Dataset Inference in Fine-Tuned Large Language Models: A Comparative Study

Gabriel Vaz de Oliveira [1], Arthur Santos Viana de Oliveira [1], Victor Aguiar Evangelista de Farias [1], Javam de Castro Machado[1], Carlos Caminha[1]

[1]Kunumi Lab
Universidade Federal do Ceará
Fortaleza – Ceará – Brasil

{gabriel.vaz, arthurviana}@alu.ufc.br,
victor.farias@ufc.br, javam.machado@lsbd.ufc.br, caminha@ufc.br

Abstract. *Dataset Inference provides a robust framework for auditing data ownership by aggregating statistical signals that traditional Membership Inference Attacks (MIAs) fail to capture. While proven for LLM pre-training, its efficacy during fine-tuning is largely unexplored. We evaluate Dataset Inference on Gemma, Llama, and Qwen models using Full Fine-Tuning (FFT), LoRA, and QLoRA. Our findings reveal a stark architectural divergence: Parameter-Efficient Fine-Tuning (PEFT) mitigates data leakage in Llama ($AUC \approx 0.50$, $p > 0.05$), while Gemma and Qwen remain highly vulnerable across all adaptation methods ($p < 0.05$). Additionally, higher learning rates accelerate data absorption, and QLoRA provides only marginal regularization compared to standard LoRA.*

1. Introduction

The widespread adoption of Large Language Models (LLMs), which are predominantly pre-trained on massive, heterogeneous datasets gathered largely through unsupervised web scraping, has raised serious privacy and copyright concerns [Das et al. 2025]. While pre-training initially equips LLMs with extensive knowledge, achieving fine-grained control over their behavior remains challenging. In most practical applications, organizations start from these base models and subsequently perform fine-tuning to more closely align the model's behavior with particular goals [Raschka 2024].

The literature has shown that fine-tuning tends to intensify the leakage of training data [Das et al. 2025]. To assess this risk, researchers use MIAs, which aim to determine whether a data point was part of a model's training set [Jagannatha et al. 2021].

The effectiveness of MIAs at the sample level is debated. High accuracy may only indicate a detection of data distribution shifts, not actual memorization. For example, Duan et al. (2024) showed that using non-member samples from other time periods lowers n-gram overlap with training data, so classifiers may pick up on temporal drift rather than true membership, causing false positives.[Duan et al. 2024].

As a robust methodological alternative to this limitation, Dataset Inference is proposed [Maini et al. 2024]. Unlike standard MIAs, which assess membership at the sample level, this approach aims to determine whether an entire *dataset* was used in model training, aligning more closely with contemporary copyright concerns. Its central intuition is

that multiple weak sample-level signals, in our case 16 per sample, can be aggregated into a stronger dataset-level indicator. While this method has been explored for pre-training, it remains underexplored for fine-tuning LLMs.

Therefore, this work seeks to answer: *(1) Which fine-tuning method (Full, LoRA [Hu et al. 2022], or QLoRA [Dettmers et al. 2023]) is more susceptible to dataset-level MIAs? and (2) how do hyperparameters, such as epochs, rank, scaling factor α, learning rate, affect Dataset Inference?* The main contributions of this work are:

1. We build a custom news dataset of approximately 9,000 articles published between 2024 and 2025, using standardization, deduplication, length filtering, and a 75%/25% split to simulate data scarcity.

2. We fine-tune models under different configurations and evaluate their susceptibility to dataset-level inference on member and non-member data.

3. We assess attack performance using Perplexity, Min-K%, and zlib, combining them in a membership model and reporting AUC and p-values across configurations.

2. Our Approach

To empirically evaluate the privacy risks of different adaptation methods, we designed a dataset-level membership detection pipeline. After collecting and standardizing articles from diverse news sources, we build isolated, balanced partitions for calibration and inference, clearly separating "Suspect" data (used in fine-tuning) from "Private" data (never seen by the model). This setup allows us to quantify how specific hyperparameters and architectural choices influence unwanted memorization of the training data.

2.1. Dataset Construction and Partitioning

To ensure data quality and avoid pre-training contamination, our dataset construction followed a compact pipeline (Figure 1). We collected news articles published between July 2024 and October 2025 [R3troR0b 2024]—dates strictly after the target LLMs' pre-training cutoffs. After deduplication and minimum-length filtering, the corpus yielded approximately 9,000 unique articles.

We first split this corpus into a *Suspect* set (75%, used for LLM fine-tuning) and a *Private* holdout set (25%, never exposed to the LLM). For dataset inference, we then sampled these base sets to build two disjoint, source-stratified partitions, both constrained to at most 1,000 samples and with a strict 50/50 mixture of *Suspect* and *Private*, following the sample-size recommendations of Maini et al. [Maini et al. 2024] for stable p-values.

