Thesis Direction Analysis

KKT-Based Privacy Attacks on LoRA Adapters

Competitive Landscape, Theoretical Lineage, and Proposed Contributions

February 2026

Contents

1 The Gap in One Sentence

Core Claim

Nobody has applied the KKT/implicit bias reconstruction framework to LoRA adapters. This enables a unique threat model (weight-only, passive, undetectable) that no existing attack addresses, and yields a testable phase transition prediction.

2 Theoretical Lineage: From Implicit Bias to LoRA

Your thesis sits at the end of a clear research arc. Each paper in this chain proves one piece; your thesis supplies the missing final link.

2.1 The Chain

1. **Haim, Vardi, Yehudai, Shamir & Irani (NeurIPS 2022)** [1].
   Foundational paper. For homogeneous networks trained with GD on binary cross-entropy, the weights converge to the KKT point of the max-margin problem:

$$\theta^* \propto \sum_{i=1}^{N} \lambda_i \, y_i \, \nabla_\theta \Phi(\theta^*; x_i) \tag{1}$$

   Optimizing (x, λ) to satisfy this condition reconstructs training images from weights.
   Setting: Full model, full training, small MLPs.

2. **Smorodinsky, Vardi & Safran (October 2024)** [2].
   First **provable** guarantees for implicit-bias privacy attacks. For 2-layer ReLU networks, they prove $\geq 25\%$ of training data is recoverable in 1D, and near-perfect membership inference in high dimensions. This puts the KKT attack on rigorous theoretical ground.
   Open questions stated: Deeper networks, broader settings, defenses.

3. **Oz, Yehudai, Vardi, Antebi, Irani & Haim (July 2024)** [3].
   Extends reconstruction to **real-world transfer learning**: frozen backbone (ViT / DINO / CLIP) + trained MLP head. Two-stage pipeline: (i) reconstruct embeddings via KKT on the MLP head, (ii) invert embeddings back to images.
   Key limitation: Only handles a **full MLP head** trained from scratch. Does **not** address LoRA fine-tuning of the backbone.

4. **Gronich & Vardi (February 2026)** [4].
   Shows Adam and Muon converge to ℓ_∞ KKT points (not ℓ_2 like SGD). Their conclusion explicitly poses:

   > *"Are training-data reconstruction attacks based on satisfaction of KKT conditions... feasible for Adam and Muon?"*

   Since most LoRA training uses Adam, this is a **directly stated open question** that your thesis can address.

5. **ImpMIA — Golbari, Wasserman, Vardi & Irani (October 2025)** [5].
   From **your lab**. Membership inference via KKT λ optimization on full ResNet-18 weights. First to show KKT works on non-homogeneous architectures at scale (25K training samples).
   Key limitation: Full model weights, not LoRA adapters.

2.2 Where Your Thesis Fits

Paper	Model	What's Attacked	Task
Haim et al. 2022	Small MLP	Full weights	Reconstruction
Smorodinsky et al. 2024	2-layer ReLU	Full weights	Provable recon.
Oz et al. 2024	Frozen ViT + MLP	MLP head only	Reconstruction
ImpMIA 2025	ResNet-18	Full weights	Membership
Gronich & Vardi 2026	Theory (Adam)	Full weights	Open question
Your Thesis	**Frozen ViT + LoRA**	**LoRA adapter**	**Recon. + Memb.**

The Missing Link

Oz et al. showed KKT reconstruction works on *frozen backbone + trained MLP head*. You replace "trained MLP head" with "LoRA adapter" — the setting people actually use and share publicly. This is the natural next step.

3 Competitive Landscape (Non-KKT Papers)

Beyond the KKT lineage, several papers attack LoRA privacy from other angles. None occupies your cell.

