TO MEMORIZE OR NOT TO MEMORIZE: AN ANALYSIS OF SUPERVISED FINE-TUNING IN LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Supervised fine-tuning (SFT) is a cornerstone technique for adapting large language models (LLMs) to specific domains and tasks. However, its propensity to induce verbatim memorization of training data poses significant risks to safety, privacy, and generalization. This paper presents an empirical analysis of the mechanisms underlying memorization within LLMs during SFT. Our findings confirm that SFT is a direct driver of memorization, with a clear positive correlation between the number of training epochs and the rate of verbatim data recall. The characteristics of the fine-tuning dataset are a critical determinant of memorization. We demonstrate that models trained on broad, open-domain datasets exhibit substantially more memorization than those trained on narrow, domain-specific ones, highlighting a crucial trade-off between model versatility and data containment. Furthermore, we indicate that verbatim memorization is suppressed when the training data includes inputs with high similarity paired with dissimilar outputs. We posit that this phenomenon is not a desirable mitigation strategy but rather a symptom of the model being exposed to conflicting data signals. These findings underscore the complex trade-offs in SFT and stress the importance of understanding these underlying dynamics to develop LLMs that are both capable and secure.

1 INTRODUCTION

Supervised fine-tuning (SFT) is a cornerstone technique for adapting large language models (LLMs) to specific domains and tasks, serving as the primary method for aligning model behavior with human preferences and injecting specialized knowledge. However, this powerful technique has a significant side effect: it often causes the model to verbatim memorize portions of its training data. This propensity for memorization poses serious risks to model safety by potentially reproducing harmful content, causing copyright infringement, compromises user privacy by leaking sensitive information, and can impair generalization by encouraging rote learning over abstract pattern recognition. While foundational research has conclusively demonstrated that LLMs memorize training data (Carlini et al., 2021), the specific mechanisms and contributing factors that govern this behavior *during the SFT process* remain underexplored, creating a critical need to move beyond simply detecting the phenomenon to actively controlling it.

This paper directly addresses this gap by presenting a rigorous empirical analysis of the mechanisms underlying memorization within SFT. Our investigation systematically tracks the temporal evolution of memorization across training epochs while simultaneously analyzing contributing factors at two distinct levels of granularity: at the *global level*, we analyze how macro-level dataset properties, such as overall diversity, influence memorization. Concurrently, at the *local level*, we examine micro-level data dynamics, specifically how a model responds to conflicting signals arising from an instance's immediate neighborhood—where highly similar inputs are paired with divergent outputs. By also tracking these effects across training epochs, we dissect the trade-offs inherent in the fine-tuning process. This investigation not only informs the development of more secure LLMs but also illuminates fundamental properties of the fine-tuning process, compelling a re-examination of SFT through the lens of its memorization behaviors.

1.1 Memorization Detection and Measurement

Memorization research began with establishing extraction attacks and three scaling laws (model capacity, data duplication, context length) (Carlini et al., 2021). Detection methods were expanded through membership inference attacks(Duan et al., 2024), prefix extraction (Hayes et al., 2025), neuron activation analysis (Slonski, 2024), etc.

The comprehensive paper (Xiong et al., 2025) mentioned a number of factors to be tested, but in the area of SFT memorization, there is a lack of comprehensive understanding. Memorization is also blamed for negatively affecting the performance in general (Bayat et al., 2024), and in the SFT of LLM (Chu et al., 2025), making the investigation of properties of SFT of LLM a timely need to safety and as a reference factor to technology selection.

It is reported that LLM memorization is influenced by the duration of training gives increasing mem accuracy versus epoch on memorization on randomly generated strings, but the setting of a small dataset with more than 100 epochs of this experiment is unrealistic (Speicher et al., 2024).

This justifies to test the trend of memorization in SFT in a more realistic settings to be supplementary to this problem. If we want to quantify memorization, we should do it within a model size and epoch that is widely accepted, and within a reasonable tuning of model size not too small and number of epochs not too big.

1.2 Memorization versus Training Dynamics and Fine-tuning Methods

Training epochs and temporal patterns are fundamental to learning dynamics, yet memorization emergence throughout fine-tuning remains under-characterized. Most studies and tech reports measure endpoint retention rather than tracking when/why memorization occurs, despite temporal understanding being critical for prevention-based approaches.

1.3 Dataset Characteristics and Diversity Effects

Dataset diversity fundamentally affects model performance and generalization (Gong et al., 2023), yet memorization relationships remain poorly understood. Current approaches build and select dataset based on diversity and has achieved some performance gain(Bukharin et al., 2024) (Wang et al., 2024).

