NO FREE EFFICIENCY: REVISITING THE TRADE-OFF BETWEEN TRAINING EFFICIENCY AND MODEL VULNERABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Training efficiency has become the central driver of recent progress in foundation models. To overcome the massive computational and data requirements of large-scale training, researchers increasingly adopt strategies such as selective data sampling, efficient pre-training, and simplified reinforcement learning pipelines. While these strategies drastically reduce overhead, they prompt a critical, yet neglected question: Is efficiency achieved at the expense of model robustness and security? To our knowledge, we present the first systematic cross-domain investigation of the efficiency–vulnerability trade-off. Across vision and language models, we show that efficiency-oriented training increases susceptibility to adversarial and privacy attacks. We characterize this vulnerability by analyzing the models' internal geometry and functional representations, demonstrating that the evaluated efficient variants consistently exhibit sharper loss geometry together with systematic changes in representational structure. We further extend our analysis to "zero RL training", finding that models trained using simplified RL recipes exhibit substantially greater susceptibility to catastrophic forgetting and more pronounced overconfidence than those trained through conventional alignment pipelines. Our findings suggest that training efficiency is rarely a "free lunch"; rather, the mechanisms that minimize computation can inadvertently compromise safety. We conclude by calling for a paradigm shift toward multi-objective training that jointly optimizes for performance, cost, and security.

1 INTRODUCTION

In recent years, foundation models have reshaped the world at large, becoming defining technologies embedded in search engines and creative tools. However, their success remains tied to scale, requiring massive datasets and tens of millions of dollars in compute.

This financial burden has sparked an "efficiency revolution," with researchers and companies exploring approaches to reach similar performance at a fraction of the cost. Recent examples include DeepSeekV3 DeepSeek-AI (2024) and Kimi 2 Thinking Kimi-Team (2025), which have drawn wide attention for their highly efficient training strategies. While exact costs are not public, GPU usage for DeepSeekV3 suggests a compute cost of roughly 5.6 million US dollars DeepSeek-AI (2024). This is a dramatic reduction from conventional models. These results have even influenced the market value of hardware companies such as NVIDIA NVI, which sparks an optimistic question: *can we truly build powerful AI systems without paying the traditional computational and financial cost?*

However, as the "no free lunch" principle suggests Wolpert & Macready (1997), efficiency improvements in complex systems can carry hidden trade-offs. While current discussions emphasize performance and scalability, the potential security and privacy implications of these efficiency techniques remain largely unexplored. This gap motivates our work: we aim to understand whether the pursuit of efficiency in foundation model training introduces new vulnerabilities that may compromise robustness and safety.

Our findings demonstrate that the representative efficiency strategies studied here carry measurable security costs. Through comprehensive cross-domain experiments, we find that the evaluated efficiency-oriented variants exhibit systematic changes in their internal geometry and functional

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.

representations, alongside greater susceptibility to adversarial and privacy attacks. In short, the same strategies that reduce training requirements can also introduce new avenues for exploitation.

We make the following contributions:

- First systematic cross-domain analysis of the efficiency–vulnerability trade-off: We analyze multiple stages of the training pipeline, including data selection, pre-training, and alignment, across computer vision and language domains.
- Theoretical intuition for the observed trade-offs: We identify links between efficiency choices and measurable changes in a model's geometric and functional properties. This provides conceptual insights into why efficient models may become more fragile.
- Hidden security costs of efficient LLM alignment: We reveal that zero-RL models can preserve reasoning utility while exhibiting substantially greater structural brittleness and overconfidence, uncovering a previously overlooked tension between alignment efficiency and model security.

Our study demonstrates that improving training efficiency is not a free gain. As foundation models continue to shape the technological landscape, understanding this hidden cost is essential for building systems that are not only efficient and powerful but also secure and trustworthy.

2 Related Work

2.1 Training-Efficiency Strategies

The substantial cost of modern model training has led to a wide range of techniques for reducing data and computational requirements. One prominent direction reduces the volume of training data through high-importance core-set selection Ghorbani & Zou (2019); Jia et al. (2019); Paul et al. (2021); Zhou et al. (2023); Ji et al. (2025), which identifies a compact subset that preserves task utility. Beyond data reduction, pre-training followed by downstream fine-tuning Dosovitskiy et al. (2021); Devlin et al. (2019); DeepSeek-AI (2024) amortizes representation learning across tasks, while parameter-efficient fine-tuning Hu et al. (2022); Zhang et al. (2024) further lowers adaptation costs by updating only a small portion of the model. A third direction concerns post-training alignment. Conventional approaches such as RLHF Christiano et al. (2017); Ouyang et al. (2022) and DPO Rafailov et al. (2023) rely on preference data and additional optimization; recent "zero RL training" methods Zeng et al. (2025); Cui et al. (2026) reduce this burden by starting directly from capable base models without initial SFT. These three lines of work define the training-efficiency landscape examined in our study.

2.2 Efficiency and Model Trustworthiness

A growing body of work shows that choices made to improve or modify training can also reshape model trustworthiness. Differentially private fine-tuning Li et al. (2022); Bao et al. (2025) can reduce memorization and membership-inference risk, whereas model distillation Ge et al. (2021); Muri et al. (2026) can transfer or amplify backdoor behavior. Data selection Wen et al. (2025); Ramesh et al. (2025) may likewise influence privacy leakage and stability. Related studies consider efficiency after training, showing that pruning Ye et al. (2019); Egashira et al. (2025), quantization Yu et al. (2024), and LLM compression more broadly Hong et al. (2024) can degrade robustness and other dimensions of trustworthiness. Collectively, these findings establish important connections between individual efficiency mechanisms and security, but they remain fragmented across particular methods, stages, and domains. Our work provides the first systematic cross-domain analysis that connects these questions across multiple stages of model training.

3 Empirical Framework

To systematically investigate the efficiency-vulnerability trade-off across model families and training stages, we combine controlled, utility-matched studies of vision models with cross-scale evaluations of representative efficient alignment recipes for LLMs.

3.1 Defining Training Efficiency

In foundation model development, training efficiency means reducing data, computation, or alignment overhead while maintaining comparable task utility. Our work focuses on three major efficiency paradigms that correspond to different stages of development, illustrated in Figure 1:

Data Selection. This stage minimizes the amount of training data used to save on memory, storage, and collection costs. We use strategies like high-importance core-set selection to find the most informative data samples. This often involves data valuation metrics like Shapley values to prioritize "high-importance" records Ghorbani & Zou (2019). This approach is ideal when collecting data is expensive or when the goal is to train on only the most critical information.

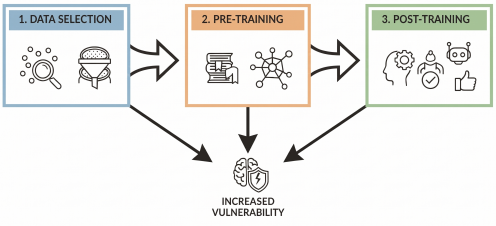


Figure 1: Lifecycle of model training.

Pre-training. This paradigm uses a two-stage Pre-training and Fine-tuning pipeline. Pre-training uses the most resources to build general knowledge from a massive dataset. Fine-tuning then uses this knowledge base to adapt the model to specific tasks with very little additional compute. Our study explores the trade-off between using this efficient pre-trained base versus the higher cost of training a model from scratch.

Post-training Alignment. Alignment elicits specialized behaviors, such as mathematical reasoning. Conventional alignment pipelines typically build on SFT and may regularize policy updates relative to a reference policy. For example, KL-regularized RLHF maximizes

$$\mathcal{J}_{\mathrm{RLHF}}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi_\theta(\cdot|x)}\left[r(x,y) - \beta \log \frac{\pi_\theta(y|x)}{\pi_{\mathrm{ref}}(y|x)}\right], \tag{1}$$

where $\mathcal{D}$ is the prompt distribution, π_θ is the trainable policy, $r(x, y)$ is the reward, π_{ref} is the reference policy, and β controls the strength of the deviation penalty. The evaluated "zero RL training" approaches reduce alignment overhead by starting directly from capable base models without preliminary SFT and simplifying conventional alignment components, including the omission of explicit KL regularization. Our comparison therefore examines whether these utility-preserving approaches incur greater security risks, treating each training pipeline as a whole rather than isolating the effect of any single omitted component.

3.2 Measuring Vulnerability

To quantify the cost of efficiency, we adopt a comprehensive approach to measure model vulnerability across three core security concerns: data privacy, robustness, and proprietary security.

Data Privacy. This dimension measures whether training data can be inferred from the model. We evaluate this risk using membership inference attacks (MIA), where an adversary attempts to determine whether a specific sample was included in the training set. For vision tasks, we adopt three computationally efficient and widely used MIAs: prediction confidence-based attack Song et al. (2019), entropy-based attack Salem et al. (2019), and modified prediction entropy-based attack Song & Mittal (2021). For math reasoning LLMs, we construct source-defined candidate-member and candidate-non-member groups and evaluate answer likelihood Fu et al. (2024) (analogous to loss-based attacks), model self-confidence Yoon et al. (2025) (analogous to confidence-based attacks), and the Min-k% probability attack, which provide complementary membership signals for discrete reasoning tasks.

Model Robustness and Resilience. This category assesses whether the model is easily manipulated by external perturbations or prone to functional degradation. We evaluate this across two dimensions: adversarial resistance and adaptation stability. To quantify susceptibility to adversarial manipulation (being "tricked"), we employ domain-specific attacks. For vision tasks, we measure how many Projected Gradient Descent (PGD) Madry et al. (2018) iterates remain correctly classified on average;

Table 1: Accuracy (training accuracy in brackets) and TPR at 0.1% FPR of membership inference for data selection with Shapley values. Higher TPR indicates greater privacy leakage.

Dataset	Value	Acc	MIA		
			conf	entr	modi
CIFAR-10	Low	48.07 (99.97)	0.12%	0.05%	0.11%
	High	47.97 (99.97)	0.16%	0.09%	0.16%
CIFAR-10 (swin)	Low	37.01 (99.95)	0.08%	0.05%	0.09%
	High	36.94 (99.93)	0.29%	0.30%	0.30%
TinyImageNet	Low	36.09 (100.00)	0.08%	0.06%	0.10%
	High	36.17 (100.00)	0.22%	0.35%	0.27%
PubFig83	Low	44.01 (100.00)	0.32%	0.32%	0.32%
	High	44.26 (100.00)	1.51%	1.43%	1.51%

Table 2: Adversarial robustness for data selection and pre-training in vision models, measured by the average number of correctly classified PGD iterates (higher is better).

Dataset	Data Selection		Pre-training	
	Value	Steps	Paradigm	Steps
CIFAR-10	Low	3.85	Scratch	3.58
	High	1.33	Pre-train	1.00
CIFAR-10 (swin)	Low	0.70	Scratch	3.22
	High	0.18	Pre-train	0.11
TinyImageNet	Low	20.89	Scratch	14.03
	High	10.46	Pre-train	3.64
PubFig83	Low	2.34	Scratch	1.83
	High	1.30	Pre-train	0.41

a lower score indicates a more fragile robustness margin. For math reasoning LLMs, we use the RobustMath dataset Zhou et al. (2024) to evaluate resilience against semantic perturbations and AutoDAN Liu et al. (2024) on AdvBench Zou et al. (2023) to evaluate jailbreak susceptibility under adversarial prompting. Additionally, we evaluate fine-tuning stability by adapting the model to a disjoint domain using LoRA Hu et al. (2022), measuring if the model maintains its integrity or suffers catastrophic performance loss under distribution shift.

