Leaner Training, Lower Leakage: Revisiting Memorization in LLM Fine-Tuning with LoRA

Anonymous ACL submission

Abstract

Memorization in large language models (LLMs) makes them vulnerable to data extraction attacks. While pre-training memorization has been extensively studied, fewer works have explored its impact in fine-tuning, particularly for LoRA fine-tuning, a widely adopted parameter-efficient method.

In this work, we re-examine memorization in fine-tuning and uncover a surprising divergence from prior findings across different fine-tuning strategies. Factors such as model scale and data duplication, which strongly influence memorization in pre-training and full fine-tuning, do not follow the same trend in LoRA fine-tuning. Using a more relaxed similarity-based memorization metric, we demonstrate that LoRA significantly reduces memorization risks compared to full fine-tuning, while still maintaining strong task performance.performance.

1 Introduction

Memorization in large language models (LLMs) has raised growing concerns, as studies have shown that these models can retain and expose training data (Carlini et al., 2023, 2019; Leybzon and Kervadec, 2024; Mireshghallah et al., 2022a). This susceptibility to memorization has led to the emergence of data extraction attacks, where adversaries exploit model generations to recover sensitive sequences from the training corpus (Carlini et al., 2021; Kassem et al., 2024; Zhang et al., 2023).

With the rapid proliferation of open-source LLMs, fine-tuning has become a standard approach for researchers and practitioners to adapt these models for task-specific applications. Fine-tuning is often performed using private or proprietary data, such as personal medical records, internal company documents, or domain-specific knowledge bases. This raises a critical concern: while fine-tuning datasets are typically much smaller than pre-training datasets, it remains unclear whether LLMs fine-tuned on limited data are still vulnerable to extraction attacks.

While memorization in pre-training has been extensively studied, far fewer works have systematically examined its effects in fine-tuning. As a pioneering study, Mireshghallah et al. (2022b) explored how different fine-tuning strategies—full-model, head-only, and bottleneck adapter-based fine-tuning (Houlsby et al., 2019)—impact membership inference attack (MIA) recall (Mireshghallah et al., 2022a) and the exposure (Carlini et al., 2019) of fine-tuning data. Their findings indicate that head-only fine-tuning leads to the highest memorization, whereas full-model and bottleneck adapter-based fine-tuning present lower but comparable risks. However, their study is constrained by the choice of evaluation metrics, as MIA recall tends to under-detect memorized data (Schwarzschild et al., 2024), potentially leading to an underestimation of actual memorization levels. In contrast, Zeng et al. (2024) analyzed memorization at the task level using a more memorization-sensitive metric based on plagiarism detection (Lee et al., 2023). Their results show that summarization and dialogue tasks induce significantly higher memorization compared to classification and translation, highlighting how task characteristics influence memorization severity.

In this work, we extend previous research by re-examining memorization in full and head-only fine-tuning while introducing an analysis of LoRA (Low-Rank Adaptation) (Hu et al., 2022) fine-tuning, a widely adopted parameter-efficient fine-tuning method. While LoRA is well known for its computational efficiency, its impact on memorization and data extraction risks remains largely unexplored. As an initial step, we use the same plagiarism-based memorization metric as in Zeng et al. (2024), uncovering a surprising divergence from previous findings regarding the susceptibility

of different fine-tuning methods to data extraction. Furthermore, we conduct a comprehensive evaluation using similarity-based metrics to compare LoRA, full, and head-only fine-tuning across model scales, data duplication levels, and hyperparameter configurations.

As the first empirical analysis of memorization in LoRA fine-tuning, our study extends prior work on full and head-only fine-tuning. Our results demonstrate that LoRA significantly reduces memorization risks compared to full fine-tuning, achieving near-zero plagiarism-based memorization while maintaining strong task performance and ensuring similarity scores remain below extraction thresholds. These findings highlight LoRA not only as a computationally efficient alternative to full fine-tuning but also as a potential privacy-preserving approach.

2 Preliminaries

2.1 Memorization in Pre-Training vs. Fine-Tuning

The concept of memorization in language models, as introduced by prior work (Carlini et al., 2023, 2021), defines a sequence s as *extractable* if there exists a prefix c that, when used as a prompt, leads the model to generate s with high probability. To quantify this memorization, they employ the framework of k-eidetic memorization, where a generated sequence s is considered memorized if it is extractable and appears in at most k instances within the training data.

To perform this data extraction, the adversary is assumed to have *black-box* access to the language model, allowing them to generate text without direct visibility into the model's internal parameters. The extraction process begins by initializing the model with a single-token prompt containing a special start-of-sentence token, followed by autoregressive sampling. Tokens are then iteratively selected using *top-*k sampling with $k = 40$, ensuring that only the most probable tokens are considered at each step.

