Security and Privacy of Machine Learning, 2025 Critique: Model & Data Privacy – (1) Stealing Part of a Production Language Model; (2) Trap-MID: Trapdoor-based Defense against Model Inversion Attacks; (3) Generative Model Inversion Through the Lens of the Manifold Hypothesis

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

CRITIQUE:(1) STEALING PART OF A PRODUCTION LANGUAGE MODEL

SUMMARY

The paper demonstrates that a black-box adversary can extract *real parameters*—not just functional behavior—from production LLMs by exploiting common API features. The core idea is to operate *top-down*: recover the full output projection matrix W (and thus the hidden width h) of transformer LMs by reconstructing full logit vectors from restricted interfaces (e.g., Top-K logprobs with *logit-bias*), then using linear-algebraic factorization to separate W from the hidden states. The authors formalize threat models for three API regimes (all logits; top-K logprobs + logit-bias; logprob-free with logit-bias), design query- and token-efficient extraction procedures, and validate them on both open-source models and multiple production models (with provider permission). They report high-fidelity reconstructions (RMS error $< 10^{-3}$ up to symmetry), recover hidden sizes, and propose mitigations (e.g., disallowing simultaneous logprobs and logit-bias, rate-limiting, or architectural changes). The work reframes model stealing from "functional imitation" [1], [2] to *parameter recovery* under realistic LLM APIs([1]–[5]).

STRENGTHS

- **Conceptual novelty.** Flips the usual bottom-up extraction (input$\rightarrow$embedding$\rightarrow \cdots$) by directly targeting the last layer via its low-rank structure ($h \ll \ell$), yielding a surprisingly effective path to *parameter* theft rather than only fidelity theft.
- **Practicality and careful engineering.** The attacks are derived for exactly the interfaces that popular providers exposed (top-K logprobs, bounded logit-bias), with clear *token* vs. *query* cost accounting that mirrors how APIs are charged/rate-limited in practice.
- **Responsible evaluation with production systems.** Coordinated disclosure and provider-confirmed results elevate the impact beyond lab settings; the paper shows tangible changes to API design decisions.
- **Security clarity.** The analysis crisply distinguishes what each API capability leaks, and why combining capabilities (logprobs *and* logit-bias) is multiplicative for the adversary.
- **Broader applicability.** Even partial recovery (hidden width, last layer up to symmetries) materially reduces "black-boxness," enabling downstream analysis and possibly facilitating other attacks.

WEAKNESSES

- **Reliance on specific API affordances.** The main efficiency gains hinge on logit-bias and top-K logprobs; once these are decoupled or restricted, attacks become costlier or numerically fragile. The paper could more deeply quantify robustness under aggressive API hardening.
- **Symmetry ambiguity and practical utility.** Recovery is up to affine/orthogonal transforms. While expected, the paper stops short of showing concrete downstream attacks that *necessitate* aligned W (e.g., targeted jailbreak construction or watermark removal efficacy).
- **Numerical stability.** Some variants require solving ill-conditioned systems; a more thorough conditioning analysis (e.g., sensitivity to softmax temperature, quantization noise, or provider-side stochasticity) would strengthen reliability claims.
- **Limited exploration of modern mitigations.** Beyond API tuning, emerging defenses (e.g., per-user noise shaping, randomized *top-K* selection, DP-inspired output

perturbations) are discussed but not experimentally stress-tested against utility.
- **Generalization to multimodal.** The discussion mentions extensions, but no empirical evidence is given for VLMs or audio-text models whose APIs expose different artifacts.