Proprietary Security. This dimension concerns the risk of unauthorized extraction of a model's proprietary knowledge and structure. We evaluate this threat using Model Stealing Attacks (MS), in which an adversary attempts to reconstruct a functional equivalent of the target model by querying its outputs. We focus on a logit-query setting, using data from the target distribution for vision tasks.

4 Controlled Study in Vision Tasks

4.1 Experimental setup

To examine how efficiency choices affect vulnerability at comparable utility, we compare a higher-cost baseline with an *Efficient Variant*. We calibrate dataset sizes and configurations so that each pair achieves nearly identical training and testing accuracy, preventing task-performance differences from confounding the security comparison. All vision results are means over three seeds; complete mean $\pm$ standard deviation results are reported in Appendix B.1.

Datasets and Architecture. We employ ResNet-18 and Swin Transformer architectures to evaluate vulnerability across distinct inductive biases, using ResNet-18 by default unless "Swin" is specified. Our experiments span three datasets: CIFAR-10 and TinyImageNet for standard object classification, and PubFig83 for fine-grained face recognition. This diversity allows us to test whether the observed patterns persist across architectures and tasks involving different levels of intra-class variation.

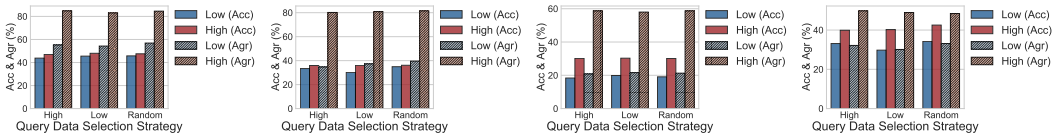
Model Training. We study efficiency through data selection and pre-training. For data selection, KNN-Shapley Jia et al. (2019) ranks the training samples: *High* uses a smaller subset with the highest values, whereas *Low* uses a larger subset with the lowest values. The two subsets are disjoint, and their sizes are intentionally calibrated to match clean utility, with all other training hyperparameters fixed within each comparison. For pre-training, the Efficient Variant fine-tunes a pre-trained model on less task-specific data, whereas the higher-cost baseline is trained from scratch on a larger dataset. We thus operationalize efficiency through reduced task-specific data and reuse of pre-trained representations.

4.2 Experimental Results

Our controlled comparisons uncover a fundamental trade-off between data efficiency and security. Despite comparable generalization and overfitting levels (the gap between training and testing accuracy), High-Shapley Efficient Variants exhibit heightened MIA vulnerability (Table 1). Relative

Table 3: Accuracy (training accuracy in brackets), membership inference (MIA), and model stealing (MS) under pre-trained fine-tuning and scratch training. Higher MIA indicates greater privacy leakage. MS reports stolen-model accuracy with target–surrogate agreement in parentheses; higher values indicate more successful extraction.

Dataset	Paradigm	Acc	MIA			MS	
			conf	entr	modi	1000	1000 (Pre-train)
CIFAR-10	Scratch	67.71 (99.91)	0.05%	0.04%	0.05%	62.60 (0.748)	65.22 (0.758)
	Pre-train	67.88 (100.00)	0.21%	0.16%	0.21%	58.05 (0.669)	67.25 (0.785)
CIFAR-10 (swin)	Scratch	73.85 (100.00)	0.12%	0.12%	0.11%	42.41 (0.446)	57.70 (0.536)
	Pre-train	74.01 (100.00)	0.54%	0.54%	0.54%	39.01 (0.437)	60.06 (0.651)
TinyImageNet	Scratch	34.91 (99.98)	0.13%	0.07%	0.14%	24.65 (0.381)	27.53 (0.379)
	Pre-train	34.93 (100.00)	0.75%	0.75%	0.67%	17.65 (0.305)	29.82 (0.529)
PubFig83	Scratch	35.19 (100.00)	0.01%	0.01%	0.01%	17.81 (0.358)	18.59 (0.299)
	Pre-train	34.79 (100.00)	0.36%	0.46%	0.37%	7.89 (0.183)	25.26 (0.533)



(a) CIFAR-10 (b) CIFAR-10 (swin) (c) TinyImageNet (d) PubFig83

Figure 2: Model stealing under Shapley-based data selection. "Acc" denotes stolen-model test accuracy, "Agr" the fraction of test samples on which target and stolen models agree, and "Random" random query selection; higher Acc and Agr indicate more successful extraction. Query pools contain 1,000 distinct images (500 for PubFig83 owing to data limitations).

to the Low-Shapley baselines, the gap is most pronounced with Swin Transformers and on challenging datasets such as TinyImageNet and PubFig83. Beyond this privacy exposure, High-Shapley variants remain correctly classified for fewer PGD iterates, revealing reduced adversarial robustness (Table 2). The model-stealing results broaden this concern further: attacks against these targets produce stolen models with higher accuracy and agreement (Figure 2). The consistency across these distinct attack surfaces establishes that importance-based data efficiency can preserve predictive utility while weakening the model's overall security posture.

The comparison of training paradigms further highlights the security cost of efficiency gains. At comparable utility, the pre-trained Efficient Variant exhibits greater MIA risk and lower adversarial robustness than the scratch-trained baseline (Tables 3 and 2). Model stealing reveals a critical nuance: although pre-trained targets resist scratch-initialized surrogates, this advantage disappears when the attacker uses the same pre-trained checkpoint. Because this checkpoint is publicly available, the informed attack requires no privileged access. Under this realistic threat model, the higher stolen-model accuracy and agreement show that the reusable prior reducing task-specific training cost can also lower the empirical barrier to functional replication.

5 Mechanistic Analysis

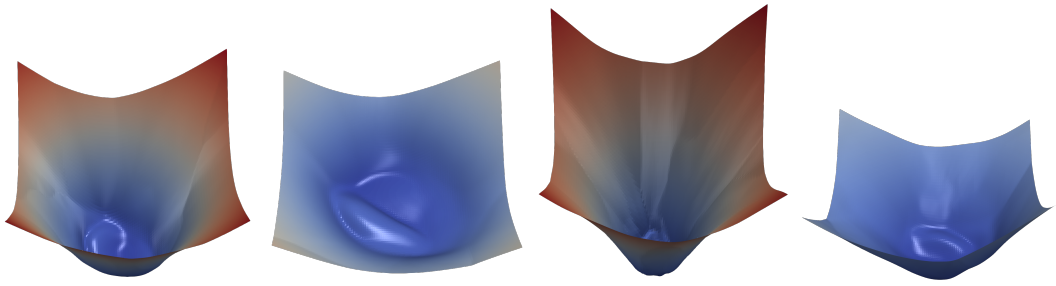
We analyze structural differences between Efficient Variants and baselines from two complementary perspectives: geometric stability and functional representation structure. These analyses identify concrete geometric and functional signatures of the evaluated Efficient Variants, providing mechanistic insight into their observed security degradation.

5.1 Geometric Instability

We hypothesize that the evaluated efficiency strategies steer models toward geometrically less stable solutions. We characterize geometric instability directly through loss-landscape sharpness and

Table 4: Sharpness, Expected Calibration Error, and PCA Participation Ratio for vision models.

Dataset	Value	Sharpness	ECE	PCA PR	Paradigm	Sharpness	ECE	PCA PR
CIFAR-10	Low	0.0410	0.382	28.490	Scratch	0.0311	0.219	14.266
	High	0.0591	0.430	10.793	Pre-train	0.1332	0.232	25.137
CIFAR-10 (swin)	Low	0.0170	0.410	13.907	Scratch	0.0010	0.060	9.897
	High	0.0189	0.436	10.896	Pre-train	0.0373	0.241	26.786
TinyImageNet	Low	0.0548	0.336	265.298	Scratch	0.0511	0.317	157.477
	High	0.0600	0.384	74.960	Pre-train	0.1201	0.251	163.206
PubFig83	Low	0.1234	0.208	48.395	Scratch	0.1349	0.350	31.940
	High	0.1475	0.299	33.936	Pre-train	0.2551	0.384	39.009



(a) High (b) Low (c) Pre-train (d) Scratch

Figure 3: Loss landscape for CIFAR-10.

examine its accompanying calibration behavior through Expected Calibration Error (ECE), which measures the alignment between predictive confidence and accuracy.

Loss Landscape Sharpness. We measure sharpness through local loss sensitivity around the model parameters θ. For a test batch b with loss $\mathcal{L}_b$, let $\widehat{\delta}_b$ denote the bounded loss-increasing perturbation found by projected gradient ascent, and let $\mathcal{B}$ denote the collection of test batches. We report

$$\widehat{S}(\theta) = \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} \left[\mathcal{L}_b(\theta + \widehat{\delta}_b) - \mathcal{L}_b(\theta) \right]. \tag{2}$$

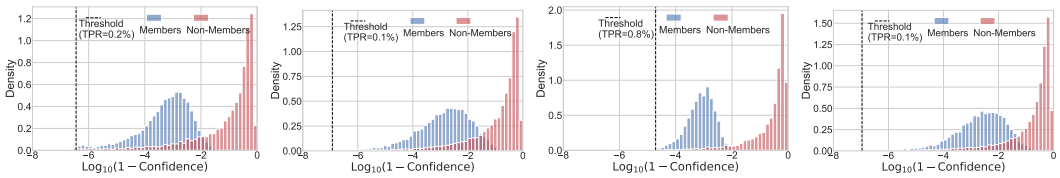
A larger value indicates greater local loss sensitivity. A local Taylor expansion connects this quantity to directional gradients and Hessian curvature (Appendix A.1), motivating its use as a measure of geometric stability. This connection suggests that prioritizing highly influential samples can concentrate gradient contributions, while fine-tuning on limited data can reduce gradient diversity and favor interpolation regimes associated with sharper solutions Ghorbani & Zou (2019); Yin et al. (2018); Feldman (2020); Kumar et al. (2022).

Across all evaluated comparisons, the High-Shapley and pre-trained variants exhibit higher sharpness than their respective baselines (Table 4) and visibly sharper local loss landscapes (Figure 3). Despite operating at different training stages, both efficiency strategies therefore exhibit the same pattern of reduced geometric flatness.

Expected Calibration Error. We form M equal-width confidence bins Guo et al. (2017):

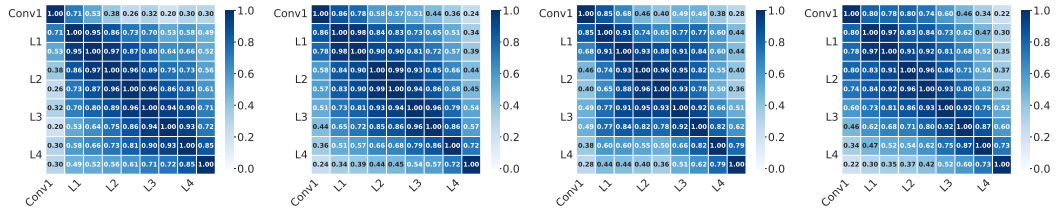
$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \left| \mathrm{acc}(B_m) - \mathrm{conf}(B_m) \right|, \tag{3}$$

where N is the total sample count, and $\mathrm{acc}(B_m)$ and $\mathrm{conf}(B_m)$ are the average empirical accuracy and confidence in bin B_m, respectively. Classical analyses show that, for separable data and positively homogeneous models, gradient-based optimization can increase weight norms and confidence after the training error vanishes Soudry et al. (2018); Lyu & Li (2020). Guided by this result, we interpret the observed behavior as follows: the evaluated efficiency strategies may accelerate this "logit explosion" without commensurate accuracy gains. Data selection can concentrate influential gradient



(a) High (confidence) (b) Low (confidence) (c) Pre-train (confidence) (d) Scratch (confidence)

Figure 4: Log-scale member/non-member confidence distributions on TinyImageNet.