In this work, we investigate memorization in fine-tuned language models using a similar data extraction setup. However, a generated sequence s is considered memorized in the fine-tuned model if it is extractable and appears in the fine-tuning data. Importantly, Variations in memorization and extractability evaluation methods can lead to significantly different interpretations of memorization risks. As we will demonstrate, even when adopting more relaxed memorization criteria, LoRA fine-tuning remains highly resistant to data extraction attacks.

2.2 Model Fine-Tuning Methods

We explore three common fine-tuning techniques for adapting pre-trained language models to specific tasks. Full fine-tuning simply updates all model parameters. Head-only fine-tuning is a lightweight approach that freezes all model layers except the final projection layer, also known as the language modeling head, which maps hidden states to token probabilities. LoRA strikes a balance by adding small trainable low-rank matrices to attention layers while keeping the rest of the model frozen, significantly reducing the number of trainable parameters. Key hyperparameters include rank, which determines the size of the low-rank matrices, the scaling factor, which adjusts the update strength, and the dropout rate (Srivastava et al., 2014), which helps prevent overfitting.

Unlike bottleneck adapter-based fine-tuning (Houlsby et al., 2019), which adds extra layers between transformer blocks (Mireshghallah et al., 2022b), LoRA directly modifies attention layers, achieving efficient fine-tuning with minimal computational overhead. This design enables faster adaptation, lower memory usage, and improved scalability, making it particularly well-suited for larger models and resource-constrained environments.

3 Related Work

While memorization in pre-training has been widely studied, fewer works have explored memorization in the context of fine-tuning. Two notable studies, Mireshghallah et al. (2022b) and Zeng et al. (2024), investigate this issue from different perspectives.

Mireshghallah et al. (2022b) examined how different fine-tuning strategies impact memorization, comparing full-model, head-only, and bottleneck adapter-based fine-tuning (Houlsby et al., 2019). Using membership inference attack (MIA) recall (Mireshghallah et al., 2022a) and the exposure metric (Carlini et al., 2019), they found a surprising result: fine-tuning only the model's head leads to the highest memorization, far more than full fine-tuning, even though it updates fewer parameters. This suggests that the common practice

of freezing all but the final layer may introduce unexpected risks of data leakage. However, their definition of memorization and choice of evaluation metrics differ from more recent advancements in the literature and our focus. Specifically, they equate memorization with the recall of the membership inference attack on the training set, which has a high false negative rate (Schwarzschild et al., 2024), potentially leading to an underestimation of memorization. To address this limitation, our study analyzes model memorization across different fine-tuning methods using more representative and informative evaluation metrics. Furthermore, we extend this investigation to state-of-the-art fine-tuning techniques, particularly LoRA, which has not been explored in prior memorization studies.

On the other hand, Zeng et al. (2024) investigated memorization at the task level, analyzing how different fine-tuning tasks affect memorization. They found that summarization and dialogue induce significantly higher memorization than classification, reading comprehension, and translation. Additionally, they confirmed that increasing model size amplifies memorization in fine-tuning—a phenomenon previously observed in pre-training (Carlini et al., 2023)—particularly in tasks prone to high memorization. However, the language models they examined, such as BART and T5, are no longer considered modern LLMs. While they did evaluate GPT-Neo-125M, its relatively small size limits its relevance to more recent large-scale models.

4 A First Look at Memorization of LLM Fine-Tuning

4.1 A Tentative Metric for Fine-Tuning Memorization

To what extent can we confirm that an LLM has memorized its training data? Using appropriate evaluation criteria is crucial, as different metrics can yield highly divergent results in data extraction attacks. Defining memorization solely as the model reproducing exact wordings from the training data, as in Carlini et al. (2023), may be too strict and could overlook instances of semantic memorization, where the model retains knowledge without verbatim reproduction. This limitation has been highlighted by Zeng et al. (2024).

To address this, Zeng et al. (2024) adopt an automated plagiarism detection approach from Lee et al. (2023), which evaluates memorization through a two-step process. First, the system retrieves the *top-n* most similar documents to a given query using the Elasticsearch (Gormley and Tong, 2015) ranking. Then, it applies the PAN 2014 competition-winning text alignment algorithm to detect and classify plagiarized text pairs within the identified candidate documents.

The detected plagiarism is classified into three types: *verbatim plagiarism*, which copies words or phrases exactly; *paraphrase plagiarism*, which modifies wording through synonym substitution, reordering, or back translation; and *idea plagiarism*, which rephrases key points in a condensed or expanded form.

