Theoretical Feasibility and Algorithmic Pathways for Training Data Reconstruction from Low-Rank Adaptation (LoRA) and Large Language Models: A Comprehensive Thesis Prospectus

1. Introduction: The Privacy Frontier in Parameter-Efficient Learning

The memorization capacity of deep neural networks has evolved from an empirical curiosity into a central pillar of learning theory and a critical vulnerability in data privacy. Modern deep learning models, particularly Large Language Models (LLMs) and diffusion models, are trained on dataset scales that defy manual curation, ingesting billions of tokens and images that inevitably contain sensitive personally identifiable information (PII), copyrighted material, and private communications. While the capacity of these models to generalize is the engine of their success, their capacity to memorize specific training examples poses a severe security risk.

The seminal work of Haim et al. (2022), titled *Reconstructing Training Data from Trained Neural Networks*, fundamentally altered the landscape of model inversion attacks.[1] Prior to this work, extracting training data typically required querying the model to obtain confidence scores or gradients (active attacks). Haim et al. demonstrated a more profound and disturbing result: the parameters of a trained binary classifier *implicitly encode* the training data samples themselves. By leveraging the implicit bias of gradient descent (GD)—specifically its tendency to converge to solutions satisfying the Karush-Kuhn-Tucker (KKT) conditions of a margin-maximization problem—they showed that high-fidelity reconstruction of training images is possible *solely from the model weights*, without any access to the training data or query access to the model during training.

This report serves as a foundational document for a Master's or Doctoral thesis in Computer Science aimed at extending the reconstruction paradigm established by Haim et al. (2022) into the era of Foundation Models. Specifically, we rigorously analyze the theoretical feasibility and algorithmic pathways for extending "weight-based reconstruction" to three frontier domains:

1. **Low-Rank Adaptation (LoRA)** in the Neural Tangent Kernel (NTK) regime.
2. **Discrete Sequence Reconstruction** from LLMs trained via Next-Token Prediction (NTP).

3. **Generative Priors**, utilizing diffusion models to regularize reconstruction in ill-posed settings.

As the field shifts from full fine-tuning to Parameter-Efficient Fine-Tuning (PEFT), methods like LoRA have become ubiquitous. LoRA reduces the number of trainable parameters by orders of magnitude ($r \ll d$). A critical, open research question is whether this massive compression acts as a privacy firewall—discarding the fine-grained details necessary for reconstruction—or merely provides a compressed, yet recoverable, encoding of sensitive data. Similarly, the shift from continuous image inputs to discrete text tokens in LLMs presents a non-trivial barrier to gradient-based reconstruction methods that assume differentiability.

This report synthesizes disparate results from optimization theory, differential privacy, and generative modeling. We dissect the mathematical stationarity conditions of LoRA updates, bridge the gap between low-rank adapters and full-model gradients using novel frameworks like "Recover-to-Forget" [2], and propose hybrid optimization objectives that fuse KKT residuals with Score Distillation Sampling (SDS) to hallucinate plausible reconstructions where information is lost. This analysis confirms that the building blocks for such a thesis are present in the current literature, and the combination of these elements represents a significant and viable research contribution.

---

2. Foundational Theory: The Geometric Mechanics of Reconstruction

To rigorously extend the work of Haim et al. (2022), one must first master the theoretical mechanism they established. Their reconstruction scheme is not a heuristic; it is a direct inversion of the optimization dynamics of gradient descent on separable data.

2.1 The Implicit Bias of Gradient Descent

The core premise relies on the phenomenon of **implicit bias** (or implicit regularization). In overparameterized neural networks, there are infinitely many parameter configurations θ that achieve zero training error. However, gradient descent does not select a solution uniformly at random. It converges to a very specific solution determined by the optimization geometry.

