

inLoRA-MINT: Internal Representations for Auditing Training Data in Domain-adapted LLMs

Gonzalo Mancera*, Julian Fierrez*, Aythami Morales*[‡], Ruben Tolosana*, Daniel DeAlcala*, Francisco Jurado[†]

*BiometricsAI, Universidad Autonoma de Madrid, Spain

[†]Computer Engineering Dept., Universidad Autonoma de Madrid, Spain

[‡]Mathematics Dept., Universidad de Las Palmas de Gran Canaria, Spain

Abstract—This work investigates the application of the Membership Inference Test (MINT) to large language models (LLMs) fine-tuned with LoRA. MINT aims to determine whether specific data samples were used during model training, addressing key concerns related to data privacy, data leakage, and model transparency. This work builds on our previous LoRA-MINT framework based on perplexity, which addressed this same auditing problem under a strict black-box assumption, relying exclusively on the perplexity of candidate samples compared against a synthetic reference distribution in the domain. We propose an approach based on extracting intermediate representations from the model, capturing internal behaviors such as activations and parameter-dependent signals. These representations enable the construction of a MINT model that learns to distinguish between training and non-training samples, without requiring full access to the model parameters. We evaluated the method in 4 LLMs (Qwen, Phi-4, DeepSeek-V3, and LLaMA 3.2-3B) fine-tuned in three datasets covering medical, mathematical, and academic reasoning tasks. The results show that MINT is effective across different architectures and domains, achieving performance scores ranging from 95% to 98%, and demonstrating robustness under varying model configurations and data distributions. By allowing efficient auditing of whether sensitive records were used during fine-tuning without requiring full model access, the proposed MINT approach provides a practical mechanism for data provenance in accordance with emerging privacy and AI governance regulations in real-world LLM deployments. For the sake of concreteness and current relevance, our discussion and experiments are centered on LoRA-adjusted LLMs, but note that most of the presented methodology is easily applicable for auditing training data given any other technique for adapting LLMs or, more generally, any AI model.

Index Terms—Membership Inference Test, Large Language Models, LoRA, Model Auditing, Data Privacy, Trustworthy AI,

I. INTRODUCTION

LLMs and VLMs are now revolutionizing multiple domains, including healthcare [1], [2], law [3], [4], human resources management [5], [6], education [7], [8], and finance [9], significantly transforming the way textual information is processed and interpreted. These recent advances have enabled the analysis of large-scale textual corpora, accelerating progress in NLP applications. However, these models require vast amounts of data during training, and the provenance and public availability of such data remain controversial.

Support by M2RAI (PID2024-160053OB-I00, MICIU/FEDER) and Cátedra ENIA UAM-VERIDAS en IA Responsable (NextGenerationEU PRTR TSI-100927-2023-2). Mancera is supported by FPI-PRE2022-104499 MICINN/FEDER and DeAlcala is supported by FPU21/05785 MIU.

Protecting sensitive information in AI applications [10], [11] is now critical in order to preserve user trust and ensure compliance with increasingly stringent social concerns about the proper use of AI systems [12]. In response to these concerns, the European Union introduced new regulations in June 2024¹, aimed at governing the deployment of artificial intelligence systems [13]. This regulatory framework reflects a larger effort to address the societal and ethical implications of AI technologies, particularly in NLP applications.

A central objective of this legislation is the protection of fundamental rights through mechanisms that promote transparency and accountability in the design and use of AI systems. Consequently, both this framework and related international guidelines emphasize the need for auditing methodologies that enable the responsible deployment of AI/NLP technologies, while also supporting a better understanding of their behavior [14] and potential risks [13].

In the context of increasing privacy concerns in AI deployments [15], a related work proposed Membership Inference Attacks (MIAs), which study whether an adversary can infer if specific data were used during model training [16]. Previous work has shown that such attacks are feasible and has proposed a variety of mitigation and evaluation strategies.

Beyond traditional MIAs, recent work on privacy in machine learning has studied the possibility of inferring whether specific data samples were used during training, highlighting both the feasibility of such an inference and its implications for data protection and model leakage [16], [17]. These approaches typically operate under adversarial assumptions, where an attacker exploits the behavior of the model to extract information about the training dataset.