Another follow-up study uses datasets with duplicated data to investigate memorization (Carlini et al., 2023)–but the presence of completely duplicated data–letting alone artificially (intentionally) duplicating a subset of the data for training is far from a realistic case–if we want to investigate memorization, we should investigate in real data.

While Large Language Model (LLM) memorization is known to be more intense for knowledge-based than for reasoning-based tasks (Wang et al., 2025), recent focus has shifted to quantifying the diversity of the underlying data rather than relying on subjective labels (Zhao et al., 2024). Methodologies for this quantification are emerging, such as diversity coefficients for pre-training (Miranda et al., 2025) and predictive diversity scores for fine-tuning (Yang et al., 2025). Despite the success of these scores in predicting model performance, their connection to memorization is still an open question. Furthermore, inspired by taxonomies that characterize properties of recall like confusion (Prashanth et al., 2025), we are also interested in how these properties relate to memorization phenomena.

1.4 *Conflicting Training Signals*

Machine learning research has established relationships between model performance, memorization, and data removal, with the field of machine-unlearning (Ginart et al., 2019) dedicated to selectively removing data influence while preserving model utility. In Large Language Models, existing work has examined memorization's impact on generalization (Bayat et al., 2024) and the adverse effects of data ambiguity on fine-tuning performance (Xu et al., 2024). However, a critical connection remains underexplored in LLMs: the relationship between data ambiguity and memorization mechanisms. When training data contains similar inputs with divergent outputs, models cannot identify generalizable patterns and must resort to memorization. Investigating how this specific form of data ambiguity—-where proximate inputs yield conflicting outputs—-influences memorization mechanisms in LLMs represents a vital extension of existing research and offers key insights into when and why models abandon pattern learning in favor of rote memorization.

This heuristic understanding is complemented by insights from statistical learning theory, which provides a more formal lens through which to view this problem. From this perspective, the diversity of outputs conditioned on similar inputs bounds the Bayesian risk, hence the lowest possible prediction error on exact training data for any classifier. A simple illustration of this is found in the relationship between classification error and the conditional probability of a class given an input, as detailed in foundational texts on pattern recognition (Devroye et al., 1996). The inherent ambiguity in the training data, where a single input x can be associated with multiple correct outputs $\{y_1, y_2, \ldots\}$, introduces a fundamental level of irreducible error. This raises a compelling question for future research: can a continuous or probabilistic version of these risk properties be developed specifically for the phenomenon of memorization in Large Language Models?

1.5 *Evaluation Frameworks*

Evaluating large language models (LLMs) requires a comprehensive understanding of their behavior, particularly memorization. However, current assessment of this phenomenon is fragmented (Xiong et al., 2025). The field often lacks consistent definitions, relies on singular metrics, and misses standardized benchmarks across various fine-tuning methods. While some research proposes complexity-based metrics to assess variables Morris et al. (2025), a comprehensive and unified framework for memorization remains underdeveloped.

This fragmentation is particularly problematic as memorization can be a nuanced, fine-grained issue. For example, recent work indicates that entity-level memorization can be highly extractable (Zhou et al., 2024). This implies that capturing these details requires a diverse and sometimes fine-grained set of easily comparable metrics. Although tools to measure text diversity exist (Shaib et al., 2025), they have not seen systematic application to LLM memorization contexts. Therefore, we propose using a broader spectrum of metrics to quantify memorization, understand their uses, and analyze their relationships.

Our work addresses these gaps through a systematic empirical analysis of memorization mechanisms during supervised fine-tuning (SFT). We provide a comprehensive investigation into the effects of training epochs, dataset diversity, and natural conflict resolution on what models memorize. To properly quantify these effects in a realistic SFT scenario, we also employ a suite of metrics including Edit Distance (Wagner & Fischer, 1974) and Effective Dimensionality, calculated using the *scikit-dimension* package (Bac et al., 2021), thereby providing a more thorough investigation than what current research trends suggest.

2 Methods

2.1 Experiment Setup

Our primary datasets for this study are derived from `alpaca-cleaned` (Taori et al., 2023), `code-alpaca` (Chaudhary, 2023), and `finance-alpaca` datasets (Bharti, 2024), and a medical QA dataset (Chen et al., 2024) . We developed a preprocessing pipeline to standardize these datasets, with the goal of producing a final version for each containing approximately 10^4 rows and texts with not too short length.

