussian at p=256. Attack error (red, right axis) and $\gamma/2$=1/512 for reference. Shaded band: attack broken but utility intact; neither curve enters it.

error from 0.92 to 0.38. Thus, increasing the noise raises the query cost quadratically rather than preventing recovery. Alternative escape routes, including input validation, zero-knowledge well-formedness proofs, function secret sharing, and fully server-side FHE, are discussed in Section 6.

6 Discussion

Coverage across the scheme class. Theorem 1 relies on three hypotheses: (P) the client observes shuffled, noisy outputs of $Wx + b$; (N) the noise satisfies the correctness bound on a known quantization lattice; and (C) the client may issue chosen-input queries. They are worth isolating because closely related schemes satisfy different subsets. Under its stated threat model, SAFHIRE [2] satisfies all three at every layer and is therefore the primary target. STIP [28] and Centaur [16] violate both (N) and (C): they use exact floating-point computation without quantization-bounded noise, keep intermediate activations encrypted or secret-shared, and reveal only the final output, so Theorem 1 does not yield a meaningful client-side attack. Fission [24] violates (P), since as an MPC protocol it never reveals a permuted layer output to the client. PermLLM [30] falls outside Definition 1, since its permutation acts on the sequence axis rather than the output-channel axis $\pi \in S_m$.

Compound leakage. The attack recovers $s_i(W, b)$ rather than the full weight matrix. In SAFHIRE, moreover, the final layer is left unshuffled to enable client-side $\arg\max$ evaluation. As a result, the recovered spectra combine with exact logit access to form a compound leakage channel, exposing both the model's functionality and its identity, as discussed in Section 5.2.

Necessary conditions for a defense. Any viable defense must invalidate at least one hypothesis of Theorem 1. The known design space offers three broad escape routes. *Secret linear transforms* mask $Wx + b$ so that, without the key, it is computationally indistinguishable from random, violating (P). *Function secret sharing* across non-colluding servers prevents any one party from observing the

permuted output, violating both (P) and (C). *Fully server-side FHE* withholds intermediate decryptions from the client, violating (C). GAZELLE [13], DELPHI [17], and CrypTFlow2 [20] instantiate variants of the second route, though at the bandwidth cost that permutation-based schemes were meant to avoid.

Per-query permutations. Because the recovered quantity is a permutation-invariant summary, the guarantee of Theorem 1 holds for any permutation strategy, including one that draws a fresh permutation for every query. Table 1 confirms this empirically: the fresh-per-query regime leaves the maximum recovery error at zero across all 21 layers, so refreshing the permutation on each query does not strengthen the scheme.

Acknowledgements. The authors would like to thank the anonymous reviewers for their helpful comments. This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. RS-2024-00399491, 75%). Jiseung Kim was also supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. RS-2026-25477662, 25%).

References

1. Balle, B., Bell, J., Gascón, A., Nissim, K.: The privacy blanket of the shuffle model. In: Boldyreva, A., Micciancio, D. (eds.) CRYPTO 2019, Part II. pp. 638–667. LNCS, Springer (2019)
2. Biswas, S., Chartier, P., Dhasade, A.B., Jurien, T., Kerriou, D., Kerrmarec, A., Lemou, M., Tranie, F., de Vos, M., Vujasinovic, M.: Practical and private hybrid ML inference with fully homomorphic encryption. CoRR **abs/2509.01253** (2025)
3. Carlini, N., Jagielski, M., Mironov, I.: Cryptanalytic extraction of neural network models. In: Micciancio, D., Ristenpart, T. (eds.) CRYPTO 2020, Part III. pp. 189–218. LNCS, Springer (2020)
4. Chaturvedi, B., Chakraborty, A., Chatterjee, A., Mukhopadhyay, D.: Model stealing attacks on fhe-based privacy-preserving machine learning through adversarial examples. IACR Cryptol. ePrint Arch. **2023**, 1665 (2023)
5. Chen, S., Fan, J.: SEEK: model extraction attack against hybrid secure inference protocols. IACR Cryptol. ePrint Arch. **2022**, 1200 (2022)
6. Chillotti, I., Gama, N., Georgieva, M., Izabachène, M.: TFHE: fast fully homomorphic encryption over the torus. J. Cryptol. **33**(1), 34–91 (2020)
7. Desfontaines, D., Pejó, B.: SoK: Differential privacies. Proc. Priv. Enhancing Technol. **2020**(2), 288–313 (2020)
8. Dwork, C., Roth, A.: The algorithmic foundations of differential privacy. Found. Trends Theor. Comput. Sci. **9**(3-4), 211–407 (2014)
9. Feldman, V., McMillan, A., Talwar, K.: Hiding among the clones: A simple and nearly optimal analysis of privacy amplification by shuffling. In: FOCS 2021. pp. 954–964. IEEE (2021)
10. Folkerts, L.W., Tsoutsos, N.G.: Testing robustness of homomorphically encrypted split model llms. In: DATE 2025. pp. 1–7. IEEE (2025)
11. Gawron, G., Stubbings, P.: Feature space hijacking attacks against differentially private split learning. In: AAAI PPAI Workshop (2022)