For a binary classification dataset $S = \{(x_i, y_i)\}_{i=1}^{n}$ and a homogeneous neural network $\Phi(\theta; \cdot)$, training with the logistic loss (binary cross-entropy) induces a specific directional convergence. Even though the loss $\mathcal{L}(\theta)$ approaches zero only as the norm of the parameters approaches infinity ($\|\theta\| \to \infty$), the *direction* of the parameter vector $\frac{\theta}{\|\theta\|}$ converges to a

stationary point of the maximum-margin problem.[1]

This is codified in the work of Lyu & Li (2019) and Ji & Telgarsky (2020), which states that under gradient flow, the normalized weights align with the direction that maximizes the margin normalized by the parameter norm.

2.2 The Stationarity Condition as a Reconstruction Key

The convergence to a maximum-margin solution implies that the final parameters satisfy the first-order stationarity conditions (KKT conditions) of the margin maximization problem.

Mathematically, this means the optimal parameter vector $\tilde{\theta}$ lies in the linear span of the gradients of the network with respect to the support vectors (the data points closest to the decision boundary).

The stationarity condition can be expressed as:

$$\frac{\theta}{\|\theta\|} \propto \sum_{i=1}^{n} \lambda_i y_i \nabla_\theta \Phi(\theta; x_i)$$

Here, $\lambda_i \geq 0$ are the Lagrange multipliers (dual variables). Crucially, the complementary slackness condition dictates that $\lambda_i > 0$ *only* for samples x_i that lie exactly on the margin (i.e., $y_i \Phi(\theta; x_i) = \gamma$). For all other samples, $\lambda_i = 0$.

The Reconstruction Insight:

Haim et al. (2022) recognized that this equation represents a system of constraints where typically the data x_i is known and the parameters θ are the result of optimization. However, in a post-training setting, the trained model parameters θ are known (observed), and the training samples x_i are the unknowns.

If the network is sufficiently overparameterized, this system of equations provides a rich signal. The reconstruction attack is formulated as an optimization problem: find a set of synthetic inputs $X' = \{x'_1, \ldots, x'_m\}$ and dual variables $\Lambda' = \{\lambda'_1, \ldots, \lambda'_m\}$ that minimize the residual of the stationarity condition:

$$\mathcal{L}_{\text{rec}}(X', \Lambda') = \quad \theta - \sum_{j=1}^{m} \lambda'_j y_j \nabla_\theta \Phi(\theta; x'_j) \quad ^2$$

When this loss is minimized to near zero, the synthetic inputs x'_j with non-zero λ'_j have been empirically shown to converge to the actual training samples x_i.[1]

2.3 Prerequisites for Extension

For this methodology to be successfully applied to new domains like LoRA or LLMs, three theoretical prerequisites must be examined:

1. **Homogeneity**: The original proofs rely on the network being homogeneous (e.g., ReLU networks without bias, or where bias is handled carefully). We must determine if LoRA-adapted Transformers satisfy or approximate this property.
2. **Margin Maximization Convergence**: We must verify if the training algorithm (e.g., LoRA fine-tuning with AdamW on cross-entropy) actually converges to a max-margin solution.
3. **Information Density**: The number of parameters (equations) must be sufficient to constrain the variables (unknown pixels or tokens). LoRA reduces the parameter count significantly, which potentially makes the system under-determined.

The remainder of this report systematically addresses these challenges for the three proposed thesis directions.

---

3. Thesis Direction 1: Low-Rank Adaptation (LoRA) in the NTK Regime

The first and most direct extension of the thesis is to apply the reconstruction framework to models fine-tuned with Low-Rank Adaptation (LoRA), specifically analyzing the problem under the Neural Tangent Kernel (NTK) regime. This section provides a deep theoretical analysis of why this is feasible and how the stationarity conditions adapt to the low-rank constraint.

3.1 LoRA Parameterization and Gradient Dynamics

LoRA freezes the pre-trained weights $W_0 \in \mathbb{R}^{d \times k}$ and injects trainable rank-r matrices $B \in \mathbb{R}^{d \times r}$ and $A \in \mathbb{R}^{r \times k}$, such that the forward pass for a layer becomes:

$$h = (W_0 + BA)x$$

where $r \ll \min(d, k)$. The effective weight update is $\Delta W = BA$.