Partition A (Calibration) is used to train the linear Membership Model with access to labels, while Partition B (Inference) is evaluated with labels hidden, and its score distributions feed the final statistical test that decides whether the dataset was used for fine-tuning.

2.2. Fine-tuning

To evaluate the impact of training dynamics on modelo memorization, we systematically varied key hyperparameters one at a time to measure how each affects model memorization and data leakage during fine-tuning, allowing us to isolate how specific constraints

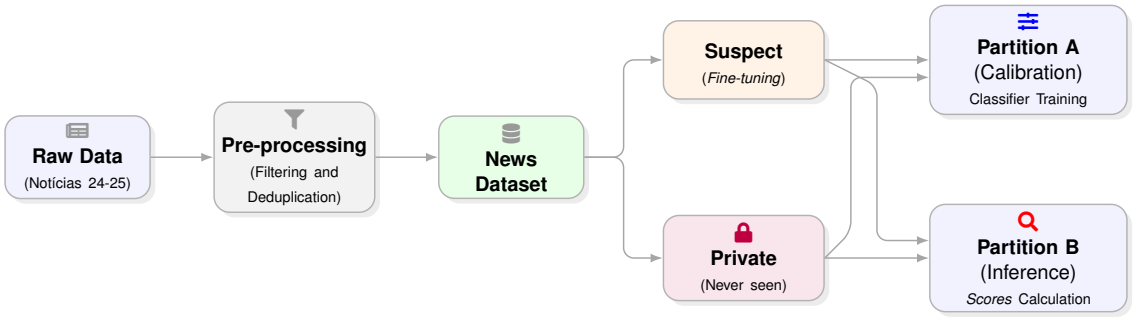

Figure 1. Construction and partitioning flow of the dataset

influence the "leakage" of training data. With this we aim to identify when models shift from task alignment to unintended memorization.

The chosen models for our experiments were the *qwen-2.5-0.5b* with 0.5 billion parameters, and the *gemma-3-1b-pt* and *llama-3.2-1B*, both with 1 billion parameters. The chosen parameters and their values were: Rank (r) {8, 16, 32, 64}, Alpha (α) {16, 32, 64, 128}, learning rate $\{1 \times 10^{-3}, 2 \times 10^{-4}, 5 \times 10^{-5}, 1 \times 10^{-5}\}$ and training epochs {1, 2, 3, 4}. The baseline configuration established was $r = 16$, $\alpha = 32$, learning rate of 2×10^{-4}, and a training duration of 3 epochs for LoRA, and QLoRA. In the case of FFT, only learning rate and epochs apply.

For the training and evaluation steps of the models, the Suspect set was used, with 75% allocated to training and 25% to evaluation. For each configuration, the model was trained with evaluation steps performed at 10% intervals. Each evaluation yielded its loss score, which was used with an **early stopping** callback to prevent overfitting.

2.3. Dataset Inference in Fine-tuned Models

In this section, we detail the methodology used to determine whether a specific dataset was used in the fine-tuning process of a language model. Unlike traditional MIAs, which seek to identify the presence of individual samples, dataset inference aggregates weak signals from multiple samples to perform a statistical test on the distribution as a whole. We adapt the approach proposed by Maini et al. to the scenario in which the hypothesis is that fine-tuning leaves detectable traces of memorization even when the base model already has prior knowledge of the data.

1. Extract the MIA metric vector per sample: All texts in partitions A and B are then processed by the full suite of Membership Inference Attacks (MIAs). This step converts each original news article into a fixed-size feature vector, where each dimension corresponds to one of 16 MIA-derived numerical signals (1 signal for perplexity, 7 for Min-K% and 7 for Max-K% where $K \in \{5, 10, 20, 30, 40, 50, 60\}$, and 1 for zlib ratio). After this stage, both A and B exist only in this vector representation, which allows us to reason entirely in terms of attack scores instead of raw text. **2. Train the Membership Model:** Using only the labeled feature vectors from Partition A (Calibration), we train a supervised linear classifier, referred to as the *Membership Model*. Its goal is to learn a weight vector w that linearly combines the 16 MIA scores into a single membership score

$$f(x) = w^\top \phi(x),$$

where $\phi(x)$ is the feature vector of sample x. The linear formulation keeps the model interpretable and mitigates overfitting, while still capturing the aggregate signal obtained by combining multiple weak MIAs. **3. Scoring and hypothesis testing on Partition B:** With w fixed, we apply $f(x)$ to all samples in Partition B (Inference), whose labels remain hidden. We then compare the score distributions of Suspect vs. Private records in B using a one-sided statistical test: the null hypothesis H_0 states that the model did *not* use the Suspect dataset during fine-tuning (no membership signal), while the alternative hypothesis H_1 assumes the opposite. A small p-value (e.g., $p < 0.05$) leads us to reject H_0 and is interpreted as statistical evidence that the dataset was consumed during fine-tuning.