(a) High (CIFAR-10) (b) Low (CIFAR-10) (c) High (TinyImageNet) (d) Low (TinyImageNet)

Figure 5: Intra-CKA under Shapley-based data selection on CIFAR-10 and TinyImageNet.

contributions, while limited-data fine-tuning can reach separation earlier and devote more updates to margin growth.

Our analysis reveals generally elevated miscalibration among Efficient Variants (Table 4 and Figure 4), while the associated separation between member and non-member confidence supplies an exploitable signal for MIA. Interestingly, TinyImageNet exhibits an inverse relationship between calibration and privacy. Although pre-training improves calibration, it heightens MIA vulnerability. We interpret this contrast through the overlap between TinyImageNet and ImageNet: pre-training supplies a strong prior that drives extreme confidence on members, widening their separation from non-members. In contrast, the baseline suffers from generalized overconfidence, obscuring the distributional distinguishability required for inference.

5.2 Functional Complexity and Redundancy

We next examine functional complexity and redundancy in relation to feature encoding and Model Stealing. We characterize these properties using Principal Component Analysis (PCA) and Centered Kernel Alignment (CKA).

Feature Dimensionality (PCA). We analyze Model Stealing through the effective feature dimension d_{eff} Maddox et al. (2020); Ansuini et al. (2019), which we estimate using the PCA Participation Ratio (PR). Let λ_i denote the eigenvalues of the empirical covariance $\mathbf{C}$ of centered penultimate-layer features. The participation ratio is $\text{PR}(\mathbf{C}) = (\sum_{i=1}^{d} \lambda_i)^2 / \sum_{i=1}^{d} \lambda_i^2$. A lower PR indicates that feature variance is concentrated in fewer dominant directions.

This spectral interpretation helps explain the data-selection results. High-Shapley selection retains easy, representative samples that preserve task utility while filtering out harder intra-class variations. The learned representation can therefore preserve the information needed for the task while concentrating feature variance into fewer directions, reducing the effective complexity faced by a naive (scratch-initialized) attacker. Consistent with this mechanism, High-Shapley variants exhibit lower PR across all evaluated settings (Table 4) and greater vulnerability to naive Model Stealing (Figure 2).

Pre-training follows a different pathway. Rather than reducing feature dimensionality, pre-trained variants maintain higher PR and resist scratch-initialized surrogates. However, an informed attacker initialized from the same public checkpoint inherits their shared representation prior and primarily needs to recover the task-specific adaptation (Table 3). These results identify compact feature geometry and shared initialization as distinct routes through which the evaluated efficiency strategies lower the empirical barrier to extraction.

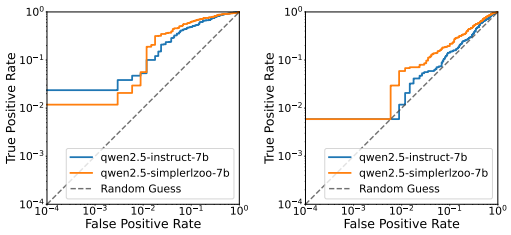
Representation Similarity (CKA). To quantify changes in the feature hierarchy, we employ linear CKA Kornblith et al. (2019), a standard representation-similarity metric. Let $X \in \mathbb{R}^{n \times d_1}$ and $Y \in \mathbb{R}^{n \times d_2}$ denote the activation matrices of two layers, with Gram matrices $K = XX^\top$ and $L = YY^\top$. Using the Hilbert–Schmidt Independence Criterion (HSIC), it is defined as

$$\mathrm{CKA}(K, L) = \frac{\mathrm{HSIC}(K, L)}{\sqrt{\mathrm{HSIC}(K, K)\,\mathrm{HSIC}(L, L)}}. \tag{4}$$

We interpret the High-Shapley pattern through simplicity bias Arpit et al. (2017); Pérez et al. (2019). By emphasizing easy, representative samples, data selection reduces the pressure for later layers to encode challenging intra-class variations. Accordingly, the Intra-CKA heatmaps in Figure 5 reveal more extensive late-layer self-similarity in High-Shapley models than in their Low-Shapley baselines, indicating that successive layers preserve similar representational geometry and contribute limited additional feature refinement. The lower PCA PR values in Table 4 reinforce this picture of a lower-dimensional, more homogeneous feature hierarchy and help explain the higher Model Stealing success observed in our experiments.

6 Analysis on Large Language Models

We now extend our evaluation to math-reasoning LLMs to examine whether the vulnerability pattern observed in vision also appears in another model domain. We focus on recent "Zero RL" or "Simple RL" approaches, which reduce alignment overhead by starting directly from capable base models without preliminary SFT and simplifying conventional alignment components. Our analysis examines whether models produced by these efficiency-oriented pipelines exhibit analogous security and structural signatures.



(a) Answer likelihood (b) Confidence

Figure 6: MIA for 7B math reasoning LLMs.

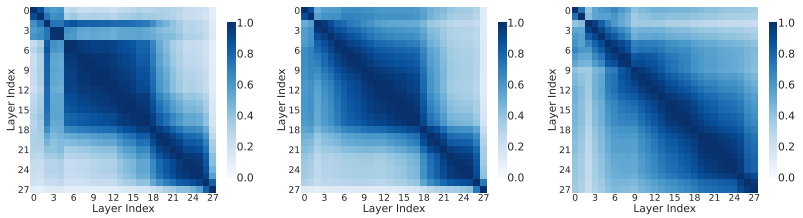
Table 5: Robustness and AutoDAN jailbreak results for 7B math reasoning LLMs. Higher accuracy indicates greater robustness; fewer steps and higher ASR indicate greater vulnerability.

Model	Robustness		AutoDAN	
	Base	Math	Steps	ASR
Instruct	99.50	89.30	4.32	91.92%
SimpleRL	98.70	82.00	0.53	100.00%
PRIME	98.20	82.30	0.41	99.04%

To investigate this question, we conduct a comparative analysis between Qwen2.5-Math-7B-Instruct and two efficiency-oriented variants, Qwen2.5-7B-SimpleRL-Zoo and Eurus-2-7B-PRIME, obtained from Qwen2.5-Math-7B-Base through their respective zero-RL alignment pipelines. As shown in Figure 6, while the Instruct baseline exhibits greater leakage under answer likelihood at stringent low-FPR thresholds, this metric is sensitive to exact ground-truth phrasing and can therefore mask the vulnerability of the RL model, whose generated trajectories need not follow the static reference text after reward optimization. Self-confidence reveals the broader risk: the efficient RL variant is more susceptible than the baseline and exhibits pronounced overconfidence in its own generated trajectories. This reversal demonstrates that the efficient RL model is more vulnerable overall, with its leakage manifested primarily through overconfidence in generated reasoning rather than verbatim memorization. Security differences also emerge in model integrity, as shown in Table 5. Although all three models perform comparably on the clean RobustBase dataset, the zero-RL models lose more accuracy on the adversarial RobustMath benchmark. AutoDAN on AdvBench provides a complementary jailbreak evaluation: SimpleRL and PRIME achieve higher attack success rates and require substantially fewer attack steps than the Instruct model. The clearest disparity appears under out-of-distribution fine-tuning on IMDB (Table 6). The Instruct model largely retains its mathematical-reasoning performance, whereas both zero-RL models lose more reasoning accuracy, with catastrophic forgetting most pronounced for PRIME. Its average score falls from 53.7 to 20.8, indicating that its specialized capabilities are more readily overwritten in this setting. We additionally evaluate privacy using the Min-k% probability attack and extend the Instruct–SimpleRL comparison

Table 6: Fine-tuning stability attack on IMDB dataset.

Dataset		GSM8K	mawps	Minerva Math	MATH 500	Olympiad Bench	AIME24	AMC23	Avg.
Qwen2.5-Math-7B-Instruct	Before	95.38	98.60	31.62	77.20	33.93	3.33	52.50	56.08
	After	94.77	98.30	27.21	78.00	32.89	10.00	45.00	55.17
Qwen2.5-7B-SimpleRL-Zoo	Before	93.40	97.09	25.74	75.40	30.37	10.00	45.00	53.86
	After	81.58	84.99	18.75	68.80	29.33	6.67	42.50	47.52
Eurus-2-7B-PRIME	Before	89.69	96.22	25.37	70.60	32.00	6.67	55.00	53.65
	After	41.77	33.22	12.87	24.40	8.15	0.00	25.00	20.77



(a) Instruct (b) SimpleRL (c) Prime

Figure 7: Intra-CKA on GSM8K.

to 32B, observing similar patterns across membership inference, adversarial robustness, jailbreak susceptibility, and fine-tuning stability; complete results are reported in Appendix B.2.

Table 7: Sharpness and ECE for 7B math reasoning LLMs.

Model	Sharpness		ECE	
	GSM8K	MATH 500	GSM8K	MATH 500
Instruct	0.0615	0.0398	0.0375	0.4006
SimpleRL	0.9120	0.5183	0.0899	0.4453
PRIME	0.5565	0.5032	0.1073	0.4492

To elucidate the structural basis of the observed vulnerabilities, we shift from behavioral outputs to the models' internal geometry, calibration, and functional representations. Table 7 reveals a pronounced geometric divergence: the Instruct model exhibits substantially flatter local geometry, whereas the zero-RL models have markedly higher sharpness. Explicit KL regularization ordinarily constrains policy deviation from a reference model; its absence from the evaluated zero-RL pipelines therefore offers a plausible optimization-level explanation for this sharper geometry. The efficient models also exhibit consistently higher ECE, indicating greater miscalibration and overconfidence. Functionally, the Intra-CKA analysis in Figure 7 reveals stronger adjacent-layer similarity earlier in the zero-RL models, suggesting a more compressed and less differentiated progression of representation processing. The same pattern appears on 500 examples sampled from the MATH test set in Appendix Figure 12. These findings demonstrate that efficiency gains in the evaluated LLM alignment pipelines are accompanied by broader security vulnerability, with sharper local geometry and less differentiated representations providing a consistent structural explanation.

7 Conclusion

In this work, we presented the first systematic evaluation of the hidden security costs associated with efficiency-driven foundation model development. Across the evaluated vision and language settings, we show that strategies used to improve training efficiency, including importance-based data selection, pre-training, and simplified alignment, can compromise robustness and privacy even when task utility is preserved.

Our results reveal a recurring "no free lunch" trade-off: across the strategies studied, lower training cost is repeatedly accompanied by greater exposure to privacy leakage and adversarial attacks. The mechanistic analyses connect this pattern to sharper local loss geometry, greater miscalibration, and systematic changes in functional representations. These findings establish that efficiency gains can carry consequential security costs. As models continue to scale, we advocate for "Robust Efficiency", a paradigm shift toward training methods that integrate stability constraints to ensure models are both cheap and secure.

REFERENCES

https://www.cbsnews.com/news/what-is-deepseek-ai-china-stock-nvidia-nvda-asml/.

Alessio Ansuini, Alessandro Laio, Jakob H. Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 6109–6119. NeurIPS, 2019.

Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron C. Courville, Yoshua Bengio, and Simon Lacoste-Julien. A Closer Look at Memorization in Deep Networks. In *International Conference on Machine Learning (ICML)*, pp. 233–242. PMLR, 2017.