Unlike full fine-tuning (FFT), where the optimization variable is the full matrix W, LoRA optimizes A and B. The gradients with respect to these adapters are derived via the chain rule from the gradient of the accumulated weight matrix $\nabla_W \mathcal{L}$:

$$\nabla_A \mathcal{L} = B^\top \nabla_W \mathcal{L}$$

$$\nabla_B \mathcal{L} = \nabla_W \mathcal{L} A^\top$$

This relationship is pivotal. It tells us that the gradients driving the updates of A and B are *projections* of the "ideal" full-rank gradient $\nabla_W \mathcal{L}$ onto the subspaces spanned by B and A.

3.2 Stationarity Conditions in the LoRA Regime

To extend Haim et al.'s work, we must formulate the stationarity condition for LoRA. According to recent convergence analyses [4], the stationarity of LoRA is often analyzed by stacking A and B into a single variable matrix. However, we can derive the condition directly from the gradients.

At a stationary point (convergence), the gradients with respect to the trainable parameters must vanish (or be parallel to the weight vector for exponential loss). Thus:

$$B^\top (\nabla_W \mathcal{L}) = 0 \quad \text{and} \quad (\nabla_W \mathcal{L}) A^\top = 0$$

If we assume the implicit bias of the underlying loss function (e.g., cross-entropy) drives the *effective* weight gradient $\nabla_W \mathcal{L}$ to align with a linear combination of support vectors (as in the full-rank case), then we can substitute $\nabla_W \mathcal{L} \approx \sum \lambda_i y_i x_i x_i^\top$ (for a linear network).

This yields a coupled system of equations for reconstruction:

$$B^\top \left(\sum_{i=1}^{n} \lambda_i y_i \nabla_W \Phi(\theta; x_i) \right) = 0$$

$$\left(\sum_{i=1}^{n} \lambda_i \, y_i \, \nabla_W \, \Phi(\theta; x_i) \right) A^\top = 0$$

Interpretation: This does *not* imply that the linear combination of data gradients is zero (as in full fine-tuning stationarity). Instead, it implies that the "residual" signal of the training data lies in the null space of $B^\top$ and the left null space of $A^\top$.

However, there is a second, more powerful constraint. The trained weights A and B themselves are the result of accumulated gradient updates. If training starts from $A = 0, B = 0$ (or random initialization), the final values of A and B are time-integrals of their gradients.

$$A_{final} \approx -\eta \sum_{t=0}^{T} B_t^\top \nabla_{W,t} \mathcal{L}$$

$$B_{final} \approx -\eta \sum_{t=0}^{T} \nabla_{W,t} \mathcal{L} A_t^\top$$

This means A and B are low-rank compressions of the history of data gradients. This is the "memory" of the training data.

3.3 LoRA in the NTK Regime: A Theoretical Sweet Spot

The thesis proposal explicitly mentions the **Neural Tangent Kernel (NTK) regime**. This is a specific mode of training where the network width is very large, and the weights move very little from initialization.

Recent Theoretical Breakthroughs: Recent work by Jang et al. (2024), *LoRA Training in the NTK Regime has No Spurious Local Minima*, provides a crucial theoretical foundation for this thesis direction.[5] They prove that:

1. Full fine-tuning in the NTK regime admits a low-rank solution with rank $r \lesssim N$ (where N is the number of data points).
2. If the LoRA rank is sufficiently large ($r \gtrsim$), LoRA optimization eliminates spurious local minima and converges to the *same* global minimum as full fine-tuning.

Implication for Reconstruction:

This is a "smoking gun" for reconstruction feasibility. If LoRA converges to the *same* functional solution as full fine-tuning in the NTK regime, then the resulting effective weights $\Delta W = BA$ must encode the same information as the full weights in FFT. Since Haim et al. proved that FFT weights encode training data via KKT conditions, it follows that **LoRA weights (BA) must also encode the training data**, provided the rank is sufficient to represent the support vectors.

If the rank r is lower than the intrinsic rank of the support vectors, LoRA acts as a bottleneck. It will learn the principal components of the support vectors. Reconstruction in this case would yield "eigen-samples"—images that represent the dominant features of the training set rather than individual samples. This aligns with the "Model Inversion" phenomenon where class averages are recovered.