Potential Improvements / Extensions

- **Hardening sensitivity map.** Provide a systematic ablation sweeping: (i) removing logit-bias, (ii) restricting K, (iii) quantizing/rounding logprobs, (iv) adding small, calibrated logit noise, and (v) randomizing vocabulary subsets per request. Plot utility–leakage frontiers to guide providers.
- **Lower bounds and optimality gaps.** Tighten query/token complexity lower bounds for parameter recovery under various API constraints, to contextualize the $< \times 2$ gaps the paper notes.
- **Symmetry breaking.** Explore language-informed priors (e.g., anisotropy of token embeddings) or cross-task probes to resolve the $h \times h$ ambiguity and turn approximate W into directly actionable structure.
- **Downstream risk demonstrations.** Show how stolen W concretely improves (a) prompt-injection success rates, (b) red-teaming coverage models, (c) output watermark removal, or (d) transfer of fine-tuning heads—turning partial theft into measured harm.
- **Noisy/quantized production settings.** Reproduce attacks with server-side logit quantization, temperature jitter, and caching/retrieval augmentation to measure real-world headwinds.
- **Multimodal API analysis.** Adapt the extraction to per-token/piece probabilities in VLMs, or to audio/text dual-heads; identify which modality couplings leak more structure.

Questions for the Authors

- Can the recovered W be used to *diagnose* or reconstruct pieces of the tokenizer (e.g., identify merge rules or special-token handling) beyond recovering ℓ?
- How does server-side speculative decoding or KV cache reuse alter leakage (e.g., if logits reflect draft-model mixture)?
- Could per-account randomized logit remapping (fixed secret permutation + slight noise) preserve utility yet break cross-query linear structure needed for SVD-based recovery?
- Do retrieval-augmented systems inadvertently amplify leakage (e.g., shifting logits in data-dependent ways that aid rank estimation)?
- Are there principled, utility-preserving constraints (e.g., DP budgets on logprob exposure) that provably raise query complexity for last-layer recovery?

Critique: (2) Trap-MID: Trapdoor-based Defense against Model Inversion Attacks

Summary.

The paper proposes **Trap-MID**, a trapdoor-based defense that aims not to suppress all private information in the model but to *mislead* Model Inversion (MI) attacks toward extracting trapdoor triggers instead of genuine private data. Concretely, the method injects class-wise blended triggers during training and co-optimizes (i) the classifier, (ii) a discriminator to encourage trigger *naturalness*, and (iii) the trigger patterns, so that trigger-injected inputs are confidently mapped to a chosen label while remaining visually indistinguishable from clean data. A simple theoretical bound relates deception success to two properties: trapdoor *effectiveness* (predictive power gap on triggered vs. benign data) and *naturalness* (KL divergence between clean and triggered distributions). Empirically, on CelebA with VGG-16 (plus Face.evoLVe/ResNet-152 in the appendix), Trap-MID reduces SOTA white-box attacks—including PLG-MI's cGAN approach—from near-perfect top-1 accuracy to single digits, and outperforms dependency-regularization and negative label smoothing baselines. The study also explores adaptive attackers and shows resilience under auxiliary-data shift. The approach fits within a broader shift from information-suppression defenses (e.g., MID) toward *misleading* strategies grounded in generative priors [6], [7], label smoothing [8], and trapdoor ideas from adversarial detection [9].

Strengths.

- *Clear insight and framing.* The central insight—guiding MI optimization to a high-confidence, high-plausibility "shortcut" manifold—is crisp and aligns with the optimization geometry of cGAN-based MI [6], [7]. The δ (effectiveness) vs. ε (naturalness) decomposition provides an interpretable "tension metric" for designing triggers rather than treating trapdoors as black magic.
- *Targeted to the actual MI loop.* By optimizing trigger naturalness with a discriminator, the method addresses a core reason prior patchy trapdoors fail against MI (the GAN discriminator rejects unnatural artifacts). This bridges a known gap between adversarial-perturbation settings and MI's realism constraints [6].
- *Broad empirical coverage.* The evaluation spans multiple MI families (GMI/KED-MI/LOMMA/PLG-MI), architectures, label-only settings, and auxiliary data shifts, with consistent and often large margins. The synthetic-distribution analysis is a nice diagnostic that the generator indeed gravitates toward public-like *triggered* regions.
- *Practicality.* No shadow attacks, no external confounder datasets, and minimal data assumptions compared to NetGuard/DCD-style misleading defenses. Training-time overhead is moderate and clearly reported.