Ergute Bao, Yangfan Jiang, Fei Wei, Xiaokui Xiao, Zitao Li, Yaliang Li, and Bolin Ding. Unlocking the Power of Differentially Private Zeroth-order Optimization for Fine-tuning LLMs. In *USENIX Security Symposium (USENIX Security)*, pp. 1569–1588. USENIX, 2025.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep Reinforcement Learning from Human Preferences. In *Annual Conference on Neural Information Processing Systems (NIPS)*, pp. 4299–4307. NIPS, 2017.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Yuchen Zhang, Jiacheng Chen, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, Jiarui Yuan, Huayu Chen, Kaiyan Zhang, Xingtai Lv, Shuo Wang, Yuan Yao, Xu Han, Hao Peng, Yu Cheng, Zhiyuan Liu, Maosong Sun, Bowen Zhou, and Ning Ding. Process Reinforcement through Implicit Rewards. *Transactions on Machine Learning Research*, 2026.

DeepSeek-AI. DeepSeek-V3 Technical Report. *CoRR abs/2412.19437*, 2024.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 4171–4186. ACL, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Kazuki Egashira, Robin Staab, Thibaud Gloaguen, Mark Vero, and Martin T. Vechev. Fewer Weights, More Problems: A Practical Attack on LLM Pruning. *CoRR abs/2510.07985*, 2025.

Vitaly Feldman. Does learning require memorization? a short tale about a long tail. In *Annual ACM Symposium on Theory of Computing (STOC)*, pp. 954–959. ACM, 2020.

Wenjie Fu, Huandong Wang, Chen Gao, Guanghua Liu, Yong Li, and Tao Jiang. Membership Inference Attacks against Fine-tuned Large Language Models via Self-prompt Calibration. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*. NeurIPS, 2024.

Yunjie Ge, Qian Wang, Baolin Zheng, Xinlu Zhuang, Qi Li, Chao Shen, and Cong Wang. Anti-Distillation Backdoor Attacks: Backdoors Can Really Survive in Knowledge Distillation. In *ACM International Conference on Multimedia (MM)*, pp. 826–834. ACM, 2021.

Amirata Ghorbani and James Y. Zou. Data Shapley: Equitable Valuation of Data for Machine Learning. In *International Conference on Machine Learning (ICML)*, pp. 2242–2251. PMLR, 2019.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. In *International Conference on Machine Learning (ICML)*. PMLR, 2017.

Junyuan Hong, Jinhao Duan, Chenhui Zhang, Zhangheng Li, Chulin Xie, Kelsey Lieberman, James Diffenderfer, Brian R. Bartoldson, Ajay Kumar Jaiswal, Kaidi Xu, Bhavya Kailkhura, Dan Hendrycks, Dawn Song, Zhangyang Wang, and Bo Li. Decoding Compressed Trust: Scrutinizing the Trustworthiness of Efficient LLMs Under Compression. In *International Conference on Machine Learning (ICML)*, pp. 18611–18633. PMLR, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations (ICLR)*, 2022.

Yixin Ji, Yang Xiang, Juntao Li, Qingrong Xia, Ping Li, Xinyu Duan, Zhefeng Wang, and Min Zhang. Beware of Calibration Data for Pruning Large Language Models. In *International Conference on Learning Representations (ICLR)*, 2025.

Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nezihe Merve Gürel, Bo Li, Ce Zhang, Costas J. Spanos, and Dawn Song. Efficient Task-Specific Data Valuation for Nearest Neighbor Algorithms. *Proceedings of the VLDB Endowment*, 2019.

Kimi-Team. Kimi K2: Open Agentic Intelligence. *CoRR abs/2507.20534*, 2025.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey E. Hinton. Similarity of Neural Network Representations Revisited. In *International Conference on Machine Learning (ICML)*, pp. 3519–3529. PMLR, 2019.

Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. Fine-Tuning can Distort Pretrained Features and Underperform Out-of-Distribution. In *International Conference on Learning Representations (ICLR)*, 2022.

Xuechen Li, Florian Tramèr, Percy Liang, and Tatsunori Hashimoto. Large Language Models Can Be Strong Differentially Private Learners. In *International Conference on Learning Representations (ICLR)*, 2022.

Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. AutoDAN: Generating Stealthy Jailbreak Prompts on Aligned Large Language Models. In *International Conference on Learning Representations (ICLR)*, 2024.

Kaifeng Lyu and Jian Li. Gradient Descent Maximizes the Margin of Homogeneous Neural Networks. In *International Conference on Learning Representations (ICLR)*, 2020.

Wesley J. Maddox, Gregory W. Benton, and Andrew Gordon Wilson. Rethinking Parameter Counting in Deep Models: Effective Dimensionality Revisited. *CoRR abs/2003.02139*, 2020.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations (ICLR)*, 2018.

Giovanni De Muri, Mark Vero, Robin Staab, and Martin T. Vechev. Pay Attention to the Triggers: Constructing Backdoors That Survive Distillation. In *International Conference on Learning Representations (ICLR)*, 2026.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*. NeurIPS, 2022.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep Learning on a Data Diet: Finding Important Examples Early. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 20596–20607. NeurIPS, 2021.

Guillermo Valle Pérez, Chico Q. Camargo, and Ard A. Louis. Deep learning generalizes because the parameter-function map is biased towards simple functions. In *International Conference on Learning Representations (ICLR)*, 2019.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*. NeurIPS, 2023.

Pranav Ramesh, Arjun Roy, Deepak Ravikumar, Kaushik Roy, and Gopalakrishnan Srinivasan. The Easy Path to Robustness: Coreset Selection using Sample Hardness. *CoRR abs/2510.11018*, 2025.

Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models. In *Network and Distributed System Security Symposium (NDSS)*. Internet Society, 2019.

Liwei Song and Prateek Mittal. Systematic Evaluation of Privacy Risks of Machine Learning Models. In *USENIX Security Symposium (USENIX Security)*. USENIX, 2021.

Liwei Song, Reza Shokri, and Prateek Mittal. Privacy Risks of Securing Machine Learning Models against Adversarial Examples. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 241–257. ACM, 2019.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, and Nathan Srebro. The Implicit Bias of Gradient Descent on Separable Data. In *International Conference on Learning Representations (ICLR)*, 2018.

Rui Wen, Michael Backes, and Yang Zhang. Understanding Data Importance in Machine Learning Attacks: Does Valuable Data Pose Greater Harm? In *Network and Distributed System Security Symposium (NDSS)*. Internet Society, 2025.

David H. Wolpert and William G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1997.

Shaokai Ye, Xue Lin, Kaidi Xu, Sijia Liu, Hao Cheng, Jan-Henrik Lambrechts, Huan Zhang, Aojun Zhou, Kaisheng Ma, and Yanzhi Wang. Adversarial Robustness vs. Model Compression, or Both? In *IEEE International Conference on Computer Vision (ICCV)*, pp. 111–120. IEEE, 2019.

Dong Yin, Ashwin Pananjady, Maximilian Lam, Dimitris S. Papailiopoulos, Kannan Ramchandran, and Peter L. Bartlett. Gradient Diversity: a Key Ingredient for Scalable Distributed Learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1998–2007. PMLR, 2018.

Dongkeun Yoon, Seungone Kim, Sohee Yang, Sunkyoung Kim, Soyeon Kim, Yongil Kim, Eunbi Choi, Yireun Kim, and Minjoon Seo. Reasoning Models Better Express Their Confidence. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*. NeurIPS, 2025.

Chong Yu, Tao Chen, Zhongxue Gan, and Jiayuan Fan. Spear: Evaluate the Adversarial Robustness of Compressed Neural Models. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1598–1606. IJCAI, 2024.

Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. SimpleRL-Zoo: Investigating and Taming Zero Reinforcement Learning for Open Base Models in the Wild. *CoRR abs/2503.18892*, 2025.

Zhi Zhang, Qizhe Zhang, Zijun Gao, Renrui Zhang, Ekaterina Shutova, Shiji Zhou, and Shanghang Zhang. Gradient-based Parameter Selection for Efficient Fine-Tuning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28566–28577. IEEE, 2024.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. LIMA: Less Is More for Alignment. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*. NeurIPS, 2023.

Zihao Zhou, Qiufeng Wang, Mingyu Jin, Jie Yao, Jianan Ye, Wei Liu, Wei Wang, Xiaowei Huang, and Kaizhu Huang. MathAttack: Attacking Large Language Models towards Math Solving Ability. In *AAAI Conference on Artificial Intelligence (AAAI)*, pp. 19750–19758. AAAI, 2024.

Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. Universal and Transferable Adversarial Attacks on Aligned Language Models. *CoRR abs/2307.15043*, 2023.

ETHICS STATEMENT

The primary ethical consideration of this work is disclosure risk: showing that the evaluated efficiency-trained models are more vulnerable could, in principle, help an adversary prioritize targets. We mitigate this risk by relying exclusively on established, publicly documented attack methodologies (membership inference, adversarial perturbations, model stealing, and jailbreak prompts) applied to publicly released models, so the study expands no offensive capability, and we disclose the findings so that practitioners adopting these increasingly popular strategies can account for their hidden security costs before deployment. All experiments use publicly available datasets and model checkpoints, including the established PubFig83 benchmark of public-figure face images; no new personal data were collected, no human subjects were involved, and no privately deployed system was attacked.

REPRODUCIBILITY STATEMENT

The evaluated architectures, datasets, and public model checkpoints are named in the main text, the utility-matched comparison protocol is specified in Section 4.1, and complete training, attack, and measurement configurations are given in Section C. Vision results are averaged over three seeds, with mean and standard deviation reported in the appendix. Upon acceptance, we will publicly release the code and evaluation scripts that reproduce all reported tables and figures.

THE USE OF LARGE LANGUAGE MODELS

We used generative AI tools to polish the writing for readability and consistency. All AI-assisted text was reviewed and verified by the authors, who take full responsibility for the final content of this work.

A DETAILS OF THEORY

A.1 LOSS LANDSCAPE SHARPNESS

Let $\mathcal{U}_\rho = \{\delta : \|\delta_j\|_2 \le \rho \text{ for every parameter tensor } j\}$ denote the tensor-wise perturbation set used by the procedure in Appendix C. For batch b, define $g_b = \nabla_\theta \mathcal{L}_b(\theta)$ and $\mathcal{H}_b = \nabla^2_\theta \mathcal{L}_b(\theta)$. A local Taylor expansion gives

$$\mathcal{L}_b(\theta + \delta) - \mathcal{L}_b(\theta) = g_b^\top \delta + \frac{1}{2}\delta^\top \mathcal{H}_b \delta + o(\|\delta\|_2^2). \tag{5}$$

Thus, the measured loss increase reflects both the local gradient and curvature. In the stationary local-minimum case, where $g_b = 0$ and $\mathcal{H}_b \succeq 0$, maximizing over a single global L_2 ball yields $\frac{1}{2}\rho^2 \lambda_{\max}(\mathcal{H}_b) + o(\rho^2)$; the tensor-wise constraint used in our experiments defines a different, implementation-aligned neighborhood. Projected gradient ascent approximates rather than exactly solves this maximization, so $\widehat{S}$ is an empirical measure of local loss sensitivity rather than an exact Hessian-eigenvalue estimate.