In line with recent research oriented to develop new technologies for auditing LLMs, the present paper addresses the challenge of identifying whether a given text was utilized in the training of a Language Model (see Fig. 1). The study aims to further highlight the versatility and applicability of the general MINT framework [18] by adapting it to LoRA fine-tuned LLMs and demonstrating its effectiveness in language generation scenarios [19], [20]. In this line, we extend our previous work LoRA-MINT [19], which relied solely on perplexity statistics computed from a synthetic reference distribution, by incorporating: 1) information from internal activations

¹<https://artificialintelligenceact.eu>

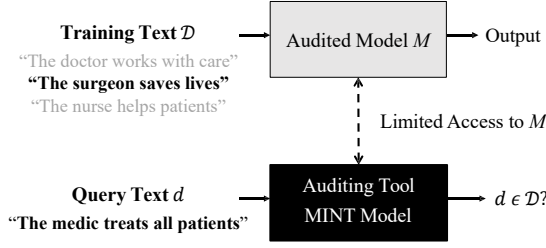


Fig. 1. MINT discerns whether given data (d) were utilized in the training process of an AI Model (M) trained with a specific database ($\mathcal{D}$).

(‘in’ in inLoRA-MINT is from ‘internal’), 2) gradient-based sensitivity, and 3) output-level signals; trading a mild increase in required model access for consistently higher detection performance across the same model families and datasets.

The main contributions of this work are as follows.

- We adapt the MINT framework to LoRA fine-tuned LLMs using internal representations, enabling membership inference without full access to model parameters.
- We evaluate 4 LoRA fine-tuned LLMs (**Qwen**, **Phi-4**, **DeepSeek-V3**, and **LLaMA 3.2-3B**) across three datasets (medical, mathematical, and academic), analyzing the effect of architecture, scale, and domain.
- Results show that this representation-based approach effectively distinguishes training from non-training samples across models and domains.

For the sake of concreteness and current relevance, our discussion and experiments are centered on LoRA-adjusted LLMs, but note that the presented methodology is easily applicable for auditing training data given any other technique for adapting LLMs or, more generally, any AI model.

II. RELATED WORKS

A. Membership Inference Test (MINT)

MINT (Membership Inference Test) [18] is a framework designed to evaluate whether a given machine learning model was trained or not on given data. Its main use as a data training auditing tool assumes some level of cooperation from the model provider, so that the development and implementation of the model follow legal and ethical requirements.

The primary goal of MINT is to establish reliable and broadly applicable evaluation procedures that can operate across multiple data modalities, including text, image, and audio domains. These procedures are designed to accommodate different degrees of transparency, depending on the amount of information disclosed by the model developer about the training process and the behavior of the system [18], [21]. Consequently, the framework spans a wide spectrum of settings, ranging from highly opaque scenarios with limited or no disclosure to fully transparent environments where complete model details are available [22], [23].

Within the MINT framework, auditors evaluate a model with full access to it or limited information about its architecture and training process. This makes MINT particularly relevant

for practical compliance settings, especially in the context of evolving data protection regulations and legal requirements.

Unlike traditional Membership Inference Attacks (MIAs), which often rely on training shadow models to approximate the behavior of the target model, MINT operates directly on the original model. This leads to different MINT methodologies depending on the amount of information available from the model being audited. As a result, the effectiveness of MINT varies depending on factors such as the level of access to the model and the type of data modality, making it adaptable to a wide range of audit scenarios [22]–[25].

B. Low-Rank Adaptation (LoRA)

Low-Rank Adaptation (LoRA) [26] is a parameter-efficient fine-tuning (PEFT) method that facilitates the specialization of massive pre-trained language models without necessitating a full gradient update of all network parameters.

Instead of modifying the primary weight tensors, LoRA introduces a bottleneck structure consisting of two low-rank matrices that approximate the weight residuals within a significantly lower-dimensional manifold. Formally, for any static weight matrix $W_0 \in \mathbb{R}^{d \times k}$, LoRA keeps W_0 frozen and optimizes two auxiliary matrices $A \in \mathbb{R}^{d \times r}$ and $B \in \mathbb{R}^{r \times k}$, where the rank $r \ll \min(d, k)$.

The final adapted weight is defined by:

$$W = W_0 + BA \quad (1)$$

This low-rank factorization drastically diminishes the volume of optimizeable parameters—frequently by multiple orders of magnitude—without compromising the model’s effectiveness on target tasks. By restricting gradient updates solely to LoRA adapters, the foundational weights remain unchanged, which facilitates efficient storage, modularity, and rapid switching between different tasks.

Furthermore, LoRA fundamentally alters the optimization landscape during the fine-tuning process, which may impact the mechanisms of data memorization and internal representation. This unique property makes LoRA-tuned architectures a compelling subject for MINT and privacy-aware technologies [27], [28], as the distinct partitioning between static and trainable components could influence how sensitive information is distributed throughout the model.

