Security and Privacy of Machine Learning, 2025 Critique: Adversarial Robustness Limits via Scaling-Law and Human-Alignment Studies

Shih-Yu Lai
National Taiwan University
Taipei, Taiwan
akinesia112@gmail.com

I. Summary of the Paper

The paper investigates why adversarial robustness on CIFAR-10 remains stubbornly below human level despite years of progress. The authors (1) derive scaling laws for adversarial training that explicitly include **data quality** of synthetic training sets (measured by FID), yielding compute-optimal allocations of model size N and data size D under an empirically approximated compute scaling FLOPs $\approx 7822\,ND$ for TRADES+PGD training, and numerically derived optima N^*, D^* as power laws of the compute budget. They instantiate two parametric forms: Approach 2 ($L = \frac{A}{N^\alpha} + \frac{B'}{D^\beta} + E'$, modeling Quality as $1/\text{FID}$, noting that this proxy measures image realism rather than strict class-faithfulness) and Approach 3 (a Kaplan-style bottleneck model in which data **quality**—not merely size—controls overfitting).

Guided by these laws, they retrain WideResNets on mixtures of real and high-quality synthetic CIFAR-10, matching prior SOTA with 20% fewer training FLOPs and then surpassing it by +3% AutoAttack accuracy; their best model reports **73.71%** AutoAttack and **79.49%** when invalid adversarial images are excluded. They also quantify adversarial-training cost (27× a single forward pass per iteration due to 10-step PGD + TRADES) and show that simply scaling compute would require on the order of 10^{30} FLOPs (3,000 years on 25,000 MI300/H100 GPUs) to reach human-level robustness.

Finally, a small human/GPT-4 study on 2,629 adversarial images misclassified by their SOTA model finds roughly **28%** of adversarial samples invalid (either deceptive or ambiguous), which contributes to an **10%** irreducible error ceiling in human performance. Average human accuracy over all adversarial images is 90%, which reflects an irreducible error ceiling partly driven by the ˜28% of invalid adversarial samples; this aligns with the scaling-law asymptote at FID $\to 0$ and is interpreted as a fundamental limit of current ℓ_∞ benchmarks and motivation to redesign attacks/evaluations that enforce validity.

II. Strengths

- **Actionable scaling perspective.** The compute-optimal N^*, D^* curves (and the 7822 ND FLOP law) provide concrete guidance for training-budget allocation under adversarial training, a contribution with immediate practical value.
- **Data-quality integration.** Modeling Quality$= 1/\text{FID}$ in the loss highlights that *better synthetic data raises the asymptote* and changes the compute frontier—moving beyond size-only laws in the robustness literature.
- **Compute accounting.** The derivation that TRADES+10-step PGD is 27× a forward pass demystifies why robust training is expensive and where efficiencies might lie.
- **Human-alignment lens.** The identification and taxonomy of “deceptive” vs “ambiguous” invalid adversarials sharpens the community’s notion of what constitutes a *valid* attack/example, and the human 90% ceiling triangulates the scaling-law predictions.
- **SOTA under a tighter budget.** Matching and then exceeding prior robustness with fewer FLOPs demonstrates the utility of the scaling prescriptions (not just descriptive fits).

III. Weaknesses / Concerns

- **Narrow threat model & dataset.** All primary conclusions rest on CIFAR-10 at 32×32 and ℓ_∞ AutoAttack/PGD. It remains unclear whether the 90% ceiling and 10^{30} FLOP extrapolation hold for ℓ_2, elastic/patch/semantic attacks, or higher-resolution datasets (CIFAR-100, ImageNet). The paper gestures at generality, but the empirical base is limited.
- **Validity judged post-hoc by small, non-blinded raters.** Only three human raters (all co-authors) plus GPT-4 were used, raising concerns about sample size, potential bias, and instructions anchoring. The validity rule (2/4 correct or 1 high-confidence human) is reasonable but ad-hoc; no inter-rater reliability or psychophysics (time-limited vs. untimed) is reported.
- **FID as “quality” proxy.** FID correlates with image realism but is not label-preservation–aware. For instance, an adversarially perturbed “airplane” image might still achieve a low FID score while visually resembling a “bird.” If the authors could incorporate additional quality metrics—such as CLIP-based class similarity scores or

direct human verification of class presence—the evaluation of robustness scaling would be more convincing. See also works proposing alternative quality or semantic metrics beyond FID, such as LPIPS [1], improved precision/recall metrics [2], and CLIP-based class consistency [3], which could strengthen the evaluation framework.