Weaknesses.

- *Theory is sufficient-but-not-necessary.* The bound uses a coarse global KL notion of naturalness and an averaged

predictive-power gap; it does not capture per-class heterogeneity or the local geometry of the MI objective (e.g., max-margin latent search in PLG-MI [7]). As a result, it is hard to use the theory *constructively* to pick hyperparameters.
- *Coupling to KD/transfer pipelines.* The defense weakens under KD-based MI (LOMMA), echoing prior backdoor fragility when the teacher never exposes triggered behavior to the student. This suggests brittleness to common deployment practices (distillation, pruning, LoRA, adapters).
- *Stability and variance.* Reported variance across random trigger initializations is nontrivial, and the best settings differ across attacks (e.g., blend ratio, loss weights). The method may require careful, attack-aware tuning to be reliably strong.
- *Modality and semantics.* The work is vision-centric with pixel-space blended triggers. It remains unclear how to design semantically natural triggers in text/graph/tabular, or for face recognition models with modern margin losses and large-scale training.
- *Detection side-effects.* While trapdoor signatures aid adversarial detection (a plus), they also create a detectable fingerprint an adaptive attacker could explicitly avoid or subtract (indeed explored partially). A broader game-theoretic treatment is missing.

POTENTIAL IMPROVEMENTS OR EXTENSIONS.

- *Constructive design from theory.* Replace global KL with *fisher-weighted* or *feature-space* divergences aligned to the attacker's discriminator/encoder; relate $\delta - \varepsilon$ to measurable margins/logit gaps used by PLG-MI [7]. Provide a recipe that maps desired deception strength to (α, β, augmentation strength) with statistical guarantees.
- *KD-robust trapdoors.* Explore *distillation-consistent* trapdoors: e.g., expose a small, privacy-safe subset of triggered behavior to the student; or encode trapdoor features in mid-level invariances that survive KD. Compare to MID/BiDO regularizers [10] in a hybrid objective that preserves deception after compression.
- *Semantic triggers.* Move beyond blended noise toward *concept triggers* (hair tint, accessory, background texture) learned via disentangled or diffusion priors; these could be more natural (smaller effective ε) and more transferable across augmentations and domains.
- *Cross-modality generalization.* Prototype language/graph/tabular Trap-MID: in NLP, inject lexical or syntactic templates tied to labels; in graphs, degree/attribute motifs; in EHR/tabular, missingness or rounding patterns. Study attacker priors in each modality.
- *Evaluation beyond FID/AA.* Incorporate privacy-specific metrics insensitive to OOD failure modes (e.g., nearest-neighbor identity leakage in face embeddings, privacy risk curves as a function of generator capacity) and causal analyses of which sensitive attributes are still leaked after deception.

QUESTIONS FOR THE AUTHORS.

- How does Trap-MID interact with *open-set* recognition or class expansion post-deployment? Do triggers for unseen identities inadvertently form new shortcuts?
- Can you characterize which classes (attributes) are hardest to protect (smallest empirical $\delta - \varepsilon$)? Is there a link to intra-class variability or head/long-tail identity frequency?
- Under multi-model ensembles (common in APIs), does diversity *reduce* or *amplify* deception? Would voting dilute trapdoor effects?
- Could a diffusion-model attacker (score-based) neutralize trapdoors by projecting onto the clean data manifold while optimizing identity loss [6]? Any initial results?
- How brittle is deception to small architectural edits or fine-tuning on new domains (e.g., ArcFace-style training on in-the-wild faces)?

CRITIQUE: (3) GENERATIVE MODEL INVERSION THROUGH THE LENS OF THE MANIFOLD HYPOTHESIS

SUMMARY

This paper offers a geometric account of why *generative* model inversion attacks (MIAs) are effective. Building on the manifold hypothesis, the authors show that backpropagating inversion-time classification loss through a generator implicitly projects noisy input-space gradients onto the generator manifold's tangent space, thereby denoising and retaining semantically aligned directions. They quantify *gradient–manifold alignment* via the cosine between the loss gradient and its tangent-space projection, and empirically observe that standard models exhibit low alignment (slightly above the $\sqrt{k/d}$ random baseline [11]).