3. Results

This section presents the empirical behavior of Dataset Inference across architectures and adaptation strategies, with emphasis on attack performance (AUC) and statistical significance (p-value) under different hyperparameter configurations.

3.1. Comparison between Full Fine-Tuning and PEFTs

The analysis of the metrics (Figure 2) suggests that Full Fine-Tuning consistently fails the hypothesis test ($p \approx 0.0$) across Gemma, Llama, and Qwen, indicating a significant tendency toward memorization and vulnerability to inference attacks. In contrast, PEFT methods (LoRA and QLoRA) appear to mitigate memorization in Llama, with AUC scores approaching the random detection line (≈ 0.50) and p-values generally remaining above the 0.05 threshold, which may reflect improved privacy protection. For Gemma and, in particular, Qwen, however, the p-values under PEFT often remain below 0.05, accompanied by higher AUC scores (frequently above 0.90), suggesting that these architectures may retain fragments of the training data independently of the adaptation method or hyperparameter settings.

3.1.1. Internal Analysis of PEFT Hyperparameters (LoRA vs. QLoRA)

When analyzing the statistical significance of inference attacks (Figure 3) using a rejection threshold of $p = 0.05$, a clear polarization emerges between architectures under PEFT methods. Llama demonstrates strong resistance to memorization, with p-values consistently above the threshold across most Rank (r) and Alpha (α) configurations, in some cases approaching 1.0, indicating failure of the Membership Model to distinguish training data.

In contrast, Gemma and Qwen exhibit structural vulnerability: all configurations yield $p \approx 0.0$, regardless of method or hyperparameters, indicating statistically significant leakage. Qwen is particularly vulnerable, combining near-zero p-values with consistently higher AUC scores than Gemma.

The resilience observed in Llama, and the vulnerability of Gemma and Qwen are confirmed by the AUC analysis (Figure 3), compared to a random baseline (AUC $=$ 0.50). Across variations in Rank and Alpha, Llama remained consistently intertwined with this baseline, demonstrating that low-rank adaptation forced the attack to operate on the margin of mere random guessing.

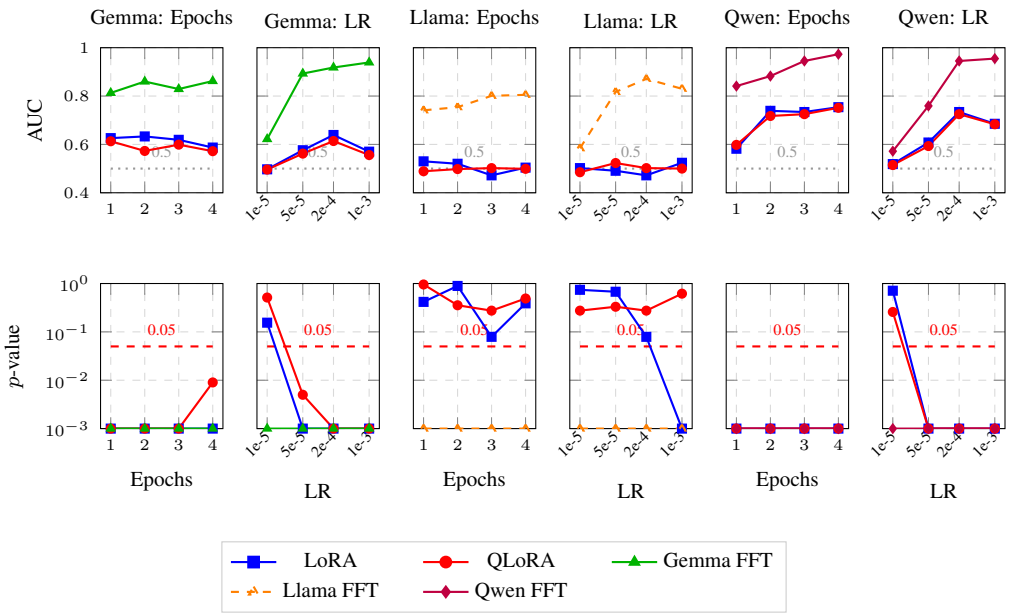

Figure 2. AUC and p-value across models and hyperparameters. Learning rate and p-values are shown in $\log_{10}$ scale. The dotted gray line at AUC = 0.5 denotes random guessing, and the red dashed line marks the significance threshold $p = 0.05$. p-values below 10^{-3} (including zeros) are clipped to 10^{-3} for visualization.