Paper	Access	Method	Goal	Domain	KKT?
PEFTLeak (CVPR'25) [6]	Training-time (FL)	Grad. inversion + poisoning	Recon.	Vision	No
LoRA-Leak (2024) [7]	Inference (queries)	Output-based MIA	Memb.	Text	No
Hu et al. (2026) [8]	Theoretical	DP analysis	Bounds	Both	No
R2F (NeurIPS'25) [9]	Proxy data	Gradient decoder	Unlearn	Text	No
DP-LoRA (ICCV'25) [10]	—	DP mechanism	Defense	Vision	No
CMU Report (2024) [11]	Inference (prompt)	Generation probe	Memor.	Text	No
Yours	**Weight-only**	**KKT optim.**	**Recon.+M**	**Vision**	**Yes**

Key differentiators.

- **PEFTLeak** requires training-time gradient interception + active poisoning. You need only the published adapter file.
- **LoRA-Leak / Hu et al.** require inference queries. You never touch the model.

- **R2F** trains a decoder on proxy data to recover gradients. You skip the decoder entirely and work directly on the final weights.
- **CMU Report** found "rank has no clear effect on memorization" — but used weak probing. If your phase diagram shows a sharp transition, you directly contradict this with a stronger method.

4 Proposed Contributions

4.1 Contribution 1: LoRA KKT Framework and Weight-Only Attack

For a LoRA-adapted model with frozen base θ_0 and adapter $\Delta W = BA$, the KKT stationarity gives:

$$\boxed{BA = \sum_{i=1}^{N} \lambda_i \, y_i \, \nabla_\theta \Phi(\theta_0 + BA;\ x_i) \ - \ \theta_0} \tag{2}$$

Reconstruct by optimizing:

$$\hat{x}, \hat{\lambda} = \underset{x \in [-1,1]^d,\ \lambda \geq 0}{\arg\min} \left\| BA - \left(\sum_{i=1}^{N} \lambda_i \, y_i \, \nabla_\theta \Phi(\theta_0 + BA;\ x_i) - \theta_0 \right) \right\|^2 \tag{3}$$

Threat model. The attacker downloads a `.safetensors` file. No model access, no queries, no training-time interception. Passive, offline, undetectable, scalable to thousands of public adapters.

Overdetermination insight. The low-rank constraint makes reconstruction *tighter*: the RHS of Eq. (2) must equal a rank-r matrix, eliminating spurious solutions that satisfy KKT but produce full-rank residuals.

4.2 Contribution 2: Phase Transition at r^*

Parameter counting on the KKT system yields:

$$\boxed{r^* \ \approx \ \frac{N \times d_{\text{input}}}{d_{\text{in}} + d_{\text{out}}}} \tag{4}$$

Prediction: Reconstruction succeeds for $r > r^*$ and fails for $r < r^*$. This is directly testable.

Example. $N = 10$ CIFAR-10 images, layer with $d_{\text{in}} = d_{\text{out}} = 768$: $r^* \approx 10 \times 3072/1536 = 20$.

Main result figure. A 2D heatmap with r vs. N, colored by reconstruction quality (SSIM), with the r^* curve overlaid. If theory matches experiment, that is a paper.

Practical guideline. "To safely publish a LoRA adapter trained on N private images, ensure $r < r^*(N)$." No existing paper provides this.

4.3 Contribution 3: Data Provenance Verification

Fix x to a candidate image, optimize only λ:

$$\hat{\lambda} = \arg\min_{\lambda \geq 0} \left\| BA - \left(\sum_{i=1}^{M} \lambda_i \, y_i \, \nabla_\theta \Phi(\theta_0 + BA;\ x_i) - \theta_0 \right) \right\|^2 \tag{5}$$

High $\hat{\lambda}_i \Rightarrow$ membership. This is a **weight-only MIA** — no inference, no queries.

Killer application. A photographer suspects their images trained a Stable Diffusion LoRA. They download the `.safetensors` file and get a membership score. No API calls, no model deployment.

Feasibility. Very high. Optimizing λ alone (with fixed x) is convex-like and far easier than joint (x, λ) reconstruction. ImpMIA (from your lab) provides engineering patterns.