III. EXPERIMENTAL FRAMEWORK

A. Problem Formulation

Given a training dataset $\mathcal{D}$ and an external dataset $\mathcal{E}$, we consider samples $d \in \mathcal{D} \cup \mathcal{E}$. We consider a base model M trained for a specific task, which is further fine-tuned using LoRA, resulting in M_{LoRA} . For each input sample d , the model produces an output y as a function of d and the learned parameters w , i.e.,

$$y = M_{\text{LoRA}}(d \mid w).$$

In this setting, we assume access to intermediate representations extracted from the model, denoted as $q(d)$, which correspond to parameter-dependent signals such as activations,

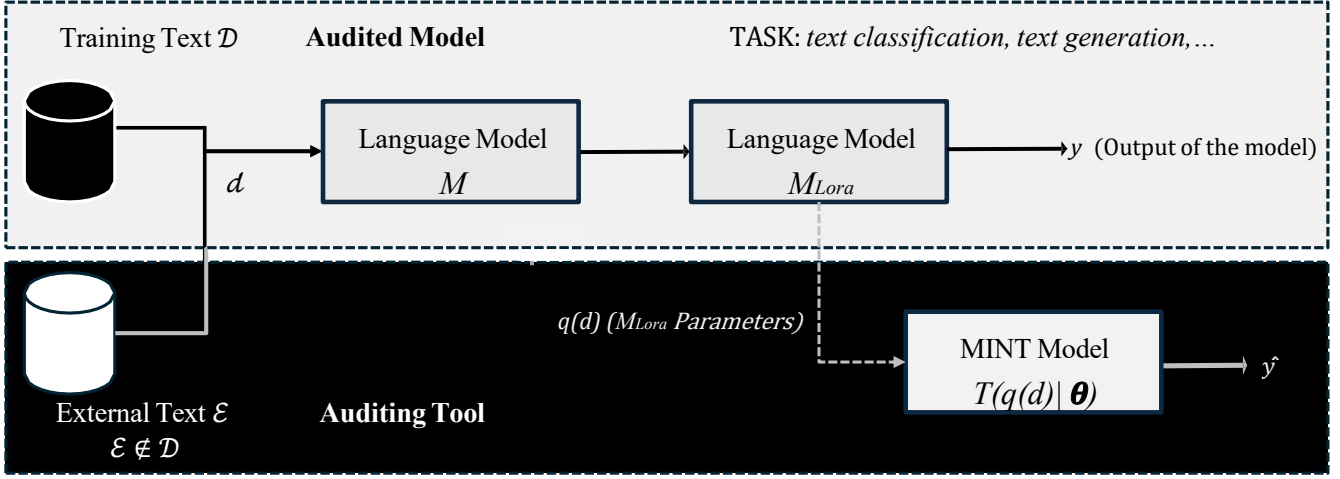


Fig. 2. The MINT model $T(q(d) | \theta)$ is trained to determine whether a given data instance d was part of the training data $\mathcal{D}$ used to develop a language model M_{LoRA} . The inputs to the MINT model are represented by $q(d)$, which consists of features extracted from M_{LoRA} , including activation-based statistics, gradient-derived sensitivity measures, and output-level features computed with respect to the sample d .

internal states, or subsets of model parameters computed when processing d . These representations capture how the model responds internally to different inputs.

We define the MINT model as a parametric function:

$$T(q(d) | \theta),$$

where θ are the learnable parameters of the model.

The MINT model is trained using representations $q(d)$ calculated on both datasets $\mathcal{D}$ and $\mathcal{E}$, ensuring that it captures differences in the internal behavior of the model for training and non-training samples.

The MINT model then predicts membership as follows:

$$\hat{y} = T(q(d) | \theta) = \begin{cases} 1 & \text{if } d \in \mathcal{D}, \\ 0 & \text{if } d \in \mathcal{E}. \end{cases}$$

This formulation relies on the use of partial parameter-level information $q(d)$ associated with each sample to distinguish between training and non-training data. This setting assumes a certain degree of collaboration from the model developer, as access to this parameter-level information is required.

As a result, this approach aligns with the principles of transparency and trustworthiness, enabling controlled auditing mechanisms in scenarios where developers expose only minimal and carefully selected internal signals. This limited disclosure is designed to avoid compromising the integrity, performance, or confidentiality of the underlying model while still supporting an effective audit.

The method can be summarized as a pipeline: (i) construct a reference set by collecting parameter-level representations $q(d)$ from samples with known membership labels, (ii) use these representations to train the MINT model $T(\cdot | \theta)$ as a binary classifier, (iii) extract the corresponding representations $q(d)$ for each candidate sample $d \in \mathcal{C}$, and (iv) predict membership by evaluating $T(q(d) | \theta)$.

IV. EXPERIMENTAL FRAMEWORK

We begin by introducing the text datasets used in this study, which consist of three datasets covering instruction-based text generation and reasoning tasks across different domains. We then describe the 4 LLMs selected for auditing, detailing their main characteristics and configurations. Finally, we present the experimental protocol followed in this work.