From this viewpoint, they posit a central hypothesis: models whose loss gradients align more strongly with the generator manifold are more vulnerable to MIAs. To test this, they introduce an *alignment-aware* training objective that encourages input-gradient alignment with an estimated natural-image manifold derived from a pre-trained Stable Diffusion VAE decoder [12]. They also propose *AlignMI*, a training-free family of methods that improve alignment at inversion time by averaging loss gradients over local perturbations (PAA) or semantic-preserving transformations (TAA). Across face recognition benchmarks and state-of-the-art generative MIAs (e.g., Plug & Play Attacks [13]), they report consistent gains, and present evidence supporting the alignment–vulnerability link, complementing prior MIA literature from direct input-space optimization [14] to GAN-based generative inversion [6].

STRENGTHS

- **Conceptual clarity via geometry.** Reframing generative MIAs as *implicit gradient projection* onto a generator manifold is an elegant, explanatory contribution. It connects an empirical recipe (optimize in latent space) to a clean geometric mechanism (tangent-space projection), helping unify disparate generative MIA techniques under one lens.

- **Actionable metric.** The alignment score is simple to compute (given a local tangent basis) and aligns with intuition: larger on-manifold components should yield more semantically faithful reconstructions. Referencing the $\sqrt{k/d}$ random-vector baseline [11] is a nice calibration.
- **Two complementary validations.** (i) A training-time objective that increases alignment and (ii) a training-free inversion-time procedure (PAA/TAA) that amplifies on-manifold components provide converging evidence for the core hypothesis. The latter is particularly practical: no model changes required.
- **Grounded system design.** Leveraging the Stable Diffusion VAE [12] for manifold tangent estimation is a pragmatic choice that scales to natural images and dovetails with the community's tooling.
- **Broad relevance.** The work connects older MIA formulations [14] and modern generative inversion [6], [13], potentially informing both stronger attacks and geometry-aware defenses.

Weaknesses

- **Manifold estimator dependence.** Alignment is measured w.r.t. (i) the generator manifold during inversion and (ii) an external VAE-manifold during training. This assumes these manifolds are good surrogates for the *private-data* manifold. Distribution shift between private data and FFHQ/LAION priors (or the VAE's latent geometry) may bias alignment estimates, particularly outside faces.
- **Computational overhead and scalability.** Constructing tangent spaces via SVD of decoder Jacobians and averaging over K transformations introduces non-trivial cost. The paper acknowledges memory/runtime constraints at high resolution, but the practical frontier (e.g., ImageNet-1k, medical images) remains unclear.
- **Causality vs. correlation.** While the study shows a relationship between alignment and MIA success (with an inverted V-shape at extremes), a principled analysis explaining *why* excess alignment can reduce attack success (beyond reduced generalization) is missing.
- **Threat model coverage.** Results primarily target white-box generative MIAs in face recognition. Black-box/label-only regimes, non-face domains (e.g., medical, OCR, satellite), and non-image modalities (audio/text) are not assessed, limiting external validity beyond the most studied setting.
- **Choice of transforms (TAA).** The semantic-preserving transforms are handpicked; their interaction with the target model's invariances and with the evaluator can subtly inflate gains (e.g., when the evaluator shares similar augmentations), risking evaluator overfitting.