A.2 EXPECTED CALIBRATION ERROR (ECE)

For vision classification, we provide an idealized margin-based interpretation of the ECE behavior rather than a guarantee that an efficiency strategy must increase miscalibration. For a sample x_i, let z_{ik} be the logit for class k and $\hat{y}_i = \arg\max_k z_{ik}$ the predicted class. Defining the pairwise margins $\Delta_{ij} = z_{i\hat{y}_i} - z_{ij}$ for $j \neq \hat{y}_i$, the corresponding softmax confidence is

$$c_i = \frac{1}{1 + \sum_{j \neq \hat{y}_i} \exp(-\Delta_{ij})}. \tag{6}$$

As long as the predicted class is unchanged, confidence increases monotonically with every margin because

$$\frac{\partial c_i}{\partial \Delta_{ij}} = c_i^2 \exp(-\Delta_{ij}) > 0. \tag{7}$$

For separable data and positively homogeneous models, gradient-based optimization of exponential-tail losses can continue increasing parameter norms and classification margins even after the training error vanishes Soudry et al. (2018); Lyu & Li (2020). This asymptotic result provides theoretical intuition for increasingly extreme confidence, but does not by itself imply poorer calibration.

The connection to ECE becomes precise when confidence grows without a corresponding improvement in correctness. For an overconfident bin B_m whose membership and accuracy remain fixed locally, its ECE contribution is

$$\mathrm{ECE}_m = \frac{|B_m|}{N}\left(\mathrm{conf}(B_m) - \mathrm{acc}(B_m)\right), \qquad \frac{\partial \mathrm{ECE}_m}{\partial\, \mathrm{conf}(B_m)} = \frac{|B_m|}{N} > 0. \tag{8}$$

Thus, margin growth enlarges the contribution of an already overconfident bin when its accuracy does not improve. This local argument does not require ECE to increase globally: predictions may move between bins, and greater confidence can reduce the gap in underconfident bins. It instead motivates our hypothesis that high-influence data selection and limited-data fine-tuning can aggravate miscalibration when their confidence shifts outpace changes in correctness, as assessed empirically in Table 4. For LLMs, we retain the same bin-based ECE definition but use the sequence-level confidence specified in Appendix C; the classification-margin derivation above is not applied to that confidence measure.

A.3 Feature Dimensionality (PCA)

Let $f(x_i) \in \mathbb{R}^d$ denote the penultimate-layer feature of sample x_i and $\boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^{N} f(x_i)$ their empirical mean. We compute the covariance $\mathbf{C} = \frac{1}{N-1}\sum_{i=1}^{N}(f(x_i) - \boldsymbol{\mu})(f(x_i) - \boldsymbol{\mu})^\top$. For its eigenvalues λ_i, define the normalized spectrum $\widetilde{\lambda}_i = \lambda_i / \sum_{j=1}^{d} \lambda_j$. The PR defined in Section 5.2 is then equivalently $\mathrm{PR}(\mathbf{C}) = (\sum_{i=1}^{d} \widetilde{\lambda}_i^2)^{-1}$. We use PR as a continuous estimate of the effective rank of the feature covariance: if variance is evenly distributed across k directions, $d_{\text{eff}} \approx k$; if a single direction dominates, $d_{\text{eff}} \approx 1$.

For a naive (scratch-initialized) attacker without a shared representation prior, the target's effective feature dimension characterizes how many independent directions must be approximated. High-Shapley selection preferentially retains easy, representative samples sufficient for task utility while omitting harder intra-class variations. If these omitted variations contribute primarily to the tail of the covariance spectrum, their removal concentrates variance in the leading eigenvalues and lowers PR. From an information-bottleneck perspective, the resulting representation encodes less within-class variation while preserving the information needed to match measured utility. Its lower effective dimension therefore provides a natural explanation for the higher stolen-model accuracy and agreement observed under naive extraction. This is an effective-complexity account of the empirical pattern rather than a universal linear query-complexity law.

The pre-training setting produces the opposite spectral pattern and requires a different explanation. Let θ_0 denote the public pre-trained initialization and write the fine-tuned target as $\theta_T = \theta_0 + \Delta_T$. A scratch-initialized attacker must learn both a useful representation and the target's task-specific behavior, whereas an informed surrogate initialized at θ_0 already inherits the shared prior and primarily needs to recover the adaptation Δ_T. This does not reduce extraction to learning only a linear head, but it narrows the empirical gap between the surrogate and target. The higher PR of the pre-trained variants is therefore consistent with their resistance to scratch-initialized attacks, while the shared initialization explains why this protection weakens under informed extraction.

A.4 Representation Similarity (CKA)

The CKA metric relies on the Hilbert–Schmidt Independence Criterion (HSIC) to measure statistical dependence between layer representations. Given activation matrices $X \in \mathbb{R}^{n \times d_1}$ and $Y \in \mathbb{R}^{n \times d_2}$, we compute the linear-kernel Gram matrices $K = XX^\top$ and $L = YY^\top$. Let $H = I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ be the centering matrix and define $K_c = HKH$ and $L_c = HLH$. The empirical HSIC is

$$\mathrm{HSIC}(K, L) = \frac{1}{(n-1)^2}\,\mathrm{Tr}(K_c L_c) = \frac{1}{(n-1)^2}\,\mathrm{Tr}(KHLH). \tag{9}$$

This value quantifies alignment between the inter-sample similarity structures of the two layers. For linear kernels, a value near 0 indicates little alignment between the centered Gram matrices, while higher values indicate stronger structural similarity. Normalizing HSIC yields a scalar in $[0, 1]$ that is invariant to orthogonal transformations and isotropic scaling. Consequently, high CKA does not establish that one layer is a literal identity mapping of another; it shows that the layers preserve similar representational geometry.

We interpret extensive late-layer similarity through simplicity bias, the tendency of neural networks to rely on the simplest predictive features available Arpit et al. (2017); Pérez et al. (2019). High-Shapley selection emphasizes easy, representative samples while omitting harder intra-class variations. Once early features are sufficient to separate this selected distribution, later layers face less pressure to reorganize inter-sample geometry, producing the more homogeneous hierarchy observed in Figure 5. Thus, high late-layer CKA provides evidence of limited additional feature refinement without requiring the stronger claim that successive layers implement exact identity mappings.

B More Results

B.1 Vision Models

Three-Seed Vision Results. This section reports the complete mean $\pm$ standard deviation results corresponding to the mean values presented in the main-text vision tables. Specifically, Appendix Table 8 expands the data-selection accuracy and membership-inference results in Table 1; Appendix Table 9 expands the adversarial-robustness results in Table 2; Appendix Table 10 expands the pre-trained fine-tuning results in Table 3; and Appendix Table 11 expands the sharpness, ECE, and PCA PR results in Table 4.

Comparison with Full-Dataset Training. Our primary data-selection study compares a small High-Shapley subset with a larger Low-Shapley subset calibrated to achieve similar clean accuracy. This utility-matched design evaluates vulnerability while controlling for predictive performance. Here, we provide a complementary comparison against conventional Full-dataset training. Unlike the primary study, these comparisons are not necessarily utility matched; they instead evaluate the practical security behavior of High-Shapley subsets at two different sizes relative to using all available training data.

Table 12 reports clean accuracy and membership-inference results. Across all evaluated datasets and architectures, the High-Shapley subsets exhibit higher membership-inference TPRs than Full-dataset training under confidence, entropy, and modified-entropy attacks. The differences remain present for the larger High-Shapley subsets, including settings in which their clean accuracy approaches that of Full-dataset training.

Table 13 shows the same behavioral pattern for adversarial robustness and model stealing. Models trained on High-Shapley subsets remain correctly classified for fewer PGD iterates than their Full-data counterparts. Using the same number of distinct query images, the corresponding stolen models also attain consistently higher accuracy and agreement, indicating greater extraction susceptibility in these evaluated settings.

Table 14 provides the corresponding structural measurements. ECE is higher for every High-Shapley subset than for Full-dataset training. Sharpness is generally higher for the High-Shapley models, although the difference becomes small for some larger subsets and reverses for PubFig83 with 3,000 samples. PCA PR is lower for the CIFAR-10 and TinyImageNet High-Shapley models, whereas the PubFig83 results are mixed. We therefore treat these measurements as complementary evidence rather than a uniform structural pattern across every dataset and subset size.

Sensitivity to Weight Pruning. We additionally evaluate sensitivity to 50% unstructured weight pruning in two representative vision settings. As shown in Table 15, the High-Shapley models and their utility-matched Low-Shapley controls achieve comparable accuracy before pruning. After pruning, however, the High-Shapley models experience substantially larger performance degradation. This result provides supplementary evidence consistent with reduced functional redundancy in models trained on small, high-importance subsets.

Table 8: Accuracy (training accuracy in brackets) and TPR at 0.1% FPR of membership inference for data selection with Shapley values (mean ± standard deviation over three seeds).

Dataset	Value	Acc	MIA (conf)	MIA (entr)	MIA (modi)
CIFAR-10	Low	48.07 ± 0.47 (99.97 ± 0.06)	$0.12\% \pm 0.01\%$	$0.05\% \pm 0.01\%$	$0.11\% \pm 0.01\%$
	High	47.97 ± 0.58 (99.97 ± 0.05)	$0.16\% \pm 0.02\%$	$0.09\% \pm 0.01\%$	$0.16\% \pm 0.01\%$
CIFAR-10 (swin)	Low	37.01 ± 0.29 (99.95 ± 0.08)	$0.08\% \pm 0.01\%$	$0.05\% \pm 0.01\%$	$0.09\% \pm 0.01\%$
	High	36.94 ± 0.26 (99.93 ± 0.07)	$0.29\% \pm 0.02\%$	$0.30\% \pm 0.02\%$	$0.30\% \pm 0.01\%$
TinyImageNet	Low	36.09 ± 0.58 (100.00 ± 0.00)	$0.08\% \pm 0.02\%$	$0.06\% \pm 0.01\%$	$0.10\% \pm 0.02\%$
	High	36.17 ± 0.46 (100.00 ± 0.00)	$0.22\% \pm 0.01\%$	$0.35\% \pm 0.02\%$	$0.27\% \pm 0.02\%$
PubFig83	Low	44.01 ± 0.55 (100.00 ± 0.00)	$0.32\% \pm 0.03\%$	$0.32\% \pm 0.03\%$	$0.32\% \pm 0.03\%$
	High	44.26 ± 0.57 (100.00 ± 0.00)	$1.51\% \pm 0.06\%$	$1.43\% \pm 0.07\%$	$1.51\% \pm 0.06\%$

Table 9: Adversarial robustness for vision models (mean ± standard deviation over three seeds).

Dataset	Value	Robustness	Paradigm	Robustness
CIFAR-10	Low	3.85 ± 0.21	Scratch	3.58 ± 0.58
	High	1.33 ± 0.07	Pre-train	1.00 ± 0.05
CIFAR-10 (swin)	Low	0.70 ± 0.13	Scratch	3.22 ± 0.11
	High	0.18 ± 0.03	Pre-train	0.11 ± 0.03
TinyImageNet	Low	20.89 ± 2.00	Scratch	14.03 ± 1.70
	High	10.46 ± 1.21	Pre-train	3.64 ± 0.13
PubFig83	Low	2.34 ± 0.25	Scratch	1.83 ± 0.11
	High	1.30 ± 0.08	Pre-train	0.41 ± 0.07

Table 10: Accuracy (training accuracy in brackets), membership inference, and model stealing for pre-trained fine-tuning and training from scratch (mean ± standard deviation over three seeds). Values in parentheses under MS denote agreement.