A. Databases

- *Medical-ol-reasoning-SFT* [29]. A medical reasoning dataset with verifiable question-answer pairs enriched with structured reasoning trajectories, supporting diagnostic and analytical reasoning in healthcare scenarios.
- *CAMEL Math Reasoning* [30]. A synthetic multi-agent dataset built from role-playing agent dialogs, emphasizing step-by-step problem-solving across mathematical tasks.
- *Maths & College Reasoning* [31]. A college-level dataset spanning algebra, calculus, probability, statistics, and differential equations, formatted as instruction-response pairs to evaluate multi-step reasoning.

B. Audited Large Language Models

Different models M are evaluated to analyze whether variations in the architecture, scale, and complexity of the audited model affect the performance of MINT. In this study, we consider a set of LLMs that have been fine-tuned using LoRA, denoted M_{LoRA} , which allows efficient adaptation while preserving the capabilities of the base model. The fine-tuning process is performed using the datasets described in Section IV-A, ensuring consistency between experimental conditions.

The selected models are chosen to provide a diverse and representative comparison between recent large language model families. These include **Qwen** [32], **Phi-4** [33], **DeepSeek-V3** [34], and **LLaMA 3.2-3b** [35]. Together, these models span

different design choices in terms of scale, training data, and architectural refinements.

C. Experimental Protocol

The audited model M_{LoRA} is obtained by fine-tuning a base language model M using Low-Rank Adaptation (LoRA) on a training set of approximately 13k samples. The LoRA configuration employs a rank of $r = 32$, a scaling factor $\alpha = 32$, and a dropout rate of 0.15, introducing trainable low-rank updates into selected transformer modules (including query, key, value, output, and feed-forward projections) while preserving the original model parameters. The fine-tuning process is carried out for 10 epochs using an AdamW-based optimizer with a learning rate of 10^{-4} .

To characterize the behavior of the audited model M_{LoRA} with respect to each input sample d , we extract a fixed-dimensional feature representation that captures multiple aspects of the model’s internal and output responses.

First, we calculate the statistical descriptors from the internal activations of the last three layers of M_{LoRA} . Let $h^{(l)}(d) \in \mathbb{R}^{T \times D}$ denote the hidden representation in layer l when processing sample d , where T is the sequence length and D is the hidden dimension. For each selected layer, we derive 4 summary statistics based on the token-averaged representation: 1) the mean value in all feature dimensions, 2) the standard deviation in all feature dimensions, 3) the ℓ_2 norm of the averaged representation, and 4) the average ℓ_2 norm computed in all token-level representations. These features capture both the global magnitude and the variability of internal activations. This results in 4 features per layer and a total of 12 activation-based features.

Second, we incorporate a sensitivity-related characteristic computed as the ℓ_2 norm of the gradient with respect to the input embeddings of M_{LoRA} , resulting in a single scalar.

Finally, we include output-based features obtained from $M_{\text{LoRA}}(d | w)$, namely prediction loss and a confidence score calculated as the maximum predicted probability averaged between tokens. These contribute two additional features.

In general, each input sample d is represented by a dimension vector of characteristics 15, combining internal activation statistics (12), a gradient-based sensitivity measure (1) and the output signal (2). This 15-dimensional representation is obtained in the experimental setting where features are extracted from the last three layers of M_{LoRA} , producing 4 statistical descriptors per layer ($4 \times 3 = 12$). This enables a comprehensive characterization of the model for subsequent membership inference.

The MINT model is a fully connected neural network that operates on the extracted feature vectors. The input dimensionality depends on the feature configuration used in each experimental setting: it is 15 when using the full representation based on the last three layers of M_{LoRA} , 7 when using a single-layer configuration and 1 when using individual signals such as gradient, loss, or probability features independently. The network consists of two hidden layers with 32 and 16 neurons, respectively. Each layer is followed by batch normalization, a

SiLU activation function, and dropout regularization with rates of 0.2 and 0.1, respectively. The final layer outputs a single scalar value corresponding to the predicted membership score.

The model is trained using a binary cross-entropy loss with logits for stable binary classification. The optimization is performed using the AdamW optimizer with a learning rate of 3×10^{-4} and a weight decay of 10^{-4} , and the training process is carried out over 30 epochs.

V. EXPERIMENTS AND RESULTS

To systematically evaluate the effectiveness of different feature configurations for membership inference, we design six experimental settings that progressively incorporate deeper representations extracted from the audited model M_{LoRA} . Experiments are conducted on three datasets (*Medical-o1-reasoning-SFT*, *CAMEL Math Reasoning*, and *Maths & College Reasoning*) and 4 LoRA fine-tuned models (**Qwen3**, **Phi-4**, **DeepSeek-V3**, and **LLaMA 3.2-3B**), using AUC as the evaluation metric. See Table I for the results.