While this plagiarism-based evaluation framework has been effective in detecting memorization in standard fine-tuning settings, the extent to which other fine-tuning methods exhibit memorization under this metric remains uncertain. To investigate this, we present preliminary results assessing plagiarism-based memorization across different fine-tuning approaches, including full fine-tuning, head-only tuning, and LoRA fine-tuning. Our evaluation follows the original setup and default hyperparameters of the plagiarism detection framework[1] (Lee et al., 2023).

4.2 Minimal Plagiarism-Based Memorization in Parameter-Efficient Fine-Tuning

Prior work has consistently shown that increasing model size and data duplication amplifies memorization during training (Carlini et al., 2023; Kandpal et al., 2022). This raises an important question: do models fine-tuned with different strategies, particularly head-only and LoRA fine-tuning, exhibit memorization trends similar to those observed in pre-training?

To investigate the impact of model size, we utilize open-source pre-trained GPT-2 models (Radford et al., 2019) of varying scales from Hugging Face: GPT-2 Small (124M parameters), GPT-2 Medium (355M parameters), GPT-2 Large (774M parameters), and GPT-2 XL (1.5B parameters). To examine the effect of data duplication, we introduce a hyperparameter ρ, which determines the duplication factor for each training sample in the fine-tuning dataset. A higher ρ increases the frequency of repeated samples while maintaining the same total dataset size. For example, in a dataset of $1,000$ samples, setting $\rho = 2$ results in 500 unique samples, each appearing twice. To mini-

[1] https://github.com/Brit7777/LM-plagiarism

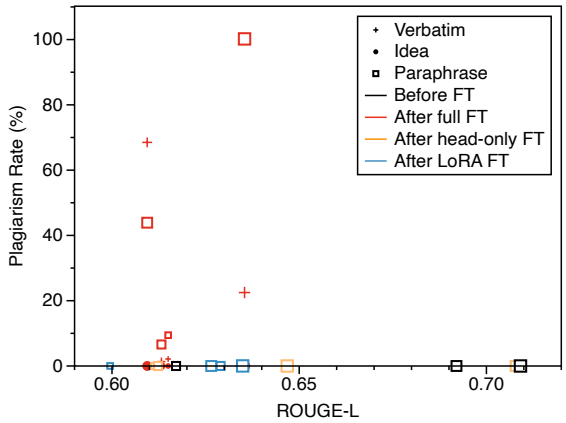
mize the likelihood that the fine-tuning data was seen during pre-training, we select a dataset released after GPT-2's training period. The `Arxiver` dataset (Alican Acar, 2024), which contains $63.4k$ arXiv papers published between January 2023 and October 2023, is a suitable choice for evaluating memorization in a scientific writing task.

We fine-tune our models for 10 epochs with a learning rate of 2×10^{-4}, using the AdamW optimizer and a cosine learning rate scheduler. Training is conducted with a batch size of 4 per device, and the sequence length is set to 512 tokens for GPT-2. Specifically, in head-only fine-tuning, we update only the last two transformer layers along with the final language modeling head, which are responsible for high-level representations and directly influence the model's token predictions. For LoRA fine-tuning, we apply a rank of $r = 16$, a scaling factor of $\alpha = 16$, and a dropout of 0.05 as a standard configuration. All GPT-2 models are fine-tuned and evaluated on a server equipped with a single NVIDIA RTX A6000 GPU (48GB VRAM), 21 vCPUs, and 83GB of RAM.

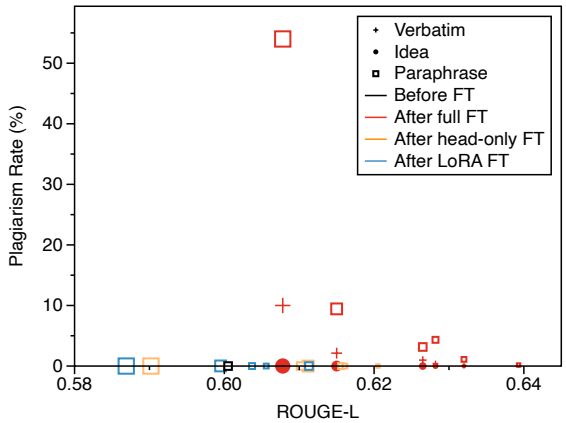
We adopt the data extraction attack mentioned in Section 2 to generate $1,000$ samples from each fine-tuned model and evaluate their plagiarism-based memorization. For text generation, rather than using the initial strategy of *top-k* sampling with $k = 40$ from Carlini et al. (2021), we refine token selection by setting *top-k* to 50 and incorporating *top-p* sampling with $p = 0.9$, ensuring that only the most probable tokens are considered while maintaining some flexibility. Additionally, we set the temperature to 0.8 to balance randomness and diversity in the generated text.