We implemented a preprocessing pipeline that iterates through the `alpaca-cleaned` Taori et al. (2023), `code-alpaca` (Chaudhary, 2023), and `finance-alpaca` datasets (Bharti, 2024). For the medical QA dataset, preprosessing script filters it to select the first 8192 patient cases that specify an age, reformats the data into an instruction-input-output structure with pandas, and saves the result as a JSON file. For each one, the script first loads the data from its JSON file and performs a detailed statistical analysis on the text in the `instruction`, `input`, and `output` columns, calculating metrics like minimum, median, mean, and maximum word/character counts, along with skewness. It then visualizes these text length distributions. The core preprocessing step involves filtering each dataset to retain only those entries where the combined word count of the instruction, input, and output fields exceeds 35 words. Finally, if the resulting filtered dataset contains more than 10,016

Table 1: Core Memorization Metrics

Metric	Definition
Full Memorization Rate (F)	The probability that the model output exactly matches the reference data. $$F = P(s_{\text{ref}} = s_{\text{out}})$$
Prefix Continuation Length (n_{pre})	The number of consecutive matching tokens from the start of the sequence until the first mismatch. $$n_{\text{pre}}(s_{\text{ref}}, s_{\text{out}}) = \max\{k : s_{\text{ref}}[j] = s_{\text{out}}[j] \text{ for all } j \in [1, k]\}$$
Memorization Ratio (M)	The proportion of the reference sequence that is memorized from the beginning. $$M = \frac{n_{\text{pre}}}{n_{\text{ref}}}$$

entries, it is randomly sampled down to that size, and the final, filtered dataset is saved as a new JSON file.

We are performing Supervised Fine-Tuning (SFT) on the *Llama-3.1-8B-Instruct* model using the `flashattn2` booster for optimized attention and a `bf16` compute type without quantization. The training is conducted on the datasets described previously with a context length limit of *2048* tokens and a batch size of *1*. This is a full-parameter fine-tuning run, as no LoRA adapter is being created. The process uses a learning rate of *1e-6* with a `cosine` scheduler and *3* warmup steps, applying gradient clipping at a max norm of *1.0*. Although planned for 20 epochs, we executed the first *3* epochs, saving a checkpoint after each one.

2.2 Quantification of Memorization

To quantify memorization and related miscellaneous factors, we use notation as such:

Memorization occurs when the model output string $s_{\text{out}} = \text{LLM}(s_{\text{input}})$ closely resembles the training (reference) data s_{ref}. For indexing, i indexes data instances, j indexes tokens within sequences, lowercase n for lengths, uppercase N for length distributions, distances prefixed with d (e.g., $d_{\text{Levenshtein}}(s_{\text{ref}}, s_{\text{out}})$). We use the convention from metric spaces to use $B(\cdot, r)$ to denote a neighborhood of radius r around a point.

We also employ metrics from two complementary areas, string processing and modern natural language understandings.

Table 2: String Processing Metrics

Metric	Notation / Description
Levenshtein distance	$d_{\text{Levenshtein}}(s_{\text{ref}}, s_{\text{out}})$
Longest common subsequence	$\text{LCS}(s_{\text{ref}}, s_{\text{out}})$
Largest shared n-grams	Measures the longest contiguous block of n tokens shared between two strings.
ROUGE scores	A set of metrics for evaluating similarity based on overlapping n-grams and subsequences.

Table 3: Natural Language Understanding Metrics

Metric	**Definition**
Text Sequence Embeddings	Vector representations of text sequences. Notation: $EMB(s_{\text{input}})$
Participation Ratio (D_{PCA})	Measures how evenly the variance of embedding data is spread across principal components. The λ_i are the eigenvalues from the PCA of the embedding data. A higher ratio indicates data uses many dimensions effectively. $$D_{\text{PCA}}(EMB(S_{\text{input}})) = \frac{\left(\sum_{i=1}^{n} \lambda_i\right)^2}{\sum_{i=1}^{n} \lambda_i^2}$$

3 Experiments and Results

3.1 More training, More Memorization

Higher epoch number usually yield higher memorization by different metrics. This property in F is shown in figure 1.