- **Experiment 1 (Antepenultimate Layer, L1)** uses only the statistical descriptors extracted from the third-to-last layer, resulting in a 7-dimensional feature vector, and achieves strong performance across all datasets (e.g., up to 0.9800 for **Phi-4** and 0.9841 for **DeepSeek-V3**), indicating that relatively shallow internal representations already contain relevant membership signals, although performance remains below deeper configurations.
- **Experiment 2 (Penultimate Layer, L2)** relies on features from the second-to-last layer, also with a 7-dimensional representation, and consistently improves over L1 (e.g., **Qwen3** improves from 0.8306 to 0.9179 on *Medical-o1*), suggesting that intermediate representations provide a better balance between abstraction and memorization.
- **Experiment 3 (Last Layer, L3)** uses the final transformer layer, again with a 7-dimensional feature vector, and further improves performance in most cases (e.g., **DeepSeek-V3** reaches 0.9846 and 0.9843 on *Medical* and *CAMEL* datasets, respectively), confirming that deeper layers encode stronger membership-related signals.
- **Experiment 4 (OnlyGrad)** evaluates a single gradient-based sensitivity feature (1-dimensional), resulting in significantly lower performance across all datasets (e.g. 0.6706 for **Qwen3** and 0.5927 for **LLaMA 3.2-3B** on *Medical-o1*), indicating that gradients alone are insufficient to reliably capture membership information.
- **Experiment 5 (GradProbLoss)** combines gradient, loss, and confidence-based features into a compact 3-dimensional representation, yielding substantial improvements over the gradient-only setting (e.g. 0.9435 for **DeepSeek-V3** on *Medical-o1* and 0.9779 for **Qwen3** on *Maths & College Reasoning*), demonstrating the complementarity of output-level and sensitivity signals.
- **Experiment 6 (All3Layers)** integrates activation-based features of the antepenultimate, penultimate, and last layers together with gradient and output signals, resulting in a 15-dimensional feature vector and achieving the best

TABLE I
PERFORMANCE COMPARISON (AUC) OF DIFFERENT MINT-BASED VARIANTS ACROSS THREE DATASETS AND 4 LoRA FINE-TUNED MODELS.
LAYER-BASED CONFIGURATIONS ARE REPORTED AS ANTEPENULTIMATE (L1), PENULTIMATE (L2), AND LAST LAYER (L3).

DATASET	PROPOSED METHOD	Qwen3	Phi-4	DeepSeek-V3	LLaMA 3.2-3b
<i>Medical-o1-reasoning-SFT</i>	Loss [36] (Relevant literature 2022)	0.7508	0.9390	0.8840	0.8000
	MinK [37] (Relevant literature 2024)	0.7690	0.9440	0.8840	0.7770
	MinK++ [38] (Relevant literature 2025)	0.7780	0.9590	0.8625	0.7812
	Antepenultimate Layer (L1)	0.8306	0.9800	0.9697	0.8811
	Penultimate Layer (L2)	0.9179	0.9788	0.9840	0.9202
	Last Layer (L3)	0.9272	0.9811	0.9846	0.9210
	OnlyGrad	0.6706	0.7132	0.6683	0.5927
	GradProbLoss	0.7920	0.6323	0.9435	0.8394
	All3Layers	<u>0.9302</u>	<u>0.9819</u>	<u>0.9851</u>	<u>0.9240</u>
<i>CAMEL Math Reasoning</i>	Loss [36] (Relevant literature 2022)	0.8838	0.9160	0.9051	0.8090
	MinK [37] (Relevant literature 2024)	0.8720	0.9160	0.9050	0.8090
	MinK++ [38] (Relevant literature 2025)	0.7823	0.8730	0.8465	0.7985
	Antepenultimate Layer (L1)	0.9294	0.9533	0.9841	0.8948
	Penultimate Layer (L2)	0.9429	0.9496	0.9865	0.9591
	Last Layer (L3)	0.9485	0.9480	0.9843	0.9630
	OnlyGrad	0.7073	0.8047	0.7895	0.7536
	GradProbLoss	0.9201	0.9400	0.9254	0.8639
	All3Layers	<u>0.9514</u>	<u>0.9578</u>	<u>0.9901</u>	<u>0.9645</u>
<i>Maths & College Reasoning</i>	Loss [36] (Relevant literature 2022)	0.9380	0.9210	0.9610	0.7850
	MinK [37] (Relevant literature 2025)	0.9440	0.9021	0.9840	0.8490
	MinK++ [38] (Relevant literature 20245)	0.9590	0.8464	0.8945	0.7954
	Antepenultimate Layer (L1)	0.9848	0.9521	0.9810	0.9747
	Penultimate Layer (L2)	0.9855	0.9801	0.9845	0.9743
	Last Layer (L3)	0.9810	0.9821	0.9866	0.9726
	OnlyGrad	0.6548	0.8424	0.7462	0.8420
	GradProbLoss	0.9779	0.9398	0.9785	0.9627
	All3Layers	<u>0.9857</u>	<u>0.9845</u>	<u>0.9901</u>	<u>0.9792</u>