Dataset	Paradigm	Acc	MIA			MS	
			conf	entr	modi	1000	1000 (Pre-train)
CIFAR-10	Scratch	67.71 ± 0.36 (99.91 ± 0.09)	$0.05\% \pm 0.03\%$	$0.04\% \pm 0.01\%$	$0.05\% \pm 0.02\%$	62.60 ± 0.27 (0.748 ± 0.005)	65.22 ± 0.20 (0.758 ± 0.004)
	Pre-train	67.88 ± 0.24 (100.00 ± 0.00)	$0.21\% \pm 0.01\%$	$0.16\% \pm 0.01\%$	$0.21\% \pm 0.00\%$	58.05 ± 0.16 (0.669 ± 0.007)	67.25 ± 0.14 (0.785 ± 0.003)
CIFAR-10 (swin)	Scratch	73.85 ± 0.35 (100.00 ± 0.00)	$0.12\% \pm 0.01\%$	$0.12\% \pm 0.01\%$	$0.11\% \pm 0.01\%$	42.41 ± 0.42 (0.446 ± 0.006)	57.70 ± 0.52 (0.536 ± 0.009)
	Pre-train	74.01 ± 0.14 (100.00 ± 0.00)	$0.54\% \pm 0.02\%$	$0.54\% \pm 0.02\%$	$0.54\% \pm 0.02\%$	39.01 ± 0.24 (0.437 ± 0.010)	60.06 ± 0.19 (0.651 ± 0.006)
TinyImageNet	Scratch	34.91 ± 0.75 (99.98 ± 0.02)	$0.13\% \pm 0.01\%$	$0.07\% \pm 0.03\%$	$0.14\% \pm 0.03\%$	24.65 ± 0.45 (0.381 ± 0.004)	27.53 ± 0.43 (0.379 ± 0.008)
	Pre-train	34.93 ± 0.38 (100.00 ± 0.00)	$0.75\% \pm 0.01\%$	$0.75\% \pm 0.01\%$	$0.67\% \pm 0.01\%$	17.65 ± 0.43 (0.305 ± 0.006)	29.82 ± 0.73 (0.529 ± 0.010)
PubFig83	Scratch	35.19 ± 0.25 (100.00 ± 0.00)	$0.01\% \pm 0.01\%$	$0.01\% \pm 0.01\%$	$0.01\% \pm 0.01\%$	17.81 ± 0.73 (0.358 ± 0.008)	18.59 ± 0.40 (0.299 ± 0.001)
	Pre-train	34.79 ± 0.33 (100.00 ± 0.00)	$0.36\% \pm 0.01\%$	$0.46\% \pm 0.02\%$	$0.37\% \pm 0.01\%$	7.89 ± 0.50 (0.183 ± 0.006)	25.26 ± 0.16 (0.533 ± 0.010)

Table 11: Sharpness, Expected Calibration Error (ECE), and PCA Participation Ratio (PCA PR) for vision models (mean ± standard deviation over three seeds).

Dataset	Value	Sharpness	ECE	PCA PR	Paradigm	Sharpness	ECE	PCA PR
CIFAR-10	Low	0.0410 ± 0.0146	0.382 ± 0.012	28.490 ± 7.284	Scratch	0.0311 ± 0.0116	0.219 ± 0.005	14.266 ± 0.160
	High	0.0591 ± 0.0141	0.430 ± 0.014	10.793 ± 0.727	Pre-train	0.1332 ± 0.0088	0.232 ± 0.001	25.137 ± 0.962
CIFAR-10 (swin)	Low	0.0170 ± 0.0005	0.410 ± 0.009	13.907 ± 0.861	Scratch	0.0010 ± 0.0002	0.060 ± 0.004	9.897 ± 1.008
	High	0.0189 ± 0.0015	0.436 ± 0.008	10.896 ± 1.633	Pre-train	0.0373 ± 0.0056	0.241 ± 0.036	26.786 ± 0.668
TinyImageNet	Low	0.0548 ± 0.0025	0.336 ± 0.021	265.298 ± 13.393	Scratch	0.0511 ± 0.0017	0.317 ± 0.002	157.477 ± 0.655
	High	0.0600 ± 0.0028	0.384 ± 0.019	74.960 ± 3.202	Pre-train	0.1201 ± 0.0033	0.251 ± 0.008	163.206 ± 1.082
PubFig83	Low	0.1234 ± 0.0054	0.208 ± 0.017	48.395 ± 1.200	Scratch	0.1349 ± 0.0052	0.350 ± 0.009	31.940 ± 0.914
	High	0.1475 ± 0.0075	0.299 ± 0.014	33.936 ± 2.311	Pre-train	0.2551 ± 0.0061	0.384 ± 0.008	39.009 ± 0.107

Table 12: Accuracy (training accuracy in brackets) and membership inference for Full-dataset training and High-Shapley subsets (mean ± standard deviation over three seeds). MIA is reported as TPR at 0.1% FPR; higher values indicate greater privacy leakage.

Dataset	Value	Acc	MIA (conf)	MIA (entr)	MIA (modi)
CIFAR-10	Full	94.84 ± 0.32 (100.00 ± 0.00)	$0.02\% \pm 0.00\%$	$0.01\% \pm 0.00\%$	$0.01\% \pm 0.00\%$
	High (20,000)	88.54 ± 0.25 (100.00 ± 0.00)	$0.07\% \pm 0.00\%$	$0.06\% \pm 0.00\%$	$0.07\% \pm 0.01\%$
	High (40,000)	93.38 ± 0.27 (100.00 ± 0.00)	$0.04\% \pm 0.01\%$	$0.02\% \pm 0.00\%$	$0.04\% \pm 0.00\%$
CIFAR-10 (swin)	Full	79.40 ± 0.31 (100.00 ± 0.00)	$0.05\% \pm 0.00\%$	$0.00\% \pm 0.00\%$	$0.05\% \pm 0.00\%$
	High (20,000)	69.86 ± 0.29 (100.00 ± 0.00)	$0.20\% \pm 0.01\%$	$0.28\% \pm 0.01\%$	$0.20\% \pm 0.01\%$
	High (40,000)	77.91 ± 0.33 (100.00 ± 0.00)	$0.18\% \pm 0.01\%$	$0.22\% \pm 0.00\%$	$0.18\% \pm 0.01\%$
TinyImageNet	Full	59.82 ± 0.25 (100.00 ± 0.00)	$0.00\% \pm 0.00\%$	$0.00\% \pm 0.00\%$	$0.00\% \pm 0.00\%$
	High (60,000)	58.01 ± 0.20 (100.00 ± 0.00)	$0.21\% \pm 0.02\%$	$0.12\% \pm 0.02\%$	$0.21\% \pm 0.02\%$
	High (80,000)	60.19 ± 0.22 (100.00 ± 0.00)	$0.22\% \pm 0.01\%$	$0.12\% \pm 0.01\%$	$0.21\% \pm 0.01\%$
PubFig83	Full	72.41 ± 0.42 (100.00 ± 0.00)	$0.00\% \pm 0.00\%$	$0.00\% \pm 0.00\%$	$0.00\% \pm 0.00\%$
	High (3,000)	61.45 ± 0.44 (100.00 ± 0.00)	$0.12\% \pm 0.02\%$	$0.12\% \pm 0.02\%$	$0.12\% \pm 0.02\%$
	High (5,000)	65.66 ± 0.45 (100.00 ± 0.00)	$0.73\% \pm 0.05\%$	$0.82\% \pm 0.06\%$	$1.90\% \pm 0.09\%$

Table 13: Adversarial robustness and model stealing using 1,000 distinct query images (500 for PubFig83) for Full-dataset training and High-Shapley subsets (mean ± standard deviation over three seeds). Higher robustness is better. MS reports stolen-model accuracy, with agreement in parentheses; higher values indicate a more successful stealing attack.

Dataset	Value	Robustness	MS
CIFAR-10	Full	18.60 ± 0.92	59.25 ± 0.23 (0.599 ± 0.002)
	High (20,000)	10.03 ± 0.88	67.05 ± 0.29 (0.712 ± 0.003)
	High (40,000)	17.60 ± 0.82	62.01 ± 0.26 (0.634 ± 0.002)
CIFAR-10 (swin)	Full	5.68 ± 0.22	41.20 ± 0.20 (0.429 ± 0.003)
	High (20,000)	2.57 ± 0.11	45.56 ± 0.21 (0.499 ± 0.004)
	High (40,000)	4.85 ± 0.19	43.10 ± 0.19 (0.458 ± 0.003)
TinyImageNet	Full	36.93 ± 1.62	19.39 ± 0.11 (0.215 ± 0.006)
	High (60,000)	29.77 ± 2.11	28.66 ± 0.12 (0.362 ± 0.009)
	High (80,000)	34.85 ± 1.73	22.93 ± 0.18 (0.267 ± 0.007)
PubFig83	Full	3.32 ± 0.11	37.71 ± 0.39 (0.396 ± 0.009)
	High (3,000)	1.27 ± 0.09	44.70 ± 0.35 (0.547 ± 0.010)
	High (5,000)	1.67 ± 0.13	42.17 ± 0.41 (0.520 ± 0.012)

Table 14: Sharpness, Expected Calibration Error (ECE), and PCA Participation Ratio (PCA PR) for Full-dataset training and High-Shapley subsets (mean ± standard deviation over three seeds).

Dataset	Value	Sharpness	ECE	PCA PR
CIFAR-10	Full	0.0253 ± 0.0078	0.033 ± 0.003	10.544 ± 0.231
	High (20,000)	0.0313 ± 0.0081	0.076 ± 0.004	9.586 ± 0.292
	High (40,000)	0.0265 ± 0.0077	0.042 ± 0.003	9.841 ± 0.221
CIFAR-10 (swin)	Full	0.0128 ± 0.0004	0.160 ± 0.002	9.048 ± 0.201
	High (20,000)	0.0152 ± 0.0006	0.245 ± 0.003	8.424 ± 0.102
	High (40,000)	0.0142 ± 0.0003	0.185 ± 0.002	8.316 ± 0.162
TinyImageNet	Full	0.0440 ± 0.0020	0.200 ± 0.002	221.403 ± 1.982
	High (60,000)	0.0482 ± 0.0024	0.220 ± 0.005	158.711 ± 2.109
	High (80,000)	0.0458 ± 0.0032	0.209 ± 0.004	193.468 ± 2.187
PubFig83	Full	0.1550 ± 0.0043	0.038 ± 0.009	41.872 ± 1.021
	High (3,000)	0.1296 ± 0.0038	0.116 ± 0.008	41.028 ± 0.987
	High (5,000)	0.1594 ± 0.0041	0.165 ± 0.008	49.366 ± 1.223

Table 15: Accuracy degradation under 50% unstructured weight pruning (mean $\pm$ standard deviation over three seeds). "Drop" denotes the decrease from pre-pruning to post-pruning accuracy; higher values indicate greater pruning sensitivity.

Dataset	Value	Before	After	Drop
CIFAR-10	Low	48.07 ± 0.47	44.83 ± 0.38	3.24 ± 0.60
	High	47.97 ± 0.58	34.12 ± 0.32	13.85 ± 0.66
TinyImageNet	Low	36.09 ± 0.58	34.54 ± 0.44	1.55 ± 0.73
	High	36.17 ± 0.46	18.87 ± 0.61	17.30 ± 0.76

Table 16: Membership inference for the 7B models using the Min-k% probability attack with $k = 15$. Higher TPR and AUC indicate greater membership leakage.