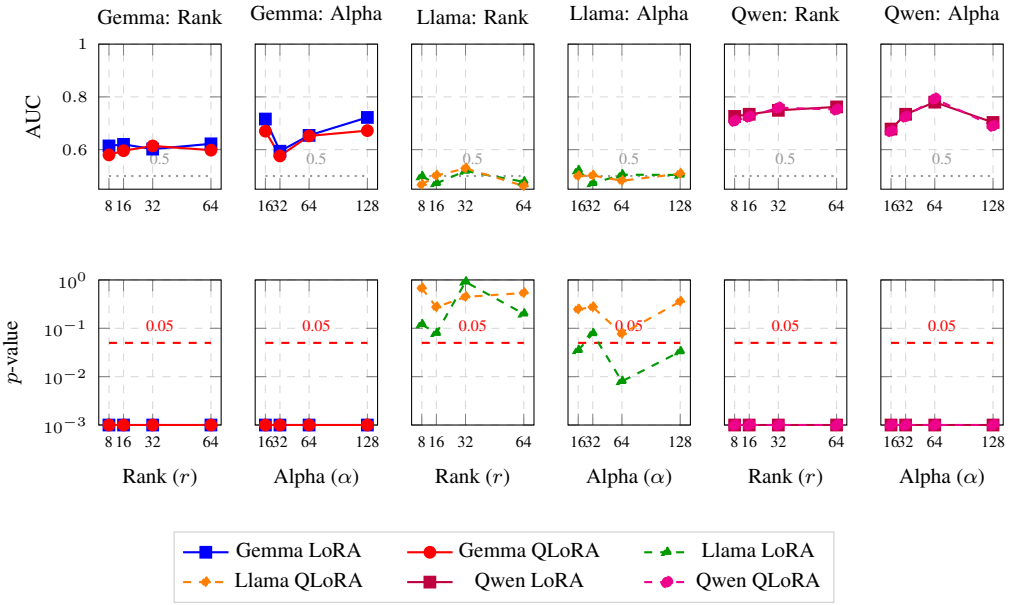

Figure 3. AUC and p-value comparison across PEFT parameters (Rank and Alpha), including Qwen. p-values are shown in $\log_{10}$ scale, with zeros clipped to 10^{-3}, and the dotted gray line at AUC = 0.5 denotes random guessing.

In turn, Gemma and Qwen deviated from this margin in all parameterizations. Gemma sustained AUC levels in the range of 0.60, reaching peaks above 0.70 at the extremes of the multiplier *Alpha* ($\alpha = 16$ and 128), with QLoRA providing a slight attenuation. Qwen showed an even more pronounced vulnerability, with AUC reaching peaks near 0.80 across configurations. Furthermore, unlike Gemma, the use of quantization (QLoRA) did not provide attenuation in Qwen, maintaining a consistently high AUC.

4. Conclusion

This study evaluates the data resilience of LLMs under fine-tuning using the Dataset Inference paradigm. Results reveal a clear architectural divergence: while FFT leads to near-total memorization across models, PEFT (LoRA and QLoRA) mitigates leakage in Llama (AUC ≈ 0.50, $p > 0.05$), but not in Gemma and Qwen, which remain vulnerable ($p < 0.05$), with Qwen exhibiting severe leakage (AUC ≈ 0.80). Learning rate emerges as a key factor: lower values (1×10^{-5}) reduce memorization, while higher rates (2×10^{-4}) amplify leakage. QLoRA provides slight regularization in Gemma but none in Qwen. These results indicate that data resilience depends more on base model architecture than on the fine-tuning strategy.

Acknowledgements

The authors thank the Kunumi Institute, and Carlos Caminha thanks FUNCAP (Fundação Cearense de Apoio ao Desenvolvimento Científico e Tecnológico) for their support of this research.

References

Das, B. C., Amini, M. H., and Wu, Y. (2025). Security and privacy challenges of large language models: A survey. *ACM Computing Surveys*, 57(6):1–39.

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. (2023). Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

Duan, M., Suri, A., Mireshghallah, N., Min, S., Shi, W., Zettlemoyer, L., Tsvetkov, Y., Choi, Y., Evans, D., and Hajishirzi, H. (2024). Do membership inference attacks work on large language models? *arXiv preprint arXiv:2402.07841*.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. (2022). Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3.

Jagannatha, A., Rawat, B. P. S., and Yu, H. (2021). Membership inference attack susceptibility of clinical language models. *arXiv preprint arXiv:2104.08305*.

Maini, P., Jia, H., Papernot, N., and Dziedzic, A. (2024). Llm dataset inference: Did you train on my dataset? *Advances in Neural Information Processing Systems*, 37:124069–124092.

R3troR0b (2024). News-dataset. `https://huggingface.co/datasets/R3troR0b/news-dataset`. [Dataset]. Accessed: 2026-03-02.

Raschka, S. (2024). *Build a large language model (from scratch)*. Simon and Schuster.