overall performance across all datasets and models (e.g. 0.9851 for **DeepSeek-V3** on *Medical-o1* and 0.9901 on both *CAMEL* and *Math* datasets), confirming that combining multi-layer internal representations with gradient and output features provides the most comprehensive characterization of model behavior for MINT.

A. Ablation Study of LoRA Target Modules

To quantify the impact of different injection points of LoRA within the transformer architecture, we conduct a structured ablation study with five configurations using the **Phi-4** model in the *Medical-o1-SF* dataset. This setup enables us to analyze how attention- and MLP-based adaptations contribute to detection performance.

We begin by evaluating the adaptation to partial attention, where only `q_proj` and `k_proj` are enabled. We then extend this to full attention adaptation by activating `q_proj`, `k_proj`, `v_proj`, and `o_proj`. Next, we isolate the contribution of the MLP layers by enabling only `up_proj` and `down_proj`. This is followed by the full MLP configuration, where all three layers of MLP (`gate_proj`, `up_proj` and `down_proj`) are activated. Finally, we evaluate the complete LoRA setup, which combines full attention with the full MLP (i.e., all three MLP layers), representing the maximum adaptation capacity. The results, summarized in Table II, reveal a clear trend. Partial attention adaptation yields limited performance (AUC = 0.6178), while full attention provides a moderate improvement (AUC = 0.6978). In contrast, MLP-based adaptations contribute significantly more, with partial

TABLE II
ABLATION STUDY OF LoRA TARGET MODULES USING THE **Phi-4** MODEL EVALUATED ON THE *Medical-o1-SF* DATASET. ALL EXPERIMENTS ARE CONDUCTED USING THE CONFIGURATION THAT AGGREGATES THE REPRESENTATIONS FROM THE LAST THREE TRANSFORMER LAYERS TOGETHER WITH THE MAXGRADPROB SCORING CRITERION. ACTIVE LoRA INJECTION POINTS ARE SHOWN IN BOLD, WHILE INACTIVE MODULES APPEAR IN GRAY. EACH CONFIGURATION ISOLATES A SPECIFIC SUBSET OF ATTENTION OR MLP PATHWAYS TO QUANTIFY THEIR CONTRIBUTION TO DETECTION PERFORMANCE.

Target Modules	AUC
<code>q_proj</code> , <code>k_proj</code> , <code>v_proj</code> , <code>o_proj</code> , <code>gate_proj</code> , <code>up_proj</code> , <code>down_proj</code>	0.6178
<code>q_proj</code> , <code>k_proj</code> , <code>v_proj</code> , <code>o_proj</code> , <code>gate_proj</code> , <code>up_proj</code> , <code>down_proj</code>	0.6978
<code>q_proj</code> , <code>k_proj</code> , <code>v_proj</code> , <code>o_proj</code> , <code>gate_proj</code> , <code>up_proj</code> , <code>down_proj</code>	0.7945
<code>q_proj</code> , <code>k_proj</code> , <code>v_proj</code> , <code>o_proj</code> , <code>gate_proj</code> , <code>up_proj</code> , <code>down_proj</code>	0.9019
<code>q_proj</code> , <code>k_proj</code> , <code>v_proj</code> , <code>o_proj</code> , <code>gate_proj</code> , <code>up_proj</code> , <code>down_proj</code>	0.9819

MLP reaching an AUC of 0.7945 and full MLP (i.e., all three layers enabled) achieving 0.9019. The best performance is obtained when combining both both attention and the full MLP stack, reaching an AUC of 0.9819.

In general, these findings indicate that MLP pathways—particularly when all three layers are enabled—play a more dominant role than attention modules in this setting, while their combination produces a strong synergistic effect.

VI. CONCLUSIONS

We presented inLoRA-MINT, a feature-based framework for membership inference on LoRA fine-tuned LLMs. For the sake of concreteness and current relevance, our discussion and experiments are centered on LoRA-adjusted LLMs, but note that most of the presented methodology is easily applicable for auditing training data given any other technique for adapting LLMs or, more generally, any AI model. Our proposed approach inLoRA-MINT combines internal activation statistics, gradient-based sensitivity measures, and output-level signals to construct expressive representations of model behavior and then a binary membership classification.