3.4 Gradient Reconstruction as a Bridge

A critical intermediate step for the thesis is bridging the gap between the low-rank adapters (A, B) and the full-rank gradients required for Haim et al.'s method.

The "Recover-to-Forget" (R2F) Insight: Very recent work by Liu et al. (Dec 2025) titled *Recover-to-Forget* [2] demonstrates that it is possible to reconstruct *full-model gradients* from LoRA updates. They train a "Gradient Decoder" network that takes LoRA parameters as input and predicts the full gradient vector.

- *Key Insight*: They show that calculating gradients with respect to LoRA parameters is effectively a "sensing" operation of the full gradient.
- *Reconstruction Pipeline*:
    1. **Step 1**: Given trained A, B, estimate the accumulated full-rank update $\Delta \hat{W} \approx BA$.
    2. **Step 2**: Apply Haim et al.'s reconstruction loss on $\Delta \hat{W}$. Treat $\Delta \hat{W}$ as the "model parameters" θ in their equation.
    3. **Step 3**: Optimize for inputs x' such that $\Delta \hat{W} \approx \sum \lambda_i \nabla \Phi(x')$.

This approach bypasses the complex stationarity of the product manifold by lifting the problem back to the full parameter space where the linear combination property holds.

---

4. Thesis Direction 2: Reconstructing Sentences from LLMs

The second proposed extension is reconstructing discrete sentences from LLMs. This transitions the problem from the continuous domain of images to the discrete domain of text, and from binary classification to Next-Token Prediction (NTP).

4.1 Implicit Bias of Next-Token Prediction (NTP)

To apply Haim et al.'s method, one must understand the stationarity conditions of the NTP loss. Thrampoulidis (2024) recently characterized the **Implicit Optimization Bias of Next-Token Prediction**.[7]

Key Theoretical Findings:

- NTP can be framed as a multiclass classification problem where the number of classes is the vocabulary size V.
- For linear models, gradient descent on NTP converges to a direction that maximizes a specific "NTP-margin."
- The solution satisfies **NTP-separability conditions**: the weights converge to a structure that separates distinct "support contexts."

Implication:

The weights of a fine-tuned LLM implicitly encode the "support contexts"—the specific prompt-completion pairs in the training data that were most difficult to predict (i.e., had the highest loss). This provides the theoretical justification that the information is *there*. The weights are effectively a linear combination of the gradients of these support contexts.

4.2 Overcoming the Discrete Barrier

The Haim et al. algorithm optimizes inputs x' via gradient descent. This works for pixels but fails for discrete tokens because the operation token_id -> embedding is non-differentiable. To reconstruct sentences, the thesis must employ **continuous relaxation strategies**.

Proposed Methodologies:

1. **Embedding Space Optimization**: Instead of optimizing discrete tokens, optimize a sequence of continuous embedding vectors $E' =$. The reconstruction loss minimizes the KKT residual with respect to these embeddings.
    - *Challenge*: The optimized embeddings might not map back to valid tokens.
    - *Solution*: Regularize the embeddings to lie close to the valid token embedding manifold, or use a "projection" step periodically.

2. **Gumbel-Softmax Relaxation**: This is a standard technique in differentiable search.[9] We represent the input as a probability distribution over the vocabulary (softmax). Using the Gumbel-Softmax trick allows gradients to flow from the reconstruction loss back to the distribution parameters, enabling the optimization of the "most likely" tokens that satisfy the KKT conditions.
3. **Language Model Priors (LAMP/DAGER)**: Recent attacks like **LAMP** (Language Model Priors) [11] and **DAGER** [12] show that text reconstruction is vastly improved by using an auxiliary LLM to guide the optimization.
    - *Algorithm*: Minimize $\mathcal{L}_{\text{rec}}(x) + \lambda \mathcal{L}_{\text{prior}}(x)$, where $\mathcal{L}_{\text{prior}}$ is the perplexity of the reconstructed text under a pre-trained (frozen) LLM.
    - This ensures that the reconstructed gradients resolve into coherent sentences rather than gibberish that happens to satisfy the mathematical constraints.