In Fig. 1, we illustrate how different fine-tuning strategies influence plagiarism-based memorization across varying levels of data redundancy and model sizes. To assess the task-specific performance of fine-tuned models, we use ROUGE-L scores as an approximate measure of model utility.

The tendency for larger models to memorize more persists in full fine-tuning, which exhibits a clear increase in verbatim and paraphrase plagiarism as model size grows. However, head-only and LoRA fine-tuning consistently maintain near-zero plagiarism rates across all model sizes. A similar trend emerges when examining the effect of data redundancy. Full fine-tuning exhibits the highest plagiarism rates as ρ increases, with paraphrase plagiarism surpassing 50% at $\rho = 20$. In contrast, head-only and LoRA fine-tuning remain largely unaffected, maintaining plagiarism rates close to zero even at the highest levels of redundancy.



(a) The dot size represents different model sizes: GPT-2 Small, GPT-2 Medium, GPT-2 Large, GPT-2 XL, with data redundancy fixed at $\rho = 10$ across all scenarios.



(b) The dot size represents varying levels of data redundancy, denoted as ρ: 1, 2, 3, 5, 10, and 20, with GPT-2 Small used across all scenarios.

Figure 1: Verbatim, idea, and paraphrase memorization across various fine-tuning methods, model scales, and data redundancy levels.

While full fine-tuning yields the highest ROUGE-L scores across all fine-tuning strategies, it also demonstrates the highest plagiarism rates, particularly in larger models and high-redundancy settings. Head-only and LoRA fine-tuning, on the other hand, achieve competitive ROUGE-L scores while significantly reducing plagiarism. LoRA fine-tuning, in particular, balances efficient training and reduced memorization, incurring minimal performance loss compared to full fine-tuning.

Unlike prior work, which found that head-only fine-tuning leads to the highest memorization among different fine-tuning methods for the same level of perplexity, while full fine-tuning and small adapter-based fine-tuning (Houlsby et al.,

2019) exhibit low data leakage under MIA recall (Mireshghallah et al., 2022a) and exposure metrics (Carlini et al., 2019), our results reveal an entirely different trend under plagiarism-based memorization evaluation. This metric effectively captures memorization in full fine-tuning, where all model parameters are updated. However, head-only and LoRA fine-tuning consistently yield zero detected plagiarism, even as fine-tuning data redundancy and model scale increase. This suggests that these parameter-efficient fine-tuning strategies, which freeze most of the model, do not exhibit direct memorization under this evaluation criterion—let alone the stricter verbatim memorization criteria originally proposed in Carlini et al. (2021).

5 Expanding the Lens: A More Permissive View of Memorization in Fine-Tuning

To further investigate memorization in fine-tuning, we extend our evaluation beyond plagiarism-based detection. A more permissive metric allows us to examine how head-only and LoRA fine-tuning strategies mitigate memorization while retaining information in a less directly extractable form. In this section, we introduce a looser memorization criterion and analyze how different fine-tuning strategies behave under this broader perspective.

5.1 What A Looser Metric Reveals About Fine-Tuning

To this end, we adopt **sentence similarity** to provide a more nuanced perspective of how closely generated outputs resemble training samples at a semantic level. This metric allows us to assess whether head-only and LoRA fine-tuning retain training data in a way that is not directly extractable but still recognizable through semantic similarity.

For each sequence generated from the fine-tuned model during the data extraction attack, we encode both the generated samples and the fine-tuning corpus into query embeddings and document embeddings, respectively, using the All-MPNet-Base-V2 model[2], a widely used sentence embedding model that maps text into a 768-dimensional dense vector space. We then compute cosine similarity scores between the query embeddings and document embeddings in a single matrix operation, efficiently capturing semantic similarity. The highest similarity score for each generated sequence with respect to a training sample serves as a quantitative measure of how closely the generated sentence aligns with a training instance. A cosine similarity threshold of 0.8 is typically used to indicate strong semantic similarity. If the top sentence similarity score of a generated sequence exceeds this threshold, it is considered to closely align with a training sample from the fine-tuning dataset.