Experiments on three benchmark datasets (*Medical-o1-reasoning-SFT*, *CAMEL Math Reasoning* and *Maths & College Reasoning*) and 4 state-of-the-art models (**Qwen3**, **Phi-4**, **DeepSeek-V3**, and **LLaMA 3.2-3B**) reveal consistent trends. Deeper transformer layers provide progressively stronger membership signals, with later layers outperforming earlier ones. Gradient-only features show limited effectiveness, while combining gradient, loss, and confidence signals leads to significant improvements, highlighting their complementarity.

The best performance is achieved by integrating activation statistics from the last three layers with gradient and output-based features, confirming that membership information is distributed across multiple levels of the model and that leveraging heterogeneous signals is essential for robust inference.

Finally, we observe differences between model families. Larger models such as **DeepSeek-V3** achieve higher and more stable AUC scores, suggesting stronger memorization effects, while smaller models such as **LLaMA 3.2-3B** show lower and more variable performance. **Qwen3** and **Phi-4** lie in an intermediate regime, indicating that model scale and architecture play a key role in privacy leakage.

In our future work, we will study and extend our training data auditing tools to evaluate other relevant risks [13] in multimodal LLMs (including VLMs [20]). Analyzing biases [6], [39] and synthetic manipulations [40] while maintaining privacy [5] are also key topics for risk evaluation and mitigation in AI systems. Finally, we will apply these risk assessment and management principles to concrete domains such as document intelligence [41], [42], e-learning [7], [8], gaming [43], and e-health [2].