4.3 Feasibility Analysis: Parameter Counting

A critical feasibility check for this thesis direction is the ratio of equations (parameters) to unknowns (data).

- **Case Study**: Fine-tuning Llama-2-7B ($d = 4096$) with LoRA rank $r = 8$ on a dataset of $N = 100$ sentences of length $L = 64$.
- **Unknowns (Data)**: The data consists of $N \times L = 6,400$ tokens.
- **Equations (Parameters)**: LoRA introduces trainable parameters $P \approx 2 \times L_{\text{layers}} \times d \times r$. For a single layer, $2 \times 4096 \times 8 \approx 65,000$ parameters.
- **Conclusion**: The system is **highly overdetermined**. The number of constraints (parameters) far exceeds the number of unknowns (tokens). Even with low-rank adapters, there is theoretically more than enough information capacity in the weights to uniquely define the small fine-tuning dataset. This suggests reconstruction is highly feasible in the few-shot fine-tuning regime.

---

5. Thesis Direction 3: Strong Priors and Diffusion-Based Inversion

The user proposed using "stronger priors" (e.g., face priors) to aid reconstruction. This is a cutting-edge direction that aligns with the concept of **Generative Model Inversion**.

5.1 The Problem of Ill-Posed Inversion

When the LoRA rank is extremely low (e.g., $r = 1$ or 2), the projection BA discards significant spatial information. The system of equations becomes underdetermined: multiple

different images could theoretically produce the same low-rank gradient update. To solve this, we need a strong regularizer to bias the solution toward the manifold of "natural images" or "faces."

5.2 Diffusion Models as Priors

Diffusion models (like Stable Diffusion) learn the implicit probability density of data $p(x)$. We can use a pre-trained diffusion model as a plug-and-play prior for the reconstruction optimization.

Score Distillation Sampling (SDS): This technique, popularized by *DreamFusion* [13], allows optimizing an image x such that it looks like a sample from a diffusion model, without generating it directly.

- *Mechanism*: We compute the gradient $\nabla_x \mathcal{L}_{SDS}$ which pushes the image x towards higher probability regions of the diffusion model (e.g., making it look more like a realistic face).

The Combined Objective:

The thesis can propose a novel hybrid loss function:

$$\mathcal{L}_{\text{total}}(x) = \underbrace{\|\theta_{\text{LoRA}} - \sum \lambda_i \nabla \Phi(x)\|^2}_{\text{KKT / Data Consistency}} + \lambda_{\text{SDS}} \underbrace{\mathcal{L}_{\text{SDS}}(x)}_{\text{Generative Prior}}$$

- The **KKT term** ensures the reconstructed image is one that *could have caused* the observed model updates. It provides the "hard evidence" or constraints.
- The **SDS term** hallucinates the missing high-frequency details (texture, lighting) that were lost due to the low-rank compression, ensuring the result is perceptually realistic.

5.3 LoRA-Fused Training-Data Generation (LoFT)

A supporting piece of evidence for this approach is the paper *LoRA-Fused Training-data Generation* (LoFT).[15] This work essentially does the inverse: it uses fine-tuned LoRA weights to generate synthetic images that act as proxies for the training data.

- *Insight*: If LoRA weights can be used to *generate* training-like data, then the inversion process is fundamentally sound. The thesis would simply formalize this as an optimization attack rather than a generative application.

---

6. Detailed Analysis of Reconstruction Feasibility from

Research Snippets

To ensure the thesis proposal is grounded in the provided research materials, we present a structured analysis of the key snippets and their implications.