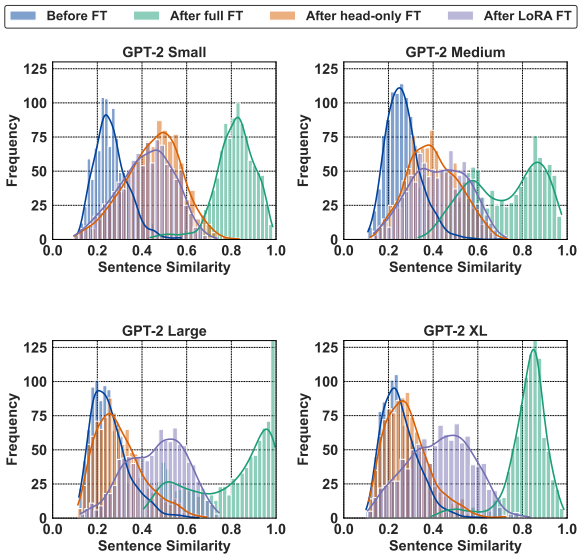


Figure 2: The histogram distribution of the top sentence similarity scores for $1,000$ generated samples from data extraction attacks, comparing different fine-tuning methods across various model sizes.

As shown in Fig. 2, full fine-tuning consistently shifts similarity scores rightward across all model sizes, with a pronounced concentration of samples exceeding the 0.8 similarity threshold. This indicates a substantial increase in direct memorization compared to the pre-trained model.

Head-only fine-tuning produces outputs that remain more similar to those of the pre-trained model, especially as model size increases. This is likely because it updates only the final layers while keeping the majority of parameters frozen. As model size grows, the proportion of trainable parameters relative to the entire model decreases, further constraining the extent to which fine-tuning can reshape the model's internal representations.

LoRA fine-tuning exhibits a distinct trend compared to both full and head-only fine-tuning. While its similarity distribution shifts slightly from the pre-training baseline, it remains considerably lower than full fine-tuning, with similarity peaks concen-

[2]https://huggingface.co/sentence-transformers/all-mpnet-base-v2

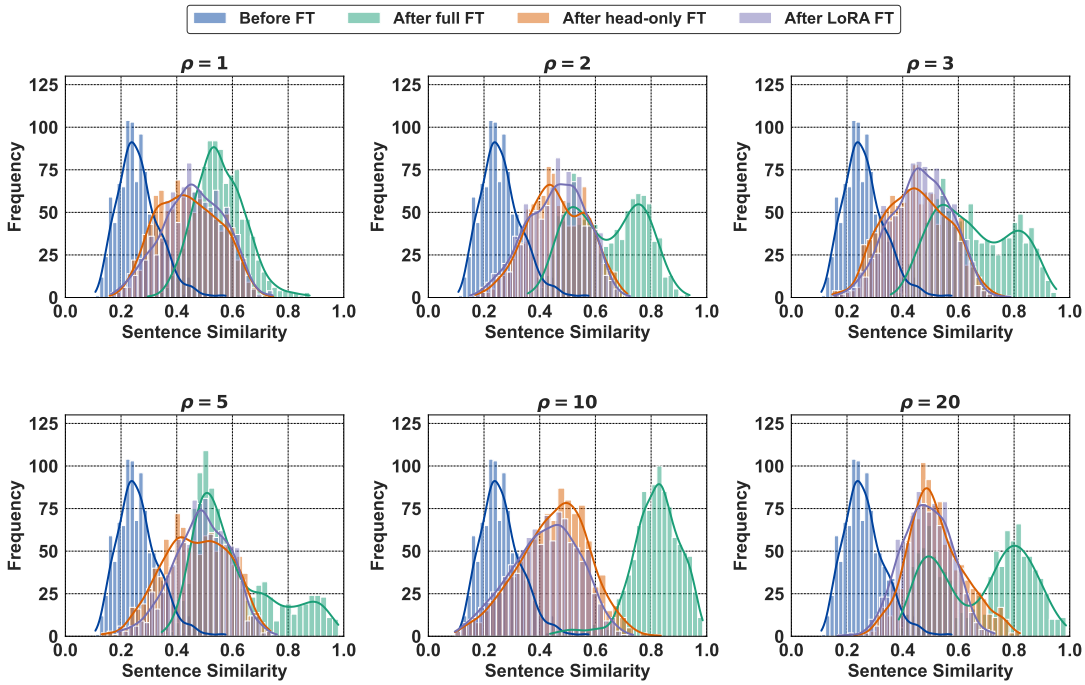


Figure 3: The histogram distribution of the top sentence similarity scores for $1,000$ generated samples from data extraction attacks, comparing different fine-tuning methods across varying levels of fine-tuning data duplication.

trated below 0.6. This suggests that LoRA enables effective adaptation while mitigating direct memorization, likely due to its parameter-efficient design, which restricts excessive retention of fine-tuning data.