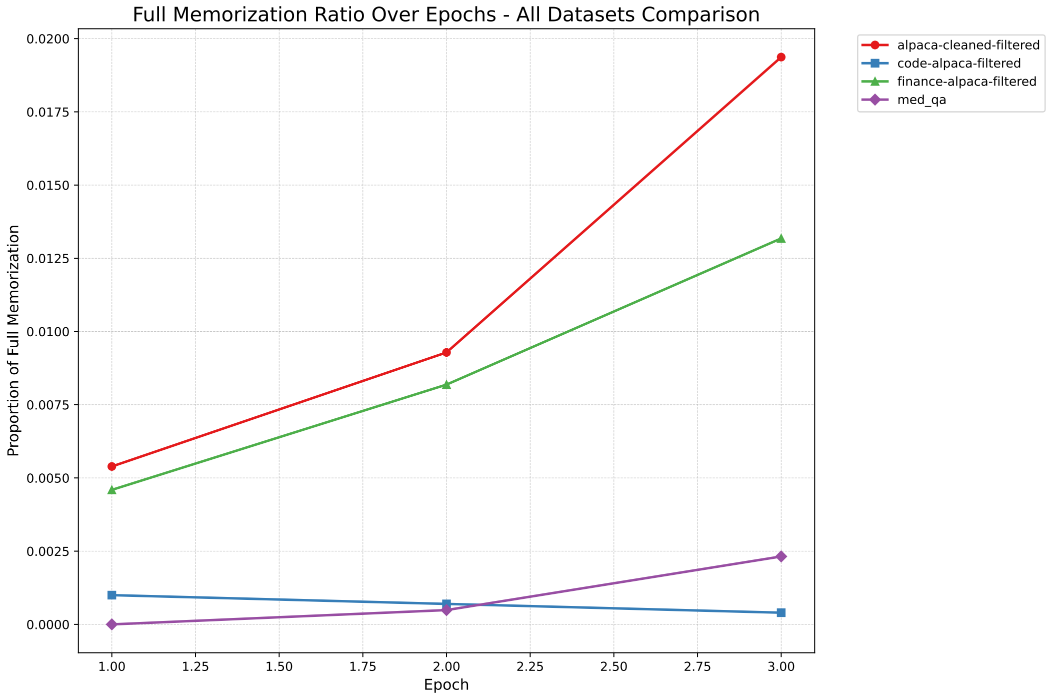

Figure 1: More Epoch

3.2 Broader Data, More Memorization

We propose that models may remember more when trained on datasets with higher diversity, based on the heuristic that a diverse dataset contains input-output pairs that are less related to each other. This reduced correlation could allow independent optimization without mutual interference, analogous to memorizing outputs conditioned on distinct inputs rather than overlapping patterns.

To test this hypothesis, we need an objective method to quantify the ”broadness” of the data set beyond simple domain comparisons. Although different domains (finance vs. medicine) may inherently vary in scope, we require a more systematic approach to measure data diversity within and across categories.

We compute embeddings for each dataset and apply Principal Component Analysis (PCA) to characterize dimensionality. Our diversity metrics are the number of PCA components required to explain 90% of variance, and the participation ratio D_{PCA}. These measures indicate whether inputs are similar to each other or more spread out in the representation space.

We chose PCA over cosine similarity analysis because in high-dimensional normalized embedding spaces, angle distributions with points uniformly distributed on the sphere concentrate heavily around $\pi/2$ due to hypersphere geometry, making cosine similarity less discriminative for measuring data relationships for a dataset as a whole.