Research Domain	Key Insight from Snippets	Implication for Thesis	Source
LoRA Theory	LoRA in NTK regime has no spurious local minima if $r \gtrsim$. It converges to the same functional solution as FFT.	Justifies that LoRA weights encode the same support vector geometry as full weights.	[5]
LoRA Gradients	Full-model gradients can be reconstructed from LoRA updates using a decoder trained on a proxy model (R2F).	Provides a concrete algorithm to "expand" LoRA weights back to full weights for reconstruction.	[2]
Implicit Bias	GD on Matrix Factorization (UV^T) minimizes nuclear norm.	Suggests LoRA learns the principal components of the data gradient covariance.	[16]
LLM / NTP	Next-Token Prediction biases linear models to max-margin solutions separating distinct contexts.	Confirms that LLM weights implicitly encode "support contexts" (training sentences).	[7]
Attacks	Discrete	Provides the	[11]

	optimization (LAMP, DAGER) can recover text from gradients.	"backend" solver needed once LoRA weights are converted to gradients.	
Priors	Score Distillation Sampling (SDS) allows using diffusion models to regularize inverse problems.	Solves the "low-rank information loss" problem by hallucinating missing details.	[13]

---

7. Experimental Roadmap and Recommendations

Based on the theoretical analysis, we propose a concrete experimental roadmap for the thesis.

7.1 Phase 1: Verification on Toy Models (LoRA NTK)

- **Setup**: Train a simple MLP with LoRA on a 2D synthetic dataset (like the "two moons" or "circles" used in Haim et al.).
- **Experiment**: Verify if the LoRA adapters A, B align with the support vectors.
- **Goal**: Empirically validate the stationarity condition derived in Section 3.2. Show that reconstruction works perfectly when r is high and degrades gracefully as r decreases.

7.2 Phase 2: Reconstruction from Vision Transformers (ViT)

- **Setup**: Fine-tune a ViT-Base using LoRA on a small subset of CIFAR-10 or CelebA (e.g., 50-100 images).
- **Method**: Use the "Gradient Bridge" approach. Treat the trained BA matrix as a gradient observation. Optimize inputs to match this gradient.
- **Extension**: Introduce the Diffusion Prior (SDS). Use a frozen Stable Diffusion model to guide the reconstruction of faces from the CelebA-trained adapter.
- **Hypothesis**: The combination of LoRA constraints and Diffusion priors will reconstruct recognizable faces even at very low ranks ($r = 4$).

7.3 Phase 3: Sentence Reconstruction from LLMs

- **Setup**: Fine-tune a small LLM (e.g., GPT-2 or Llama-tiny) on a set of 10-20 "canary" sentences containing random PII.

- **Method**: Implement the R2F strategy to estimate the full gradient direction from the LoRA adapter. Then, use a discrete search (like LAMP) to find sentences that produce this gradient.
- **Goal**: Demonstrate that "safe" PEFT fine-tuning can leak verbatim training sentences.

7.4 Discussion of Limitations

The thesis must candidly address limitations.

- **Rank Bottleneck**: If r is too small, information is strictly lost. The reconstruction will effectively be a "compression" of the dataset (e.g., the average face).
- **Optimization Difficulty**: Reconstructing discrete text is an NP-hard problem; relaxations are approximate and sensitive to initialization.
- **Memory**: While LoRA is memory-efficient, the *attack* (reconstruction) requires backpropagating through the model, which can be expensive.

---

8. Conclusion

The proposed thesis topic is theoretically robust and sits at the cutting edge of current research in trustworthy AI. The convergence of results from **implicit bias theory** (guaranteeing the signal exists), **LoRA dynamics** (guaranteeing the signal is preserved in the adapters), and **generative priors** (providing the tools to retrieve it) creates a fertile ground for discovery.

By proving that LoRA adapters are not merely functional updates but *compressed artifacts of training data*, this research will have profound implications for the privacy of fine-tuned Foundation Models. It suggests that distributing a LoRA adapter may be functionally equivalent to distributing the private dataset itself, necessitating new defense mechanisms such as differential privacy or gradient sanitation even for parameter-efficient learning.

This report confirms that the "data reconstruction from LoRA" hypothesis is not just plausible—it is the logical conclusion of the current trajectory of deep learning theory.