For both head-only and LoRA fine-tuning, similarity distributions remain relatively stable across model sizes, with very few samples exceeding the 0.8 threshold. This finding aligns with the plagiarism-based memorization results in Fig. 1, suggesting that samples surpassing a similarity score of 0.8 are more likely to be flagged under plagiarism-based detection.

Similar patterns are observed in Fig. 3, where increasing ρ in full fine-tuning shifts sentence similarity scores rightward, with more samples exceeding 0.8. When $\rho \geq 5$, generated samples with near-perfect similarity (≈ 1.0) begin to appear. In contrast, head-only and LoRA fine-tuning maintain stable similarity distributions with minimal deviation as ρ increases. Even at the highest duplication level ($\rho = 20$), the vast majority of generated samples under these methods remain below the 0.8 similarity threshold, with distribution peaks consistently around 0.4 to 0.5.

5.2 LoRA Mitigates Memorization Even in High-Memorization Tasks

Besides the scientific writing task using the `Arxiver` dataset, we are also interested in the high-memorization tasks explored in Zeng et al. (2024). Following the same setup, we fine-tune GPT-2 on a dialogue task using the `ChatDoctor-HealthCareMagic` dataset (Li et al., 2023), which comprises $112k$ user queries and doctor responses, to investigate how the number of training samples influences the sentence similarity distribution across different fine-tuning methods.

Fig. 4 illustrates that after full fine-tuning, sentence similarity scores increase, with the $10k$ training set exhibiting the most concentrated distribution and the highest median value around 0.7. In contrast, head-only fine-tuning results in a moderate increase in similarity but with greater variance, indicating a less consistent adaptation. LoRA fine-tuning follows a similar trend, maintaining a broader distribution while keeping the highest similarity score below 0.8. This suggests that LoRA achieves a balance between adaptation and generalization, effectively preventing extractable memorization.

To evaluate the utility and generation capability of the models before and after fine-tuning, we con-

Metrics	Before	Full FT			Head-only FT			LoRA FT		
	FT	$1k$	$5k$	$10k$	$1k$	$5k$	$10k$	$1k$	$5k$	$10k$
BLEU ↑	0.008	0.007	0.006	0.006	0.007	0.007	0.008	0.007	0.007	0.007
R1 ↑	0.208	0.210	0.212	0.214	0.210	0.214	0.214	0.209	0.212	0.213
R2 ↑	0.017	0.017	0.018	0.019	0.017	0.017	0.019	0.017	0.018	0.018
RL ↑	0.081	0.080	0.082	0.084	0.080	0.082	0.081	0.082	0.081	0.082
LO ↑	0.025	0.022	0.023	0.024	0.026	0.023	0.022	0.026	0.022	0.023
FRE ↑	58.75	61.86	61.71	60.17	63.84	68.32	67.70	63.35	64.70	61.15
SMOG	11.32	10.82	10.94	11.01	10.51	9.80	9.87	10.62	10.39	10.91
DC	9.326	9.171	9.295	9.442	8.955	8.633	8.540	9.183	9.004	9.424

Table 1: Performance of models before and after fine-tuning across different fine-tuning strategies in terms of task-specific performance and linguistic complexity. Arrows indicate whether higher (↑) or lower (↓) values are preferable for each metric.

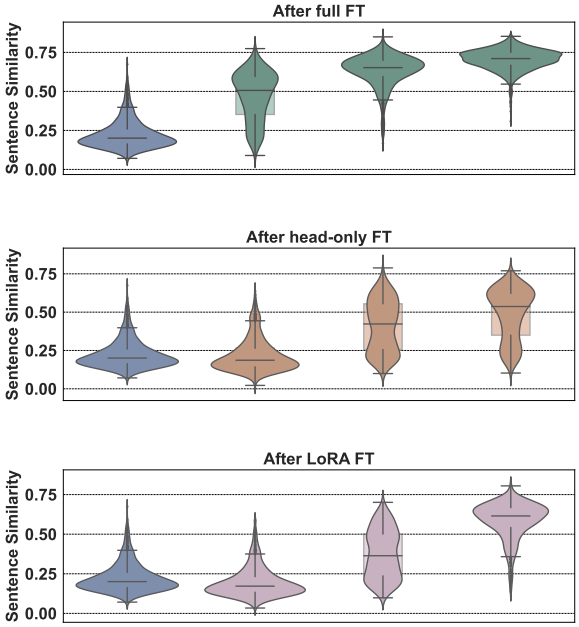


Figure 4: Violin plots depicting the top sentence similarity scores of generated samples before and after fine-tuning GPT-2 Small on the ChatDoctor dataset. Each row represents a different fine-tuning method, while the columns correspond to training sizes of $1k$, $5k$, and $10k$ samples, respectively. The width of each violin represents the density of similarity scores at different similarity levels. The middle line marks the median.