12. Jagielski, M., Carlini, N., Berthelot, D., Kurakin, A., Papernot, N.: High accuracy and high fidelity extraction of neural networks. In: Capkun, S., Roesner, F. (eds.) USENIX Security 2020. pp. 1345–1362. USENIX Association (2020)
13. Juvekar, C., Vaikuntanathan, V., Chandrakasan, A.P.: GAZELLE: A low latency framework for secure neural network inference. In: Enck, W., Felt, A.P. (eds.) USENIX Security 2018. pp. 1651–1669. USENIX Association (2018)
14. Karchmer, A.: Theoretical limits of provable security against model extraction by efficient observational defenses. In: IEEE SaTML 2023. pp. 605–621. IEEE (2023)
15. Lukas, N., Zhang, Y., Kerschbaum, F.: Deep neural network fingerprinting by conferrable adversarial examples. In: ICLR 2021. OpenReview.net (2021)
16. Luo, J., Chen, G., Zhang, Y., Liu, S., Wang, H., Yu, Y., Zhou, X., Qi, Y., Xu, Z.: CENTAUR: bridging the impossible trinity of privacy, efficiency, and performance in privacy-preserving transformer inference. In: Che, W., Nabende, J., Shutova, E., Pilehvar, M.T. (eds.) ACL 2025. pp. 22751–22770. ACL (2025)
17. Mishra, P., Lehmkuhl, R., Srinivasan, A., Zheng, W., Popa, R.A.: Delphi: A cryptographic inference service for neural networks. In: Capkun, S., Roesner, F. (eds.) USENIX Security 2020. pp. 2505–2522. USENIX Association (2020)
18. Naveed, M., Kamara, S., Wright, C.V.: Inference attacks on property-preserving encrypted databases. In: Ray, I., Li, N., Kruegel, C. (eds.) ACM CCS 2015. pp. 644–655. ACM (2015)
19. Orekondy, T., Schiele, B., Fritz, M.: Knockoff nets: Stealing functionality of black-box models. In: CVPR 2019. pp. 4954–4963. IEEE (2019)
20. Rathee, D., Rathee, M., Kumar, N., Chandran, N., Gupta, D., Rastogi, A., Sharma, R.: CrypTFlow2: Practical 2-party secure inference. In: Ligatti, J., Ou, X., Katz, J., Vigna, G. (eds.) ACM CCS 2020. pp. 325–342. ACM (2020)
21. Shokri, R., Stronati, M., Song, C., Shmatikov, V.: Membership inference attacks against machine learning models. In: SP 2017. pp. 3–18. IEEE (2017)
22. Thomas, R.K., Zahran, L., Choi, E., Potti, A., Goldblum, M., Pal, A.: Hidden no more: Attacking and defending private third-party LLM inference. In: ICML 2025. Proceedings of Machine Learning Research (2025)
23. Tramèr, F., Zhang, F., Juels, A., Reiter, M.K., Ristenpart, T.: Stealing machine learning models via prediction apis. In: Holz, T., Savage, S. (eds.) USENIX Security 2016. pp. 601–618. USENIX Association (2016)
24. Ugurbil, M., Mouris, D., Santos, M.B., Cabrero-Holgueras, J., de Vega, M., Sengupta, S.: Fission: Distributed privacy-preserving large language model inference. In: IEEE CNS 2025. pp. 1–9. IEEE (2025)
25. Wang, P., Dong, B., Cai, Y., Zhang, Z., Liu, J., Xue, H., Wu, Y., Zhang, Y., Zhang, Z.: Game of arrows: On the (in-)security of weight obfuscation for on-device TEE-shielded LLM partition algorithms. In: Bauer, L., Pellegrino, G. (eds.) USENIX Security 2025. pp. 279–298. USENIX Association (2025)
26. Xiao, H., Suh, G.E., Devadas, S.: Formal privacy proof of data encoding: The possibility and impossibility of learnable encryption. In: Luo, B., Liao, X., Xu, J., Kirda, E., Lie, D. (eds.) ACM CCS 2024. pp. 1834–1848. ACM (2024)
27. Yan, H., Li, X., Li, H., Li, J., Sun, W., Li, F.: Monitoring-based differential privacy mechanism against query flooding-based model extraction attack. IEEE Trans. Dependable Secur. Comput. **19**(4), 2680–2694 (2022)
28. Yuan, M., Zhang, L., Cheng, Y., Song, M.H., Xing, G., Li, X.Y.: STIP: Three-party privacy-preserving and lossless inference for large transformers in production. In: NDSS (2026), accepted paper
29. Zama AI: Concrete: TFHE compiler for encrypted computation, v2.11 (2024)