Potential Improvements or Extensions

- **Manifold-agnostic alignment proxies.** Explore score-based or diffusion-model Jacobians (via denoisers) as lighter-weight tangent approximations, or Fisher–Rao / NTK local subspaces as surrogates for perceptual manifolds; compare against the VAE decoder.
- **Curvature-aware sampling.** Replace isotropic PAA/TAA with curvature-adaptive neighborhoods (e.g., along principal geodesic directions estimated from J_G or via retractions), which could reduce the number of samples K while improving SNR.
- **Generalization to harder threat models.** Evaluate alignment–vulnerability links under label-only MIAs and query-limited black-box settings, where gradient signals are estimated by priors or finite differences; test whether alignment still predicts success.
- **Defense design from misalignment.** The geometry suggests defenses that *de-align* input gradients from plausible data manifolds (e.g., training-time penalties that rotate gradients off manifold, or inference-time randomization that injects off-manifold components), measured against accuracy/utility.
- **Beyond faces and images.** Validate on non-face image tasks (fine-grained species, medical), and preliminary studies on audio/text, where manifolds and priors differ markedly; assess sensitivity to prior mismatch.
- **Theory of the trade-off.** Develop a stylized model linking alignment, margin, curvature, and generalization to explain the observed inverted V-shaped vulnerability curve; relate to bias–variance and double-descent phenomena.

Questions for Authors

- How sensitive are alignment measurements to the *choice* of manifold estimator (StyleGAN vs. diffusion vs. VAE) and to prior–private distribution gaps?
- Can we estimate alignment *without* full Jacobian SVD (e.g., randomized sketching or Hutchinson-type probes) while preserving ordering across models?
- In black-box or label-only MIAs, does a proxy for alignment (computed w.r.t. a public prior) still predict attack success, or does estimator mismatch dominate?
- Could a defender exploit *targeted misalignment* (e.g., adversarially tilting gradients off-manifold) with minimal accuracy loss, and how would AlignMI adapt?
- What governs the inverted V-shape: manifold curvature, class entanglement, or evaluator bias? Can we predict the peak alignment that maximizes leakage?

References

[1] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, "Stealing machine learning models via prediction apis," in *USENIX Security Symposium*, 2016.

[2] M. Jagielski, N. Carlini, D. Berthelot, A. Kurakin, and N. Papernot, "High-fidelity model extraction attacks against machine learning models," in *USENIX Security Symposium*, 2020.

[3] OpenAI, J. Achiam *et al.*, "Gpt-4 technical report," 2023.

[4] R. Anil *et al.*, “Palm 2 technical report,” 2023.
[5] S. Zanella-Beguelin *et al.*, “Leakage of dataset properties in multi-party machine learning,” in *ACM CCS*, 2021.
[6] Y. Zhang, R. Jia, H. Pei, W. Wang, B. Li, and D. Song, “The secret revealer: Generative model-inversion attacks against deep neural networks,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
[7] X. Yuan, K. Chen, J. Zhang, W. Zhang, N. Yu, and Y. Zhang, “Pseudo label-guided model inversion attack via conditional generative adversarial network,” in *AAAI Conference on Artificial Intelligence (AAAI)*, 2023.
[8] L. Struppek, D. Hintersdorf, and K. Kersting, “Be careful what you smooth for: Label smoothing can be a privacy shield but also a catalyst for model inversion attacks,” in *International Conference on Learning Representations (ICLR)*, 2024.
[9] S. Shan, E. Wenger, B. Wang, B. Li, H. Zheng, and B. Y. Zhao, “Gotta catch ’em all: Using honeypots to catch adversarial attacks on neural networks,” in *ACM Conference on Computer and Communications Security (CCS)*, 2020.
[10] T. Wang, Y. Zhang, and R. Jia, “Improving robustness to model inversion attacks via mutual information regularization,” in *AAAI Conference on Artificial Intelligence (AAAI)*, 2021.
[11] R. Vershynin, *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.
[12] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
[13] L. Struppek, D. Hintersdorf, A. de Almeida Correia, A. Adler, and K. Kersting, “Plug & play attacks: Towards robust and flexible model inversion attacks,” in *International Conference on Machine Learning (ICML)*, 2022.
[14] M. Fredrikson, S. Jha, and T. Ristenpart, “Model inversion attacks that exploit confidence information and basic countermeasures,” in *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2015.