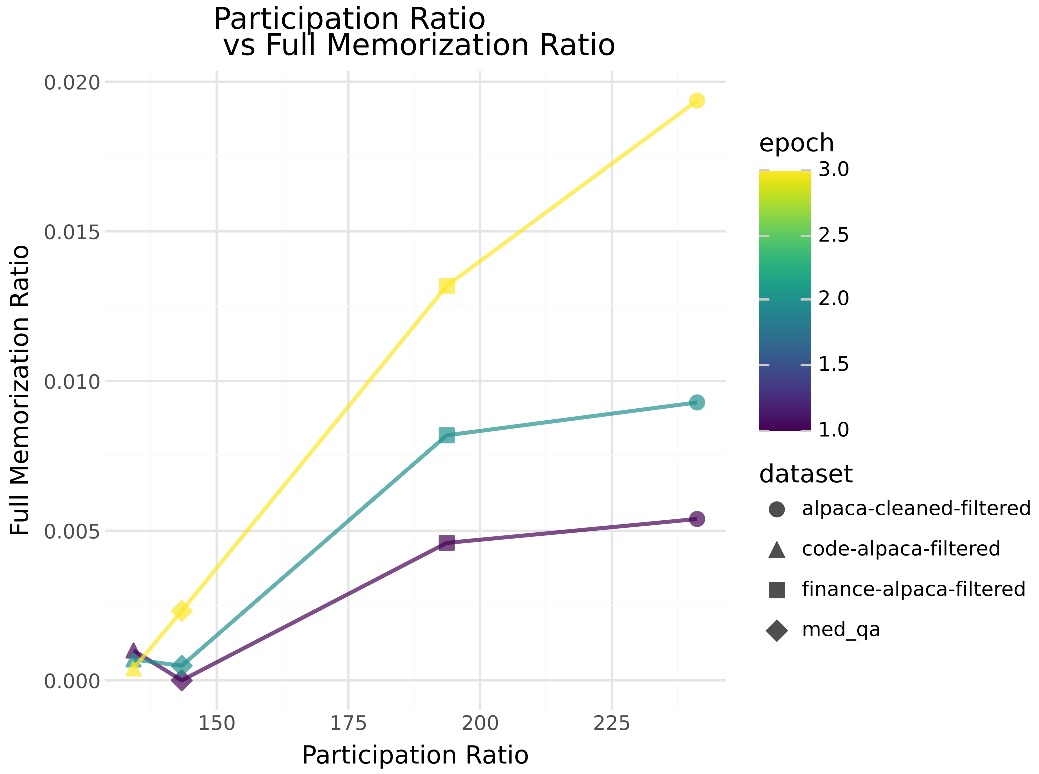

Figure 2: Memorization is higher at higher Participation Ratio of Data

Testing across four datasets, we observe that memorization increases with both higher participation ratios and greater numbers of components needed to capture 90% variance. This relationship remains consistent when controlling for training epochs (1, 2, and 3), supporting the hypothesis that data diversity enhances memorization through reduced interference between training examples.

3.3 Behavior Within a Single Dataset: Conflicting Entries Limits Memorization

For each point in the data set i, we compute $s_{input}[i]$ and find its neighbors in all data S_{input} that are within the distance of a threshold THS, denote the indices $I := B(s_{input}[i], r = THS)$ We compute the statistics of the distances $d[i][I] = [d_{Levenshtein}(s_{out}[i'], s_{out}[i]) \; \forall i' \in I]$, for the data that has such neighbors, the density at high $min(d[i][I])$ and high M must be low as a result of confusion.

The algorithm to characterize this feature is illustrated as follows

Algorithm 1 Analyzing Conflicting Entries and Their Impact on Memorization

```
Require: Training dataset $\mathcal{D} = \{(s_{\text{input}}[i], s_{\text{ref}}[i])\}_{i=1}^{N}$
Require: Distance threshold $\tau$ for neighbor detection
Require: Edit distance matrices $d_{\text{prefix}}$, $d_{\text{suffix}}$ (precomputed)
Initialize neighbor sets $\mathcal{N} = \{\}$ {For each data point}
Initialize statistics $\mathcal{S} = \{\}$ {Conflict measures}
for $i = 1$ to $N$ do
    $I_{\text{prefix}} \leftarrow B(s_{\text{input}}[i], \tau) = \{j : d_{\text{prefix}}[i, j] < \tau$ and $j \neq i\}$ {Find prefix neighborhood}
    $I_{\text{neighbors}} \leftarrow \emptyset$
    for $j \in I_{\text{prefix}}$ do
        if $d_{\text{suffix}}[i, j] < \tau$ then
            $I_{\text{neighbors}} \leftarrow I_{\text{neighbors}} \cup \{j\}$
        end if
    end for
    $\mathcal{N}[i] \leftarrow I_{\text{neighbors}}$ {Store neighbors with both prefix and suffix similarity}
    if $|\mathcal{N}[i]| > 0$ then
        $\mathbf{d}_{\text{ref}}[i] \leftarrow [d_{\text{Levenshtein}}(s_{\text{ref}}[i], s_{\text{ref}}[j]) : j \in \mathcal{N}[i]]$
        $\mathbf{d}_{\text{prefix}}[i] \leftarrow [d_{\text{prefix}}[i, j] : j \in \mathcal{N}[i]]$
        $\mathbf{d}_{\text{suffix}}[i] \leftarrow [d_{\text{suffix}}[i, j] : j \in \mathcal{N}[i]]$
        $\mathcal{S}[i] \leftarrow \{$
            $n_{\text{neighbors}} : |\mathcal{N}[i]|,$
            $d_{\text{prefix}}^{\min} : \min(\mathbf{d}_{\text{prefix}}[i]),$
            $d_{\text{prefix}}^{\max} : \max(\mathbf{d}_{\text{prefix}}[i]),$
            $d_{\text{prefix}}^{\mu} : \mathbb{E}[\mathbf{d}_{\text{prefix}}[i]],$
            $d_{\text{ref}}^{\min} : \min(\mathbf{d}_{\text{ref}}[i]),$
            $d_{\text{ref}}^{\max} : \max(\mathbf{d}_{\text{ref}}[i]),$
            $d_{\text{ref}}^{\mu} : \mathbb{E}[\mathbf{d}_{\text{ref}}[i]]$
        $\}$
    else
        $\mathcal{S}[i] \leftarrow \{$all statistics specified$\}$
    end if
end for
return $\mathcal{N}, \mathcal{S}$ {Neighbor lists and conflict statistics}
```

Hypothesis: Points with high confusion (similar inputs mapping to dissimilar outputs) will exhibit low memorization density, as the model struggles to distinguish between conflicting training signals.

Density Analysis: We expect $P(\text{memorized} \mid m_i \geq \tau_{\min}, M_i \geq \tau_{\max})$ to be lower than the baseline memorization rate.