עבודות שצוטטו

1. THE_PAPER.pdf
2. Recover-to-Forget: Gradient Reconstruction from LoRA for Efficient LLM Unlearning - arXiv, 2026 ,18 נרשמה גישה בתאריך ינואר, https://arxiv.org/html/2512.07374v1
3. Implicit Bias of Deep Learning Optimization: A Mathematical Examination - DataSpace, 2026 ,18 נרשמה גישה בתאריך ינואר, https://dataspace.princeton.edu/handle/88435/dsp01xs55mg46c
4. [2512.18248] On the Convergence Rate of LoRA Gradient Descent - arXiv, נרשמה גישה בתאריך ינואר 18, 2026, https://arxiv.org/abs/2512.18248

5. LoRA Training in the NTK Regime has No Spurious Local Minima - Princeton University, 2026 ,18 נרשמה גישה בתאריך ינואר, https://collaborate.princeton.edu/en/publications/lora-training-in-the-ntk-regime-has-no-spurious-local-minima/
6. LoRA Training in the NTK Regime has No Spurious Local Minima - arXiv, נרשמה 2026 ,18 גישה בתאריך ינואר, https://arxiv.org/html/2402.11867v3
7. [2402.18551] Implicit Optimization Bias of Next-Token Prediction in Linear Models - arXiv, 2026 ,18 נרשמה גישה בתאריך ינואר, https://arxiv.org/abs/2402.18551
8. Implicit Bias of Next-Token Prediction - arXiv, 2026 ,18 נרשמה גישה בתאריך ינואר, https://arxiv.org/html/2402.18551v1
9. Universal Adversarial Suffixes Using Calibrated Gumbel–Softmax Relaxation - arXiv, 2026 ,18 נרשמה גישה בתאריך ינואר, https://arxiv.org/html/2512.08123v1
10. Security-Enhanced and Privacy-Preserving Federated Learning by Tianyue Chu - IMDEA Networks Principal, 2026 ,18 נרשמה גישה בתאריך ינואר, https://dspace.networks.imdea.org/bitstream/handle/20.500.12761/1957/PhD_Thesis_Tianyue_Chu.pdf?sequence=1
11. LAMP: Extracting Text from Gradients with Language Model Priors, נרשמה גישה 2026 ,18 בתאריך ינואר, https://proceedings.neurips.cc/paper_files/paper/2022/file/32375260090404f907ceae19f3564a7e-Paper-Conference.pdf
12. DAGER: Exact Gradient Inversion for Large Language Models - OpenReview, 2026 ,18 נרשמה גישה בתאריך ינואר, https://openreview.net/forum?id=CrADAX7h23&referrer=%5Bthe%20profile%20of%20Martin%20Vechev%5D(%2Fprofile%3Fid%3D~Martin_Vechev1)
13. Rethinking Score Distillation as a Bridge Between Image Distributions - NIPS papers, 2026 ,18 נרשמה גישה בתאריך ינואר, https://proceedings.neurips.cc/paper_files/paper/2024/file/3b62bca132cf5c8973b09a2fc6dc8ca6-Paper-Conference.pdf
14. DiffusionBlend: Learning 3D Image Prior through Position-aware Diffusion Score Blending for 3D Computed Tomography Reconstruction - NIPS papers, נרשמה 2026 ,18 גישה בתאריך ינואר, https://proceedings.neurips.cc/paper_files/paper/2024/file/a30769d9b62c9b94b72e21e0ca73f338-Paper-Conference.pdf
15. LoFT: LoRA-Fused Training Dataset Generation with Few-shot Guidance - arXiv, 2026 ,18 נרשמה גישה בתאריך ינואר, https://arxiv.org/html/2505.11703v1
16. Implicit Bias of Gradient Descent - Emergent Mind, ,18 נרשמה גישה בתאריך ינואר 2026, https://www.emergentmind.com/topics/implicit-bias-of-gradient-descent
17. Implicit Regularization in Deep Matrix Factorization - Princeton University Library, 2026 ,18 נרשמה גישה בתאריך ינואר, https://oar.princeton.edu/bitstream/88435/pr1qs0p/1/ImplicitRegularizationDeepMatrixFactor.pdf
18. [2407.11424] Model Inversion Attacks Through Target-Specific Conditional Diffusion Models, 2026 ,18 נרשמה גישה בתאריך ינואר, https://arxiv.org/abs/2407.11424