30. Zheng, F., Chen, C., Han, Z., Zheng, X.: Permllm: Private inference of large language models within 3 seconds under WAN. CoRR **abs/2405.18744** (2024)

A CIFAR-10 / ResNet-20 Experimental Details

Training details for the 20 ResNet-20 checkpoints used in Section 4: SGD with learning rate 0.1, momentum 0.9, weight decay 10^{-4}, MultiStepLR milestones at epochs 5 and 8, batch size 128, 10 epochs, and seeds $0, \ldots, 19$, yielding mean test accuracy 82.8%. Theorem 1 is deterministic under the correctness bound, so zero-error recovery does not depend on teacher accuracy; longer training affects only the fingerprint-separation constant $\Delta_{\mathcal{F}}^{\phi}$ in Section 5.1. The CIFAR-100 comparison there, trained for 200 epochs, gives the larger value $\Delta_{\mathcal{F}}^{\phi} = 0.565$ and still separates all 190 pairs. All 21 convolution modules use `bias=False`, so before batch-normalization folding we have $b = 0$ and the raw convolution reduces to the exact weight multiset. In the artifact code, BN layers are folded into the preceding convolution via the standard deployment transform $W \mapsto \gamma W/\sigma_{\mathrm{BN}}$ and $b \mapsto \gamma(\mu_{\mathrm{conv}} - \mu_{\mathrm{BN}})/\sigma_{\mathrm{BN}} + \beta$. Accordingly, the recovered object is the folded affine pair (W', b') that the client actually sees as the layer output. Residual branches $y = F(x) + x$ are probed branchwise, consistent with SAFHIRE's per-layer ciphertext reply pattern. The per-layer input dimensions d_l range from 16 for `layer2.0.shortcut` to 576 for `layer3.2.conv2`, yielding the per-model budget $\sum_{l=1}^{21}(d_l + 1) = 5{,}712$ from Theorem 1.

B Tightness of the Correctness Bound

Proposition 6. *If $\Delta > \gamma/2$, there exists η with $\|\eta\|_\infty < \Delta$ under which no deterministic lattice-valued algorithm recovers* $\mathrm{sort}(W_{:,i} + b)$ *correctly.*

Proof. Fix $\eta^* \in (\gamma/2, \min(\gamma, \Delta))$, which is nonempty since $\Delta > \gamma/2$. Two instances share the same observation $o := k\gamma + \eta^*$: instance A with true value $k\gamma$ and noise η^*, and instance B with true value $(k{+}1)\gamma$ and noise $\eta^* - \gamma$, where $|\eta^* - \gamma| = \gamma - \eta^* < \gamma/2 < \Delta$ since $\eta^* < \gamma$. Distinct true values and identical observation force any deterministic algorithm to err on at least one instance. Since sort is deterministic, the ambiguity carries over to the sorted spectrum. □