We plot the normalized minimum of edit distance between the output of a candidate and all other outputs corresponding to the candidates within this neighborhood d_{ref}^{min} against the memorization ratio $M := n_{pre}/n_{ref}$. The conditional expectation (estimated from bins with widths 0.2) of the Nd_{min} decreases with respect to M. And the top-right corner of the density plot shows low density.

3.4 Discussion

Our empirical investigation reveals a multi-faceted relationship between supervised fine-tuning (SFT) and memorization, highlighting critical trade-offs for developing capable and secure LLMs. The results confirm expected behaviors while also uncovering more complex dynamics, compelling a more nuanced view of the SFT process.

3.4.1 Epochs, Diversity, and the Propensity for Memorization

The observation that *memorization intensifies with additional training epochs* is an expected, yet foundational, result. This establishes a distinct ”cost” associated with extended training, where each

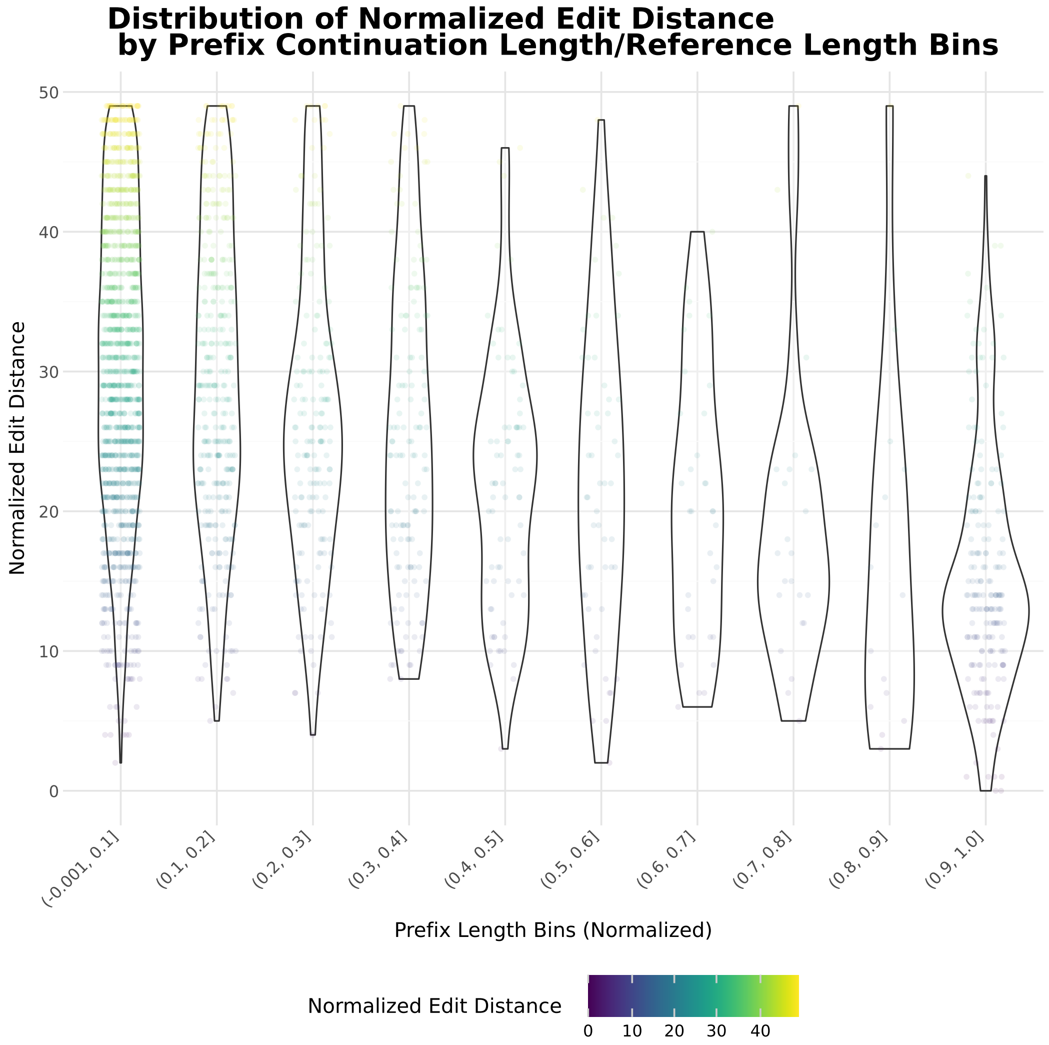

Figure 3: Alpaca-cleaned-filtered, Similar results are observed on the other data sets on the third epoch as well

pass over the data reinforces rote learning alongside desired adaptations. This monotone relationship provides a critical baseline against which our other findings can be interpreted.