- **Convert-loss-to-accuracy details matter.** As Approach 2 converts predicted losses to accuracies via an appendix mapping, the reported asymptotes and compute thresholds depend on that calibration; more visibility into sensitivity/uncertainty would help.
- **Compute claims depend on fixed algorithms.** The 10^{30} FLOP figure presumes today's TRADES/PGD-style pipelines and WideResNet-like backbones; more efficient verifiably robust training or architectures (e.g., certifiably robust networks, diffusion-regularized training) could shift the scaling frontier substantially.

IV. Potential improvements or extensions

1) **Validity-aware benchmarks.** Formalize an *attack-side* constraint that enforces perceptual label preservation—e.g., (i) object-mask consistency (segment the labeled object in the clean image and enforce preservation); (ii) perceptual metrics tuned for class identity (LPIPS-ID, CLIP-based class logits consistency); (iii) counterfactual generation with a class-conditional generative prior that disallows class swaps—to replace the current ℓ_∞ proxy. This would align evaluation with the paper's human-alignment motivation.
2) **Larger, blinded human studies.** Use preregistered protocols with 50 AMT raters, inter-rater agreement (Fleiss'), time-budgeted trials, and randomized clean/adversarial presentation to (a) validate the 10% invalid rate and (b) stratify by "deceptive vs. ambiguous" failure modes at scale.
3) **Beyond FID for "quality."** Replace or augment FID with class-faithfulness metrics: clean-label consistency under small transformations, classifier-two-sample tests conditioned on label, or human-verified class presence; re-fit Approaches 2–3 with these measures to see whether asymptotes move.
4) **Threat-model diversity.** Replicate scaling-law fits under ℓ_2, patch, spatial, and semantic attacks; report whether the compute-optimal N^*, D^* exponents and asymptotes change. If only ℓ_∞ exhibits the 90% ceiling, the conclusion should be scoped accordingly.
5) **Architectural & algorithmic knobs.** Stress-test the laws across different backbones (ConvNeXt, ViT) and training losses (e.g., MART, GAIRAT) and report whether the 7822 ND constant or exponents persist. A negative result would be as informative as a positive one for theory building.
6) **Uncertainty bands.** Provide confidence intervals for asymptotic accuracy and compute projections via bootstrapped fits; report how calibration (loss→accuracy map) and outlier runs influence the 10^{30} claim.
7) **Data-mix ablations.** Systematically vary real:synthetic ratios and generator types at fixed FLOPs to disentangle benefits from sheer data volume vs. generative diversity (Table 1 shows a wide quality range ripe for such analysis).

V. Questions for the authors

- How sensitive are your fitted exponents and asymptotes to the loss→accuracy conversion and to excluding low-FID datasets (e.g., EDM-5)? Could you add uncertainty bands on Figure 1?
- If we substitute FID with a class-preserving metric (e.g., CLIP-ID consistency), do Approaches 2–3 still predict a 90% ceiling?
- What changes in the scaling picture under ℓ_2 or spatial attacks, or on higher-res datasets (CIFAR-100, ImageNet-100)?
- Your FLOP accounting assumes 10-step PGD TRADES; how do results change with stronger/cheaper inner maximizers (e.g., FAB, Square-Attack) or single-step adversarial training with random starts?
- In the human study, why fix the validity rule at "2/4 or 1 high-confidence"? Did you test alternatives (majority vote, confidence-weighted voting), and how robust is the 10% invalidity rate to such choices?

Bottom line. This paper productively reframes progress in adversarial robustness as a *compute–data–quality* problem and spotlights a human-alignment failure in current ℓ_∞ evaluations. Its pragmatic scaling guidance and SOTA results are strong, but broader threat models, larger human studies, and label-preserving validity constraints are needed before translating the reported 90% ceiling into a general "limit."

References

[1] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 586–595.

[2] T. Kynkäänniemi, T. Karras, S. Laine, J. Lehtinen, and T. Aila, "Improved precision and recall metric for assessing generative models," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019, pp. 3927–3936.

[3] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.