C Additional Experimental Validation

Noise distribution robustness. We test six bounded noise distributions with $\|\eta\|_\infty < \Delta$ against the per-layer noise-injection defense of Section 5.3. Under zero noise, ResNet-20 and ResNet-56 retain 82.3% and 82.2% accuracy at both $m = 100$ and $m = 1{,}000$. Under Gaussian noise, both remain at 82.2% for $m = 100$, then drop to 59.0% for ResNet-20 and 48.9% for ResNet-56 at $m = 1{,}000$. Heavier-tailed distributions collapse earlier: at $m = 1{,}000$, triangular reaches 41.3% / 29.8%, Laplace 21.6% / 13.6%, and uniform 18.7% / 12.5%. The

Table 4. Logit distillation test accuracy (%) on CIFAR-10 under four spectrum-prior ablations (mean ± std, 10 seeds). Teacher accuracies: 82.92% (ResNet-20) and 82.47% (ResNet-56).

Architecture	q	Baseline	First	Col-norms	Full
ResNet-20	1,000	47.21 ± 1.21	48.60 ± 1.51	50.09 ± 0.94	51.11 ± 1.06
	2,000	56.16 ± 1.60	58.84 ± 1.20	59.87 ± 0.91	60.07 ± 0.82
	5,000	71.58 ± 1.67	72.71 ± 2.14	74.66 ± 0.50	74.12 ± 0.97
ResNet-56	2,000	53.28 ± 1.68	53.60 ± 1.10	55.38 ± 1.37	54.53 ± 1.64
	5,000	70.99 ± 1.45	72.46 ± 2.07	73.50 ± 0.92	72.53 ± 1.10

worst-case deterministic envelope $\eta_k \in \{+B, -B\}$ is most damaging, reaching 76.8% / 75.6% at $m = 100$ and 10.0% for both architectures at $m = 1{,}000$. No distribution reverses this ordering, supporting the distribution-agnostic guarantee of Theorem 1.

ImageNet-scale experiments. Table 2 uses torchvision `IMAGENET1K_V1` weights (eight CNNs and ViT-B/16). The ResNet-50 per-layer curve (Fig. 1) uses a fixed 500-image Imagenette-val subset (50/class, `random.seed(0)`) with top-1 against real ImageNet-1K labels via the WNID→index map. The lineage variant V_1^{ft} is V1 after two epochs of SGD on the same subset (LR 10^{-3}, batch 32).

Bias-having MLP. To validate the bias case of Theorem 1, we train a two-layer MLP with biases on MNIST (97.7% test accuracy; `fc1` : 128×784, `fc2` : 64×128, 8-bit precision). For every column, $\mathrm{sort}(W_{:,i}+b) \neq \mathrm{sort}(W_{:,i})+\mathrm{sort}(b)$, confirming that the bias cannot be separated from the recovered multiset; round-then-sort recovers $\mathrm{sort}(W_{:,i}+b)$ exactly on all columns of both layers.

D Distillation Ablation

We compare four knowledge-distillation variants that differ only in the spectral prior drawn from the recovered $L(W, b)$: *Baseline* (logits only); *First* ($\mathrm{sort}(W_{:,i})$ from `conv1`); *Col-norms* (column norms across all 21 convolutional layers); *Full* (column norms + per-layer weight-value remapping). A *Scrambled* control replaces teacher column norms with those of an independently trained model to isolate structural vs. teacher-specific gains.

On ResNet-20, *Full* is strictly best at $q \leq 2{,}000$; *Col-norms* becomes competitive at $q = 5{,}000$ (74.66 ± 0.50 vs. 74.12 ± 0.97). On ResNet-56, *Col-norms* already matches or exceeds *Full* at both budgets. The *Scrambled* control (74.02 ± 1.10 at $q = 5{,}000$) is within one std of *Col-norms*, indicating the gain is largely structural.

Configuration. Teachers: `resnet20_seed0.pt` (82.92%) and `resnet56_seed0.pt` (82.47%); students trained from scratch with SGD (LR 0.05, momentum 0.9, wd 10^{-4}, cosine anneal, 100 epochs, batch 128). KL on softmax logits at T=4; column-norm regularizer weight λ=5 for *Col-norms*/*Full*. Query budgets drawn from a fixed 5,000-image random subset (`random.seed(0)`); 10,000-image test set; 10 seeds.