More revealing is the strong correlation we identify between *dataset diversity and memorization*. Our use of PCA-based metrics provides quantitative support for the heuristic that broader, more varied datasets encourage memorization. This phenomenon can be understood through the lens of learning interference. In narrow, domain-specific datasets (e.g., medical QA), a high degree of conceptual overlap among data points incentivizes the model to learn underlying patterns. Conversely, in broad, open-domain datasets (e.g., Alpaca), data points are often more isolated in the semantic space. The model has less incentive to find unifying principles and can more readily memorize each input-output pair as an isolated fact. This dynamic presents a significant trade-off for practitioners: a diverse dataset may enhance a model's versatility, but it does so at the cost of heightened memorization, increasing the risks of privacy leaks and the reproduction of undesirable content.

3.4.2 The Paradox of Conflicting Training Signals

A more complex finding is that *verbatim memorization is suppressed when training data contains similar inputs paired with dissimilar outputs*. At first glance, this might appear to be a potential mitigation strategy. We posit, however, that this is not a desirable mechanism but rather a *symptom of model confusion*.

When a model is presented with conflicting signals—where nearly identical prompts correspond to disparate target outputs—it cannot form a stable, generalizable mapping. The optimization process is subjected to competing gradients, which prevents the model from converging to a confident prediction for any single output. Consequently, it fails to reliably memorize any specific response. This outcome is not indicative of a ”forgetting” mechanism but rather of an impaired ability to learn effectively from ambiguous or contradictory data. This phenomenon highlights a fundamental tension: while data consistency may facilitate memorization, data inconsistency can hinder learning altogether, leading to unreliable model behavior. Therefore, attempting to leverage such conflicting data points as a tool to reduce memorization might be counterproductive, degrading overall model performance.

3.4.3 Implications for Safe and Effective SFT

These findings have direct implications for the practical application of SFT.

1. *Controlling Training Duration.* Practitioners should treat training epochs as a finite resource that directly contributes to a ”memorization budget.” Exceeding this budget, especially on sensitive data, may yield only marginal performance gains at a high cost to data security.

2. *Evaluating Dataset Diversity as a Risk Factor.* Dataset diversity should be considered a key risk factor in memorization audits. Our PCA-based approach offers a concrete method for quantifying this risk prior to fine-tuning, allowing for more informed decisions about dataset curation and the balance between model generality and data containment.

3. *Using Data Conflicts as a Diagnostic Tool.* The suppression of memorization via conflicting signals should be leveraged as a diagnostic tool. Identifying these data points can help curators pinpoint ambiguity, noise, or inherent task difficulty within their datasets. Resolving these conflicts, rather than introducing them, is essential for building robust and predictable models.

In conclusion, managing memorization is not a matter of applying a single technique but requires navigating a complex interplay of training duration, dataset characteristics, and intra-dataset dynamics. Our work contributes to a clearer understanding of this landscape, emphasizing that the path toward creating models that are both highly capable and fundamentally secure lies in a more profound and quantitative understanding of the data used to shape them.

4 Conclusion

This paper analyzes the memorization in Large Language Models (LLMs) during supervised fine-tuning (SFT), addressing gaps in current research. Our findings confirm that SFT directly drives memorization, with a positive correlation between training epochs and verbatim data recall. We demonstrate that the characteristics of the fine-tuning dataset are crucial determinants: broad, open-domain datasets induce substantially more memorization than narrow, domain-specific ones, highlighting a trade-off between model versatility and data containment. Furthermore, we observe that verbatim memorization is suppressed when training data includes highly similar inputs paired with dissimilar outputs, which we interpret as a symptom of conflicting data signals rather than a desirable mitigation strategy. These insights underscore the complex dynamics of SFT, suggesting that a finer look at these properties helps understand memorization, but a finer look at the memorization helps understand SFT as well.

REFERENCES

Jonathan Bac, Evgeny M. Mirkes, Alexander N. Gorban, Ivan Tyukin, and Andrei Zinovyev. Scikit-dimension: A python package for intrinsic dimension estimation. *Entropy*, 23(10), 2021. ISSN 1099-4300. doi: 10.3390/e23101368. URL https://www.mdpi.com/1099-4300/23/10/1368.

Reza Bayat, Mohammad Pezeshki, Elvis Dohmatob, David Lopez-Paz, and Pascal Vincent. The pitfalls of memorization: When memorization hurts generalization, 2024. URL https://arxiv.org/abs/2412.07684.

Gaurang Bharti. finance-alpaca (revision 51d16b6), 2024. URL https://huggingface.co/datasets/gbharti/finance-alpaca.