4.4 Contribution 4 (Extension): Adam/Muon KKT Reconstruction

Gronich & Vardi [4] show Adam converges to ℓ_∞ margin KKT points and explicitly ask whether reconstruction is feasible. Since most LoRA training uses Adam:

- Replace the ℓ_2 KKT loss with ℓ_∞ KKT conditions
- Test whether reconstruction quality differs under Adam vs. SGD training
- This directly answers an open question from the Vardi group

This is a stretch goal — not required for the thesis, but would strengthen it significantly.

5 The Strong Framing

Weak (Method-Centric)

"We apply KKT reconstruction to LoRA adapters for the first time."
Reviewer: "So what? Is it better? Why should I care?"

Strong (Insight-Centric)

"LoRA adapters at rank $r > r^*$ are vulnerable to a passive, weight-only attack that requires no model access. We derive r^* analytically and validate it experimentally, providing the first privacy guideline for safe LoRA deployment."

Three claims that stand independently of the method:

1. A new threat model exists (weight-only, passive, undetectable)
2. There is a sharp phase transition at r^*
3. The theory predicts the experiments

KKT is the tool. The **findings** are the contribution.

What KKT uniquely reveals.

- **Which data leaks:** The KKT conditions encode support vectors — the hardest, most atypical (often most private) training examples. Generic data near the centroid has $\lambda_i \approx 0$ and cannot be reconstructed.
- **When data leaks:** The r^* formula gives a precise, falsifiable prediction.
- **The threat is invisible:** No queries, no API calls, no interaction with the model owner.

6 Robustness: Why This Thesis Produces Results Either Way

Outcome	What You Get	Strength
Reconstruction works above r^*	Attack + privacy guideline	Strong
r^* prediction matches experiments	Validated theory + guideline	Very strong
Provenance verification competitive AUC	Practical tool + threat model	Strong
Reconstruction fails even above r^*	Positive privacy result	Moderate

Minimum Viable Thesis

Even if reconstruction fails entirely: the LoRA KKT derivation, the r^* formula, the negative result ("optimization landscape, not information content, is the bottleneck"), and the empirical (r, N) characterization are sufficient for a Master's thesis. Combined with provenance verification, it reaches a workshop paper.

7 Comparison to Original Plan

Original Direction	Problem	Status
Gradient Bridge (R2F decoder $\rightarrow$ inversion)	Noise cascade; PEFTLeak overlap; decoder adds complexity	**Replaced** by direct KKT
LoRA in NTK regime	Requires impractically high r	**Absorbed** into r^* analysis
Generative priors (SDS)	Incremental ("we added a regularizer")	**Dropped**

8 Roadmap to a Top-Venue Paper

A good thesis answers a question. A strong *paper* tells a complete story: phenomenon, theory, attack, defense. This section lays out what it takes to go from "MSc thesis" to "ICML/NeurIPS submission."

8.1 The Five Components

#	Component	What It Delivers	Ef
Theorem (linear case)	Turns the r^* heuristic into a provable result	2–3 weeks	1 Ph tra fig
The "main result" — theory predicts experiments	1–2 weeks	Real-world demo	2 Re fro Hu Fa ad
2–3 weeks			3
4	Defense + tradeoff	SVD truncation closes the attack; utility cost measured	1 5
Provenance tool	Weight-only MIA with AUC evaluation	1 week	

Each component is described below with the standard it must meet.

8.2 Component 1: Prove r^* for the Linear Case

The r^* formula (Eq. 4) is currently a parameter-counting argument. A reviewer at a top venue will call it a heuristic. To elevate it:

Target Theorem

Theorem (informal). Consider a linear model $f(x; W) = Wx$ with LoRA parameterization $W = W_0 + BA$, trained on N samples (x_i, y_i) with binary cross-entropy via gradient descent. If $r \geq r^* = N \cdot d_{\text{input}}/(d_{\text{in}} + d_{\text{out}})$, the KKT system has a unique solution and reconstruction succeeds. If $r < r^*$, the system is underdetermined and reconstruction is information-theoretically impossible.