Model	TPR at 0.1% FPR	AUC
Qwen2.5-Math-7B-Instruct	0.89%	0.823
Qwen2.5-7B-SimpleRL-Zoo	1.77%	0.857

B.2 Large Language Models

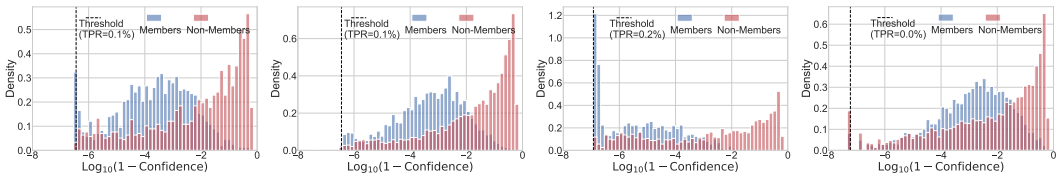
Min-k% Membership Inference on Math-Reasoning LLMs. We complement the answer-likelihood and self-confidence attacks in the main text with the Min-k% probability membership-inference attack, using $k = 15$. For each sequence, this attack selects the 15% of tokens assigned the lowest probabilities by the model and aggregates their log probabilities into a membership score. The attack is motivated by the tendency of models to assign higher probabilities to sequences encountered during training; here, we evaluate it using the source-defined groups specified in Appendix C. We report the area under the receiver operating characteristic curve (AUC) and the true-positive rate (TPR) at a stringent false-positive rate (FPR) of 0.1%.

As shown in Table 16, the 7B SimpleRL checkpoint exhibits higher membership leakage than the Instruct checkpoint: the TPR increases from 0.89% to 1.77%, and the AUC increases from 0.823 to 0.857. This result is consistent with the privacy trend observed using self-confidence and provides additional evidence that the evaluated simplified-RL checkpoint is more susceptible to membership inference. Because this comparison uses existing checkpoints rather than a controlled component-level ablation, we interpret the result as an association with the simplified alignment recipe rather than as evidence that any single omitted training component directly causes the increased leakage.

Sensitivity to Low-Bit Quantization. We evaluate the sensitivity of the 7B math-reasoning models to 4-bit NormalFloat (NF4) quantization as a supplementary deployment-stage stress test. Table 17 reports performance before and after quantization across seven reasoning benchmarks. Individual benchmark scores exhibit some variation in both directions, so we focus on the aggregate performance change rather than requiring uniform degradation on every task.

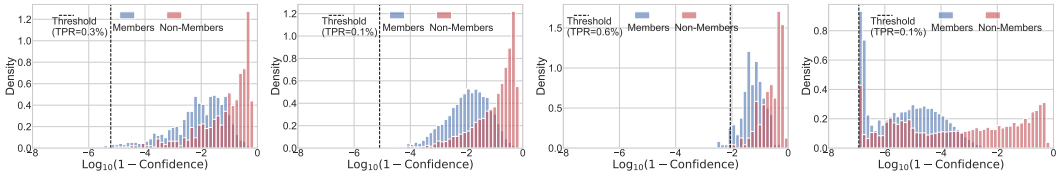
The Instruct model retains nearly all of its average performance after quantization, decreasing from 56.08 to 55.96. In comparison, SimpleRL decreases from 53.86 to 51.78, while PRIME decreases from 53.65 to 47.76. The larger aggregate degradation of the simplified-RL checkpoints provides supplementary evidence consistent with greater sensitivity to parameter perturbations. We treat this result as a quantization-sensitivity observation rather than as a direct security vulnerability or a general claim about all efficiently aligned models.

Evaluation at the 32B Scale. We extend the comparison to Qwen2.5-32B-Instruct and Qwen2.5-32B-SimpleRL-Zoo to examine whether the vulnerability pattern observed at 7B persists at a larger model scale. Table 18 combines confidence-based and Min-k% membership inference, adversarial robustness, and AutoDAN jailbreak results. SimpleRL exhibits higher membership leakage under both attacks. Both models achieve identical accuracy on RobustBase, while SimpleRL performs worse on RobustMath. Under AutoDAN, SimpleRL also attains a higher attack success rate with



(a) High (confidence) (b) Low (confidence) (c) Pre-train (confidence) (d) Scratch (confidence)

Figure 8: Member and non-member confidence distributions on a logarithmic scale for CIFAR-10 with ResNet-18 under data selection and pre-training.



(a) High (confidence) (b) Low (confidence) (c) Pre-train (confidence) (d) Scratch (confidence)

Figure 9: Member and non-member confidence distributions on a logarithmic scale for CIFAR-10 with Swin under data selection and pre-training.

substantially fewer optimization steps. These results follow the same overall pattern as the 7B evaluation across complementary threat models.

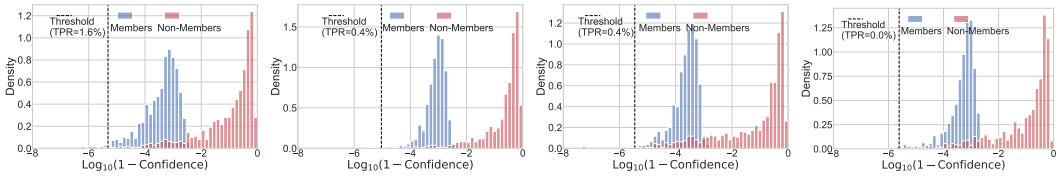
We additionally evaluate stability under out-of-distribution fine-tuning on IMDB. As shown in Table 19, the Instruct checkpoint retains most of its mathematical-reasoning performance after adaptation, whereas SimpleRL exhibits a substantially larger average decline across the seven benchmarks. This larger-scale result is consistent with the greater fine-tuning instability observed for the 7B simplified-RL checkpoints. Because the comparison uses existing checkpoints rather than a controlled training ablation, we interpret the result as an association with the evaluated alignment recipes.

B.3 Additional Analysis

Confidence Distributions. Figures 8, 9, and 10 extend the confidence-distribution analysis in the main text to CIFAR-10 with ResNet-18, CIFAR-10 with Swin, and PubFig83, respectively. Each figure compares the utility-matched High- and Low-Shapley models as well as the pre-trained and scratch-trained models. Across these datasets and architectures, the efficient variants exhibit a stronger concentration toward extreme confidence values than their corresponding baselines. These distributional shifts are consistent with their generally higher ECE and confidence-based membership-inference leakage reported in Tables 11, 8, and 10. Together with the TinyImageNet visualization in Figure 4, these results show that the observed overconfidence is not restricted to a single dataset or architecture.

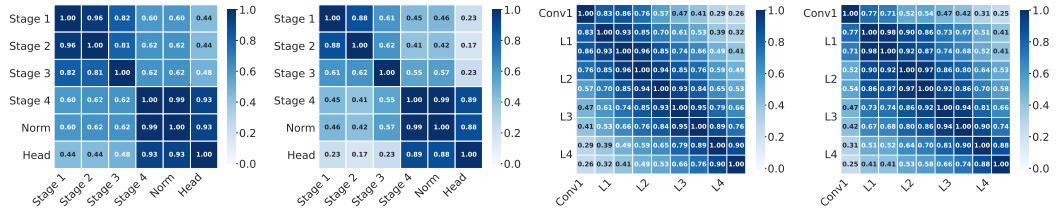
Representation Similarity. Figure 11 presents additional Intra-CKA heatmaps for the High- and Low-Shapley models on CIFAR-10 with Swin and on PubFig83. Relative to the Low-Shapley controls, the High-Shapley models exhibit more extensive similarity across later layers, indicating that successive layers perform less differentiated transformations. This pattern complements the ResNet-18 and TinyImageNet results in Figure 5 and supports our interpretation that High-Shapley selection yields a less differentiated feature hierarchy.

Figure 12 extends the language-model representation analysis from GSM8K in Figure 7 to 500 examples sampled from the full MATH test set. The SimpleRL and PRIME checkpoints again show stronger similarity between neighboring layers earlier in the network than the Instruct baseline. Observing the same qualitative structure on a second mathematical-reasoning benchmark



(a) High (confidence) (b) Low (confidence) (c) Pre-train (confidence) (d) Scratch (confidence)

Figure 10: Member and non-member confidence distributions on a logarithmic scale for PubFig83 under data selection and pre-training.



(a) High (C-10 (swin)) (b) Low (C-10 (swin)) (c) High (PubFig83) (d) Low (PubFig83)

Figure 11: Intra-CKA under Shapley-based data selection on CIFAR-10 (C-10) with Swin and PubFig83.

provides additional evidence that the simplified-RL checkpoints develop more homogeneous internal representations rather than the effect being specific to GSM8K.

Activation Sparsity under Data Selection. Table 20 reports the fraction of zero ReLU activations across five network stages for the utility-matched High- and Low-Shapley models. The High-Shapley models generally exhibit lower activation sparsity, particularly in the later stages, indicating that a larger proportion of their ReLU units remain active. This result shows that data selection changes the density of internal activations. Because activation density alone does not determine representational complexity or model-stealing difficulty, we treat it as a complementary descriptive observation rather than independent evidence of reduced functional redundancy.

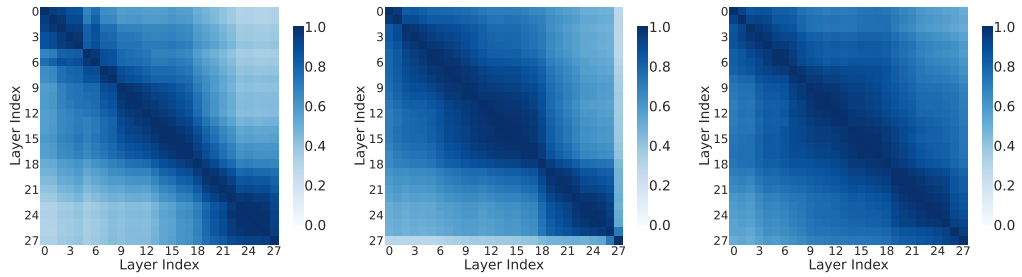
Loss Landscapes. Figures 13 and 14 visualize the local loss geometry for TinyImageNet and PubFig83. For both data-selection comparisons, the High-Shapley models lie in sharper basins than their utility-matched Low-Shapley controls. Likewise, the pre-trained fine-tuned models exhibit sharper local geometry than the corresponding scratch-trained models. These visualizations agree with the quantitative sharpness measurements in Table 11 and extend the CIFAR-10 landscapes in Figure 3. The consistent direction across datasets and training paradigms supports the conclusion that the evaluated efficiency mechanisms are associated with geometrically less stable solutions.

C Essential Experimental Settings

C.1 Vision Models

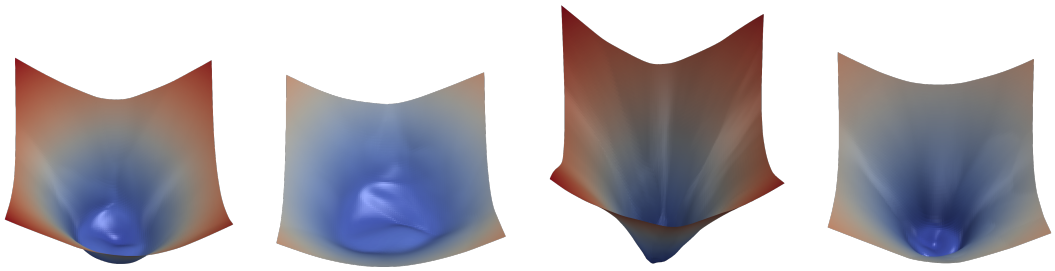
Models and data. ResNet-18 is the default architecture for CIFAR-10, TinyImageNet, and PubFig83; we additionally evaluate Swin-Tiny on CIFAR-10. KNN-Shapley uses $K = 6$. High- and Low-Shapley subsets contain the highest- and lowest-ranked examples, respectively, are disjoint, and have intentionally different sizes to match clean utility. In the pre-training comparison, the fine-tuned model uses public ImageNet-pretrained weights and a smaller task-specific subset, whereas the baseline is initialized randomly and trained from scratch on a larger subset. The backbone and classifier are optimized jointly. Table 21 reports the training-set sizes after excluding reserved model-stealing queries. The PubFig83 test set contains 830 images (10 per class).