sider two aspects: (1) the **model's alignment with the expected context** in test data and (2) the **linguistic quality** of the generated text. For the first aspect, given reference labels in the dataset, we use BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) scores to quantify the similarity between the generated text and reference labels. For the second aspect, we assess linguistic coherence and readability using Lexical Overlap (LO) (Barzilay and Lapata, 2005), which quantifies word repetition across consecutive sentences. Additionally, we employ traditional readability metrics, including Flesch Reading Ease (FRE) (Flesch, 1948), where higher values indicate greater readability and simpler sentence structures; SMOG Index (Mc Laughlin, 1969), which measures syntactic complexity, with higher values signifying increased difficulty; and the Dale-Chall Readability Score (DC) (Dale and Chall, 1948), which assesses comprehensibility, where a higher score indicates more complex and less accessible text.

Correspondingly, we observe in Table 1 that full fine-tuning with larger training sizes achieves the highest content relevance for the fine-tuning task, as reflected in the highest ROUGE scores. In terms of linguistic complexity, readability improves after fine-tuning, with the FRE score increasing, especially for head-only fine-tuning. Meanwhile, the SMOG and DC scores decrease, indicating a shift toward simpler language. Overall, full fine-tuning maximizes task-specific performance but also significantly increases memorization, similar to pre-training, whereas head-only and LoRA fine-tuning maintain competitive model utility with minimal data extractability.

5.3 Hyperparameter Effects on Memorization in LoRA Fine-Tuning

Having established that LoRA reduces memorization, we next explore how its hyperparameters influence both utility and extraction susceptibility. In this experiment, we tune the hyperparameters involved in LoRA fine-tuning, including the rank r, the scaling factor α, and the dropout rate, to evaluate their impact on the fine-tuned model's utility and its resilience to data extraction attacks in terms

r	α	Dropout	BLEU	ROUGE-L	Mean	Std	Min	Max	Range
4	16	0.05	0.4507	0.5881	0.4792	0.0842	0.1567	0.7046	0.5479
16	16	0.05	0.4456	0.5854	0.4641	0.0888	0.1068	0.7513	0.6445
32	32	0.05	0.4528	0.5921	0.4639	0.0795	0.1519	0.7337	0.5818
4	8	0.1	0.4408	0.5772	0.4485	0.0923	0.1833	0.6761	0.4928
8	16	0.1	0.4442	0.5816	0.4621	0.0971	0.1611	0.7194	0.5583
16	16	0.1	0.4434	0.5837	0.4660	0.0905	0.1474	0.7047	0.5573

Table 2: BLEU and ROUGE-L scores, along with the statistical summary of sentence similarity scores, for fine-tuned models under different LoRA hyperparameter settings using GPT-2 Small on the Arxiver dataset.

of sentence similarity. The rank r defines the dimensionality of the low-rank adaptation matrices, with a higher r enabling more expressive adaptation. The scaling factor α controls the magnitude of LoRA updates, where a larger α strengthens adaptation but may reduce generalization. The dropout rate determines the fraction of LoRA updates deactivated during training, with higher dropout improving regularization but potentially limiting fine-tuning effectiveness (Srivastava et al., 2014).

Table 2 demonstrates the higher rank and scaling factors tend to improve ROUGE-L scores, with the configuration $r = 32$, $\alpha = 32$, dropout$= 0.05$, achieving the highest ROUGE-L score of 0.5921. The sentence similarity statistics reveal that models with a lower rank ($r = 4$) and smaller scaling factors generally exhibit lower mean similarity scores, suggesting reduced memorization. However, models with higher ranks and larger scaling factors tend to have slightly higher maximum similarity values, though still well below extraction thresholds.

Overall, increasing rank and scaling factor enhances content alignment (higher ROUGE-L scores) but introduces a slight increase in memorization risk. Dropout helps regularize memorization, reducing extreme similarities while maintaining task performance, reinforcing its role in mitigating overfitting-induced memorization. Despite these variations, all maximum similarity scores remain below 0.8, supporting the conclusion that LoRA fine-tuning effectively minimizes extractable memorization compared to full fine-tuning. This suggests that LoRA's low-rank adaptation mechanism naturally constrains memorization, even when hyperparameters are adjusted for stronger adaptation.