Why this is tractable. For linear models, the KKT system becomes a system of linear equations. The rank condition for unique solvability reduces to a standard linear algebra argument (rank of the Jacobian matrix). Smorodinsky et al. [2] proved reconstruction for 2-layer ReLU — the linear case is strictly easier.

Why this matters. Even a theorem for the simplest case gives you:

- A **provable lower bound**: "below r^*, *no algorithm* can reconstruct" (information-theoretic, not method-specific)
- **Credibility**: the empirical r^* on nonlinear models is then a "conjecture validated by experiment," which is the standard pattern in deep learning theory
- A direct comparison to Smorodinsky et al. (they prove recovery for full models; you prove recovery for LoRA)

8.3 Component 2: The Phase Transition Figure

This is the centerpiece of the paper — the figure reviewers will remember.

Experiment. Train LoRA adapters at every (r, N) pair:

- $r \in \{2, 4, 8, 16, 32, 64, 128\}, \quad N \in \{5, 10, 25, 50, 100\}$
- Architecture: MLP (where theorem holds), then ViT + LoRA (practical setting)
- Run KKT reconstruction on each; measure SSIM, PSNR, LPIPS

The figure. A 2D heatmap: x-axis = rank r, y-axis = dataset size N, color = reconstruction quality. Overlay the theoretical curve $r^* = N \cdot d/(d_{\text{in}} + d_{\text{out}})$.

What makes it strong.

- If the empirical boundary matches the theoretical curve $\Rightarrow$ validated theory
- Directly contradicts CMU Report [11] ("rank has no effect") with a sharper method
- Produces an actionable guideline: "keep $r < r^*(N)$ to prevent leakage"

8.4 Component 3: Real-World Demonstration

Theory and controlled experiments are necessary but not sufficient. The result that makes the paper *matter*:

The "Holy Shit" Figure

Download 5 public LoRA adapters from HuggingFace or CivitAI (face LoRAs, style LoRAs, character LoRAs). Run your attack. Show reconstructed training images next to the originals (or, if originals are unknown, show recognizable faces/content recovered from the adapter alone).

Why this is essential. A controlled experiment on CIFAR-10 proves a scientific point. Reconstructing real faces from a public adapter proves a **societal point** — one that policymakers, journalists, and practitioners understand immediately.

Fallback. If real adapters are too hard (architecture mismatch, too many training steps), create a realistic simulation: fine-tune ViT-B/16 with LoRA on a private face dataset, publish the adapter, and attack it. This is one step removed from "in the wild" but still compelling.

8.5 Component 4: Defense and Utility Tradeoff

An attack paper without a defense discussion is incomplete. The natural defense:

SVD truncation. Before publishing an adapter BA, compute $\text{SVD}(BA) = U\Sigma V^\top$ and truncate to rank $r' < r^*$. This provably defeats the attack (by your own theorem) while preserving the dominant directions of the adapter.

Experiment.

- Measure task accuracy (on the fine-tuning task) as a function of truncation rank r'
- Measure attack success (SSIM of reconstructed images) as a function of r'
- Plot both on the same axes: the **privacy–utility tradeoff curve**
- Identify the sweet spot where accuracy is preserved but reconstruction fails

Why this closes the loop. The paper now tells a complete story:

1. Here is the attack (KKT reconstruction from LoRA)
2. Here is when it works ($r > r^*$)
3. Here is how to prevent it (truncate to $r < r^*$)
4. Here is what it costs (utility drop of $X\%$)

8.6 Component 5: Provenance Verification

The membership inference variant (Contribution 3, Section 4) serves as:

- **Insurance**: if full reconstruction fails, MIA still works (lower bar)
- **Practical tool**: copyright verification from weights alone
- **Comparison point**: benchmark against LoRA-Leak [7] (output-based) and ImpMIA [5] (full weights) to show the weight-only LoRA setting is both feasible and different

8.7 What Separates "Good Thesis" from "Strong Paper"

Element	Good Thesis	Top Venue Paper
KKT derivation for LoRA	✓	✓
r^* formula (heuristic)	✓	✓
Phase transition heatmap	✓	✓
Provenance verification	✓	✓
r^* theorem (linear case)	—	✓
Real-world adapter demo	—	✓
Defense + utility tradeoff	—	✓
Contradicts CMU finding	—	✓

The bottom four rows are what separate a solid thesis from a submission that reviewers take seriously. Each one independently strengthens the paper; together they make it airtight.