Optimization and repeated runs. Vision models use cross-entropy and SGD with momentum 0.9, weight decay 10^{-4}, and batch size 128. Data-selection runs use initial learning rate 0.1 for ResNet-18 and 10^{-3}



(a) Instruct (b) SimpleRL (c) Prime

Figure 12: Intra-CKA on 500 examples sampled from the MATH test set.



(a) High (b) Low (c) Pre-train (d) Scratch

Figure 13: Loss landscape for TinyImageNet.

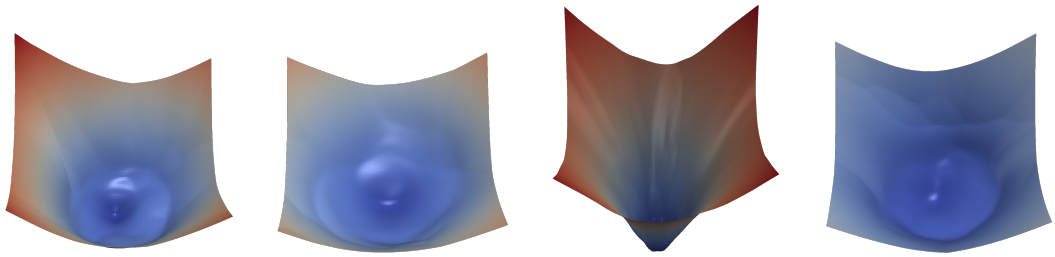
for Swin-Tiny. They run for 300 epochs on CIFAR-10, 100 on TinyImageNet, and 200 on PubFig83, with tenfold learning-rate reductions at 75% and 90% of training. Scratch training uses the same architecture-specific initial learning rates. Fine-tuning/scratch comparisons run for 300 epochs on CIFAR-10 and 200 epochs on TinyImageNet and PubFig83 with the same relative milestones; fine-tuning uses initial learning rate 0.01 for ResNet-18 and 10^{-4} for Swin-Tiny. Full-data controls use the same architecture, optimizer, learning-rate schedule, augmentation, epochs, and seeds as the corresponding High- and Low-Shapley experiments; only the training-set size changes. All reported three-seed vision results use seeds $\{0, 1, 2\}$ with fixed hyperparameters and report the sample standard deviation. Models within each comparison use corresponding seeds.

Membership inference and adversarial robustness. Vision MIA balances members from the model's actual training subset against non-members from its evaluation set and reports TPR at 0.1% FPR. The three scalar scores are ground-truth confidence, prediction entropy, and modified prediction entropy. We evaluate untargeted ℓ_∞ PGD with cross-entropy, $\epsilon = 8/255$, step size $\epsilon/20$, uniform random initialization, and 40 updates. For N evaluated examples, the reported robustness score is

$$R_{\mathrm{PGD}} = \frac{1}{N} \sum_{i=1}^{N} \sum_{t=0}^{39} \mathbf{1}\{\widehat{y}_\theta(x_i^{(t)}) = y_i\},$$

where $\widehat{y}_\theta(x)$ is the model's predicted class. This is the average number of correctly classified iterates across the 40 pre-update checks; higher values indicate greater robustness. The procedure does not stop after the first misclassified iterate.

Model stealing. Attack images are drawn from the target distribution and excluded from target training. The data-selection experiments use 1,000 distinct images for CIFAR-10 and TinyImageNet and 500 for PubFig83; the fine-tuning/scratch experiments use 1,000. The surrogate matches the target architecture and minimizes mean L_1 distance between raw target and surrogate logits. We consider both random initialization and, for informed attacks, the same public pre-trained initialization used by the target. Surrogates train for 100 epochs with batch size 128. ResNet-18 uses SGD (learning rate 0.1, momentum 0.9, weight decay 5×10^{-4}, tenfold decay at epochs 75 and 90); Swin-Tiny uses AdamW (learning rate 5×10^{-5}, weight decay 0.05, cosine decay). Reported query counts refer to distinct underlying images. Because augmented images are queried again during surrogate training rather than caching target logits, Q images produce approximately $100Q$ image-level target evaluations.



(a) High (b) Low (c) Pre-train (d) Scratch

Figure 14: Loss landscape for PubFig83.

C.2 Structural Measurements

Sharpness and calibration. We compute sharpness using Equation 2 over the complete test set. Perturbations start at zero and receive five normalized projected-gradient-ascent updates of size 2×10^{-4}. Each trainable parameter tensor is separately projected onto an L_2 ball of radius 10^{-3}; original parameters are restored after each batch. Vision uses cross-entropy, while LLMs use shifted next-token cross-entropy. Vision ECE uses 15 equal-width confidence bins and maximum-class probability; LLM ECE uses ten bins and the mean probability of greedily generated tokens. LLM calibration is evaluated on GSM8K and the MATH 500 benchmark.

PCA, CKA, and pruning. PCA participation ratio is computed from centered penultimate-layer features over the complete vision evaluation set. Linear CKA compares the initial convolution and eight residual blocks of ResNet-18; the language-model analysis uses the final input-token representation from every Transformer layer. LLM CKA uses 500 test prompts per benchmark, sampled with seed 0 and shared across models; the MATH analysis draws these prompts from the full MATH test set. The maximum input length is 512. For pruning, we independently remove the smallest-magnitude 50% of weights in every convolutional and linear layer, including classifiers, patch embeddings, attention projections, and MLPs. Biases, normalization and positional parameters are excluded, and accuracy is evaluated immediately without recovery fine-tuning.

C.3 Language Models

Checkpoints and membership inference. We evaluate `Qwen/Qwen2.5-Math-7B-Instruct`, `hkust-nlp/Qwen-2.5-7B-SimpleRL-Zoo`, and `PRIME-RL/Eurus-2-7B-PRIME`, together with the corresponding 32B Instruct and SimpleRL checkpoints. MIA candidates are drawn from difficulty levels 3–5 of MATH: the candidate-member source combines the training split and the first 4,500 test entries, while the remaining 500 test entries provide candidate non-members. Availability after filtering yields 339 examples per class. These source-defined groups provide the labels used throughout the LLM MIA evaluation. Answer likelihood is the mean conditional log probability of the reference solution; self-confidence is the mean conditional log probability of a greedily generated continuation; Min-k% with $k = 15$ averages the lowest-probability 15% of retained reference-answer tokens. A balanced one-dimensional logistic regressor orients each score, and AUC and interpolated TPR are computed on this set. Reference sequences are truncated to 1,536 tokens; self-confidence uses prompts of at most 1,024 tokens and at most 512 generated tokens.

Robustness and jailbreak evaluation. Reasoning robustness is measured on RobustMath. Each question is placed in the user turn of a Qwen chat prompt with the system instruction to reason step by step and place the final answer in `\boxed{}`. We greedily generate one response per question (temperature 0, top-$p = 1$, seed 0) with at most 1,024 new tokens. Accuracy compares the extracted final answer with the reference using normalized numerical or symbolic equivalence; empty answers and grading timeouts are incorrect. AutoDAN-HGA is evaluated on 520 AdvBench prompts for at most 100 search iterations, using population size 16, elite fraction 0.05, crossover probability 0.5, five crossover points, mutation probability 0.01, and seed 20. Responses use temperature 0.7, top-$p = 0.9$, and at most 64 new tokens. ASR uses all prompts, while average steps uses successful attacks and counts a success in the initial population as step zero.

IMDB adaptation and quantization. For the fine-tuning-stability evaluation, models undergo two epochs of causal language-model adaptation on the shuffled IMDB training split. We use 4-bit NF4 loading with bfloat16 computation and LoRA on the `q_proj`, `k_proj`, `v_proj`, and `o_proj` modules ($r = 8$, $\alpha = 16$, dropout 0.05). The learning rate is 2×10^{-4}, per-device batch size is one, gradient accumulation is four, warmup is ten steps, and seed is 42. The adapter is merged into the bfloat16 base model before evaluation. The separate quantization sensitivity experiment loads the original checkpoints in 4-bit NF4 with float16 computation and uses matched decoding settings for the quantized and unquantized evaluations.

Table 17: Mathematical-reasoning performance before and after 4-bit NF4 quantization. Higher values are better.

Model	State	GSM8K	MAWPS	Minerva	MATH 500	Olympiad	AIME24	AMC23	Avg.
Instruct	Before	95.38	98.60	31.62	77.20	33.93	3.33	52.50	56.08
	After	95.00	98.60	31.25	79.20	33.48	6.67	47.50	55.96
SimpleRL	Before	93.40	97.09	25.74	75.40	30.37	10.00	45.00	53.86
	After	90.30	97.29	26.10	71.20	30.96	6.67	40.00	51.78
PRIME	Before	89.69	96.22	25.37	70.60	32.00	6.67	55.00	53.65
	After	80.44	77.87	24.26	66.00	31.56	6.67	47.50	47.76

Table 18: Membership inference, adversarial robustness, and AutoDAN jailbreak susceptibility for 32B models. Confidence-based MIA reports TPR at 1.0% FPR; Min-k% MIA ($k = 15$) reports TPR at 0.1% FPR and AUC. Higher robustness accuracy is better; higher MIA TPR, AUC, and ASR and fewer AutoDAN steps indicate greater vulnerability.

Model	Conf. MIA	Min-k% MIA		Robustness		AutoDAN	
	TPR	TPR	AUC	Base	Math	Steps	ASR
Instruct	0.59%	0.59%	0.802	99.50	97.00	1.32	98.08%
SimpleRL	0.88%	1.77%	0.833	99.50	93.70	0.23	100.00%

Table 19: Fine-tuning stability of Qwen2.5 32B models under IMDB adaptation. "Before" and "After" report mathematical-reasoning benchmark accuracy before and after fine-tuning.

Model	State	GSM8K	MAWPS	Minerva	MATH 500	Olympiad	AIME24	AMC23	Avg.
Instruct	Before	95.91	98.31	42.28	80.40	37.78	13.33	62.50	61.49
	After	95.07	98.31	37.87	77.60	36.30	10.00	52.50	58.24
SimpleRL	Before	95.91	98.31	31.99	77.60	39.70	10.00	52.50	58.00
	After	85.37	96.66	21.69	50.20	17.93	3.33	30.00	43.60

Table 20: Fraction of zero ReLU activations under data selection. Lower values indicate that a larger proportion of units are active.

Dataset	Value	0	1	2	3	4
CIFAR-10	Low	0.512	0.509	0.561	0.677	0.781
	High	0.489	0.421	0.498	0.526	0.587
TinyImageNet	Low	0.519	0.505	0.465	0.624	0.806
	High	0.531	0.440	0.467	0.506	0.684

Table 21: Training-set sizes in the vision experiments.

Dataset / architecture	Full	High	Low	Fine-tuning	Scratch
CIFAR-10 / ResNet-18	49,000	1,180	5,000	1,850	4,000
CIFAR-10 / Swin-Tiny	49,000	700	5,000	180	28,000
TinyImageNet / ResNet-18	99,000	12,000	50,000	4,700	20,000
PubFig83 / ResNet-18	6,970	1,950	4,000	2,800	3,680