6 Concluding Remarks

In this paper, we revisit the impact of different fine-tuning strategies on LLM memorization. Our findings show that LoRA fine-tuning, designed for computationally efficient and effective training, achieves performance comparable to full fine-tuning while significantly reducing memorization risks. Notably, it leads to near-zero plagiarism-based memorization and effectively prevents strict verbatim memorization. Across various model sizes, data duplication levels, and hyperparameter configurations, LoRA fine-tuning consistently results in lower maximum similarity scores, demonstrating its ability to mitigate memorization risks. These findings contrast with prior work, highlighting a different perspective on fine-tuning's role in model memorization.

7 Limitations

While our study provides empirical insights into the impact of LoRA fine-tuning on model memorization, a more rigorous theoretical analysis is needed to explain why LoRA fine-tuning consistently maintains memorization below the threshold of extractable memorization.

Our experiments are limited to GPT-2, which, while widely studied, does not fully capture the behavior of larger and more advanced LLMs. Extending this analysis to state-of-the-art models, such as LLaMA 3 and other alternatives, will provide a broader understanding of LoRA's effectiveness at scale and across different architectures.

While our findings suggest that LoRA fine-tuning mitigates memorization risks, it remains unclear whether more advanced data extraction attacks, such as those proposed in Zhang et al. (2023); Kassem et al. (2024), could enhance the extractability of fine-tuning data. Future research should investigate the interaction between LoRA fine-tuning and advanced extraction methods to determine whether further safeguards are needed to prevent memorization risks.

References

Muhammet Hatipoglu Alican Acar, Alara Dirik. 2024. Arxiver. https://huggingface.co/datasets/neuralwork/arxiver.

Regina Barzilay and Mirella Lapata. 2005. Modeling local coherence: An entity-based approach. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 141–148. Association for Computational Linguistics.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *Proceedings of the 28th USENIX Conference on Security Symposium*, pages 267–284, USA. USENIX Association.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650. USENIX Association.

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational Research Bulletin*, pages 37–54.

Rudolf Franz Flesch. 1948. A new readability yardstick. *The Journal of Applied Psychology*, 32 3:221–33.

Clinton Gormley and Zachary Tong. 2015. *Elasticsearch: The Definitive Guide*, 1st edition. O'Reilly Media, Inc.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *International conference on machine learning*, pages 2790–2799. PMLR.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Nikhil Kandpal, Eric Wallace, and Colin Raffel. 2022. Deduplicating training data mitigates privacy risks in language models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 10697–10707.

Aly M Kassem, Omar Mahmoud, Niloofar Mireshghallah, Hyunwoo Kim, Yulia Tsvetkov, Yejin Choi, Sherif Saad, and Santu Rana. 2024. Alpaca against vicuna: Using llms to uncover memorization of llms. *arXiv preprint arXiv:2403.04801*.

Jooyoung Lee, Thai Le, Jinghui Chen, and Dongwon Lee. 2023. Do language models plagiarize? In *Proceedings of the ACM Web Conference 2023*, pages 3637–3647. Association for Computing Machinery.

Danny D. Leybzon and Corentin Kervadec. 2024. Learning, forgetting, remembering: Insights from tracking LLM memorization during training. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 43–57. Association for Computational Linguistics.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6).

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81. Association for Computational Linguistics.

G Harry Mc Laughlin. 1969. SMOG grading-a new readability formula. *Journal of reading*, 12(8):639–646.

Fatemehsadat Mireshghallah, Kartik Goyal, Archit Uniyal, Taylor Berg-Kirkpatrick, and Reza Shokri. 2022a. Quantifying privacy risks of masked language models using membership inference attacks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8332–8347. Association for Computational Linguistics.

Fatemehsadat Mireshghallah, Archit Uniyal, Tianhao Wang, David Evans, and Taylor Berg-Kirkpatrick. 2022b. An empirical analysis of memorization in fine-tuned autoregressive language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1816–1826. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, page 311–318. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI Blog*, 1(8):9.

Avi Schwarzschild, Zhili Feng, Pratyush Maini, Zachary Chase Lipton, and J Zico Kolter. 2024. Rethinking LLM memorization through the lens of adversarial compression. In *The Thirty-eighth Annual*

Conference on Neural Information Processing Systems.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. 15(1):1929–1958.

Shenglai Zeng, Yaxin Li, Jie Ren, Yiding Liu, Han Xu, Pengfei He, Yue Xing, Shuaiqiang Wang, Jiliang Tang, and Dawei Yin. 2024. Exploring memorization in fine-tuned language models. pages 3917–3948, Bangkok, Thailand.

Zhexin Zhang, Jiaxin Wen, and Minlie Huang. 2023. ETHICIST: Targeted training data extraction through loss smoothed soft prompting and calibrated confidence estimation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12674–12687. Association for Computational Linguistics.