8.8 Execution Order

1. **Phase 0 (1–2 days):** Gate experiment. KKT reconstruction on MLP + LoRA, $r > r^*$, $N \leq 10$. If this fails, the attack direction collapses — pivot to provenance-only thesis.
2. **Phase 1 (2 weeks):** Phase transition sweep on MLPs. Produce the heatmap figure. Validate or falsify r^*.
3. **Phase 2 (2–3 weeks):** Prove the linear-case theorem. This can run in parallel with Phase 1 (theory + experiments simultaneously).
4. **Phase 3 (2 weeks):** Scale to ViT + LoRA. Repeat the phase transition on a real architecture. Implement provenance verification.
5. **Phase 4 (1–2 weeks):** Real-world demo. Download public adapters, run the attack. Implement SVD truncation defense and measure the tradeoff.
6. **Phase 5 (1 week):** Write-up. With all results in hand, the paper writes itself: the story is already structured.

Total Timeline

8–10 weeks from Phase 0 to submission-ready draft. This is aggressive but realistic if Phase 0 succeeds. The proof (Phase 2) is the main risk — budget extra time or drop it for the thesis version and add it for the conference submission.

9 Next Step: Phase 0

Before any of this matters, one experiment must succeed:

Phase 0: The Gate Experiment

Run Haim et al.'s reconstruction on a small MLP + LoRA adapter with $r > r^*$ and $N \leq 10$. If this does not produce recognizable images, the KKT approach on LoRA collapses entirely.
This takes 1–2 days. Do it first.

References

[1] N. Haim, G. Vardi, G. Yehudai, O. Shamir, M. Irani. "Reconstructing Training Data from Trained Neural Networks." NeurIPS, 2022.

[2] D. Smorodinsky, G. Vardi, O. Safran. "Provable Privacy Attacks on Trained Shallow Neural Networks." arXiv:2410.15002, October 2024.

[3] N. Oz, G. Yehudai, G. Vardi, I. Antebi, M. Irani, N. Haim. "Reconstructing Training Data From Real-World Models Trained with Transfer Learning." arXiv:2407.15845, July 2024.

[4] E. Gronich, G. Vardi. "The Implicit Bias of Adam and Muon on Smooth Homogeneous Neural Networks." arXiv:2602.16340, February 2026.

[5] Y. Golbari, N. Wasserman, G. Vardi, M. Irani. "ImpMIA: Leveraging Implicit Bias for Membership Inference Attack under Realistic Scenarios." arXiv:2510.10625, October 2025.

[6] H.U. Sami et al. “Gradient Inversion Attacks on Parameter-Efficient Fine-Tuning.” CVPR, 2025.

[7] D. Ran et al. “LoRA-Leak: Membership Inference Attacks Against LoRA Fine-tuned Language Models.” arXiv, 2024.

[8] Y. Hu, J. Düngler, B. Schölkopf, A. Sanyal. “LoRA and Privacy: When Random Projections Help (and When They Don’t).” arXiv:2601.21719, January 2026.

[9] Liu et al. “Recover-to-Forget: Gradient Reconstruction from LoRA for Efficient LLM Unlearning.” NeurIPS Workshop, 2025.

[10] Y.-L. Tsai et al. “Differentially Private Fine-Tuning of Diffusion Models.” ICCV, 2025.

[11] “Extraction of Training Data from Fine-Tuned Large Language Models.” CMU Technical Report, 2024.