REFERENCES

- [1] S. Romero-Tapiador *et al.*, “Are Vision-Language Models ready for dietary assessment? Exploring the next frontier in AI-powered food image recognition,” in *IEEE/CVF CVPR Workshops*, 2025.
- [2] S. Romero-Tapiador, R. Tolosana, A. Morales *et al.*, “Personalized weight loss management through wearable devices and artificial intelligence,” *Computers in Biology and Medicine*, vol. 209, p. 111676, 2026.
- [3] A. Peña *et al.*, “Continuous document layout analysis: Human-in-the-loop AI-based data curation, database, and evaluation in the domain of public affairs,” *Information Fusion*, vol. 108, p. 102398, 2024.
- [4] M. Lopez-Duran *et al.*, “Benchmarking graph neural networks for document layout analysis in public affairs,” in *ICDAR Workshops*, 2025.
- [5] G. Mancera, A. Morales, J. Fierrez *et al.*, “PBa-LLM: Privacy-and bias-aware NLP using named-entity recognition (NER),” in *ICDAR*, 2025.
- [6] A. Peña, J. Fierrez, *et al.*, “Addressing bias in LLMs: Strategies and application to fair AI-based recruitment,” in *AAAI AIES*, 2025.
- [7] R. Daza, A. Becerra, R. Cobos, J. Fierrez, and A. Morales, “A multi-modal dataset for understanding the impact of mobile phones on remote online virtual education,” *Scientific Data*, vol. 12, p. 1332, 2025.
- [8] J. Irigoyen *et al.*, “AIriskEval-edu: New dataset for risk assessment in AI-mediated K-12 educational explanations,” in *IEEE ICCST*, 2026.
- [9] K. Du, Y. Zhao, R. Mao, F. Xing, and E. Cambria, “Natural language processing in finance: A survey,” *Information Fusion*, vol. 115, 2025.
- [10] A. Morales *et al.*, “SensitiveNets: Learning agnostic representations with application to face recognition,” *IEEE TPAMI*, vol. 43, no. 6, June 2021.
- [11] I. Serna *et al.*, “Sensitive loss: Improving accuracy and fairness of face representations with discrimination-aware DL,” *Artif. Intell.*, 2022.
- [12] A. Peña *et al.*, “Human-centric multimodal machine learning: Recent advances and testbed on AI-based recruitment,” *SN Comput. Sci.*, 2023.
- [13] J. Irigoyen *et al.*, “Overview of risk assessment and management for intelligent systems under the AI Act and beyond,” in *IEEE ICCST*, 2026.
- [14] I. Serna, A. Morales, and J. Fierrez, “Unraveling machine behavior by multi-level bias analysis and detection: Methodology and application to computer vision,” *Pattern Recognition*, 2027.
- [15] R. Busch *et al.*, *Privacy & Security in Biometric Tech.* Springer, 2027.
- [16] R. Shokri *et al.*, “Membership inference attacks against machine learning models,” in *Proc. of IEEE Symposium on Security and Privacy*, 2017.
- [17] L. Watson, C. Guo, G. Cormode *et al.*, “On the importance of difficulty calibration in membership inference attacks,” in *ICLR*, 2022.
- [18] D. DeAlcala, A. Morales *et al.*, “Is my data in your AI? membership inference test (MINT) applied to face biometrics,” *IEEE Access*, 2025.
- [19] G. Mancera, D. DeAlcala, A. Morales *et al.*, “Auditing training data in domain-adapted LLMs: LoRA-MINT,” in *IEEE COMPSAC*, 2026.
- [20] D. DeAlcala *et al.*, “Is my vision-language data in your AI? Membership Inference Test (MINT) Demo 2,” in *IEEE COMPSAC*, 2026.
- [21] G. Mancera *et al.*, “Is my text in your AI model? gradient-based membership inference test applied to LLMs,” in *IEEE CAI*, 2026.
- [22] D. DeAlcala, G. Mancera *et al.*, “A comprehensive analysis of factors impacting membership inference,” in *CVPR Workshops*, 2024.
- [23] D. DeAlcala *et al.*, “Active membership inference test (aMINT): Enhancing model auditability with multi-task learning,” in *ICCV*, 2025.
- [24] D. DeAlcala, A. Morales *et al.*, “gMINT: Gradient-based membership inference test applied to image models,” in *CVPR Workshops*, 2025.
- [25] G. Mancera *et al.*, “Membership inference test: Auditing training data in object classification models,” in *AAAI Work. on Deployable AI*, 2025.
- [26] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang *et al.*, “LoRA: Low-rank adaptation of large language models,” *ICLR*, 2022.
- [27] M. Gomez-Barrero, J. Galbally *et al.*, “Privacy-preserving comparison of variable-length data with application to biometric template protection,” *IEEE Access*, vol. 5, pp. 8606–8619, June 2017.
- [28] E. Macias *et al.*, “Privacy-aware synthetic tabular data generation for healthcare: Application to sepsis detection,” *Bioengineering*, 2026.
- [29] J. Chen, Z. Cai, K. Ji, X. Wang, W. Liu *et al.*, “HuatuoGPT-o1, towards medical complex reasoning with LLMs,” *arXiv:2412.18925*, 2024.
- [30] G. Li, H. Hammoud *et al.*, “CAMEL: Communicative agents for mind exploration of large language model society,” *NeurIPS*, vol. 36, 2023.
- [31] Ajibawa-2023, “Maths-College: A mathematics reasoning dataset,” Hugging Face Datasets, 2023, accessed: 2026-07-03. [Online]. Available: <https://huggingface.co/datasets/ajibawa-2023/Maths-College>
- [32] J. Bai, S. Bai *et al.*, “Qwen technical report,” *arXiv:2309.16609*, 2023.
- [33] M. Abdin *et al.*, “Phi-4 technical report,” *arXiv:2412.08905*, 2024.
- [34] A. Liu *et al.*, “Deepseek-v3 technical report,” *arXiv:2412.19437*, 2024.
- [35] A. Grattafiori *et al.*, “Llama 3 herd of models,” *arXiv:2407.21783*, 2024.
- [36] Y. Liu, Z. Zhao, M. Backes, and Y. Zhang, “Membership inference attacks by exploiting loss trajectory,” in *ACM CCS*, 2022.
- [37] W. Shi, A. Ajith, M. Xia, Y. Huang, D. Liu, T. Blevins *et al.*, “Detecting pretraining data from large language models,” in *ICLR*, 2024.
- [38] J. Zhang, J. Sun *et al.*, “Min-k%++: Improved baseline for pre-training data detection from large language models,” in *ICLR*, vol. 2025, 2025.
- [39] J. Tello, M. de la Cruz, T. Ribeiro *et al.*, “Symbolic AI (LFIT) for XAI to handle biases,” in *ECAI Workshops*, ser. CEUR-WS, vol. 3523, 2023.
- [40] P. Korshunov *et al.*, “DeepID challenge of detecting synthetic manipulations in ID documents,” in *IEEE ICCV Workshops*, 2025.
- [41] J. Muñoz-Haro, R. Tolosana *et al.*, “Privacy-aware detection of fake identity documents: methodology, benchmark, and improved algorithms (FakeIDet2),” *Information Fusion*, vol. 128, p. 103969, 2026.
- [42] M. Lopez-Duran *et al.*, “Comparative study of domain-adapted VLMs for general document visual question answering,” in *ICDAR*, 2026.
- [43] R. Daza *et al.*, “Evaluating Social Engineering Risks in AI-based Interaction using Biometrics and a Gaming Setup,” in *ICCST*, 2026.