Alexander Bukharin, Shiyang Li, Zhengyang Wang, Jingfeng Yang, Bing Yin, Xian Li, Chao Zhang, Tuo Zhao, and Haoming Jiang. Data diversity matters for robust instruction tuning, 2024. URL https://arxiv.org/abs/2311.14736.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models, 2021. URL https://arxiv.org/abs/2012.07805.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models, 2023. URL https://arxiv.org/abs/2202.07646.

Sahil Chaudhary. Code alpaca: An instruction-following llama model for code generation. https://github.com/sahil280114/codealpaca, 2023.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. Huatuogpt-o1, towards medical complex reasoning with llms, 2024. URL https://arxiv.org/abs/2412.18925.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, and Yi Ma. Sft memorizes, rl generalizes: A comparative study of foundation model post-training, 2025. URL https://arxiv.org/abs/2501.17161.

Luc Devroye, László Györfi, and Gábor Lugosi. *A Probablistic Theory of Pattern Recognition*, volume 31. 01 1996. ISBN 978-1-4612-6877-2. doi: 10.1007/978-1-4612-0711-5.

Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. Do membership inference attacks work on large language models?, 2024. URL https://arxiv.org/abs/2402.07841.

Antonio Ginart, Melody Y. Guan, Gregory Valiant, and James Zou. Making ai forget you: Data deletion in machine learning, 2019. URL https://arxiv.org/abs/1907.05012.

Youdi Gong, Guangzhen Liu, Yunzhi Xue, Rui Li, and Lingzhong Meng. A survey on dataset quality in machine learning. *Information and Software Technology*, 162:107268, 2023. ISSN 0950-5849. doi: https://doi.org/10.1016/j.infsof.2023.107268. URL https://www.sciencedirect.com/science/article/pii/S0950584923001222.

Jamie Hayes, Marika Swanberg, Harsh Chaudhari, Itay Yona, Ilia Shumailov, Milad Nasr, Christopher A. Choquette-Choo, Katherine Lee, and A. Feder Cooper. Measuring memorization in language models via probabilistic extraction, 2025. URL https://arxiv.org/abs/2410.19482.

Brando Miranda, Alycia Lee, Sudharsan Sundar, Allison Casasola, Rylan Schaeffer, Elyas Obbad, and Sanmi Koyejo. Beyond scale: The diversity coefficient as a data quality metric for variability in natural language data, 2025. URL https://arxiv.org/abs/2306.13840.

John X. Morris, Chawin Sitawarin, Chuan Guo, Narine Kokhlikyan, G. Edward Suh, Alexander M. Rush, Kamalika Chaudhuri, and Saeed Mahloujifar. How much do language models memorize?, 2025. URL https://arxiv.org/abs/2505.24832.

USVSN Sai Prashanth, Alvin Deng, Kyle O'Brien, Jyothir S V, Mohammad Aflah Khan, Jaydeep Borkar, Christopher A. Choquette-Choo, Jacob Ray Fuehne, Stella Biderman, Tracy Ke, Katherine Lee, and Naomi Saphra. Recite, reconstruct, recollect: Memorization in lms as a multifaceted phenomenon, 2025. URL https://arxiv.org/abs/2406.17746.

Chantal Shaib, Joe Barrow, Jiuding Sun, Alexa F. Siu, Byron C. Wallace, and Ani Nenkova. Standardizing the measurement of text diversity: A tool and a comparative analysis of scores, 2025. URL https://arxiv.org/abs/2403.00553.

Eduardo Slonski. Detecting memorization in large language models, 2024. URL https://arxiv.org/abs/2412.01014.

Till Speicher, Mohammad Aflah Khan, Qinyuan Wu, Vedant Nanda, Soumi Das, Bishwamittra Ghosh, Krishna P. Gummadi, and Evimaria Terzi. Understanding memorisation in llms: Dynamics, influencing factors, and implications, 2024. URL https://arxiv.org/abs/2407.19262.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

Robert A. Wagner and Michael J. Fischer. The string-to-string correction problem. *J. ACM*, 21 (1):168–173, January 1974. ISSN 0004-5411. doi: 10.1145/321796.321811. URL https://doi.org/10.1145/321796.321811.

Peiqi Wang, Yikang Shen, Zhen Guo, Matthew Stallone, Yoon Kim, Polina Golland, and Rameswar Panda. Diversity measurement and subset selection for instruction tuning datasets, 2024. URL https://arxiv.org/abs/2402.02318.

Xinyi Wang, Antonis Antoniades, Yanai Elazar, Alfonso Amayuelas, Alon Albalak, Kexun Zhang, and William Yang Wang. Generalization v.s. memorization: Tracing language models' capabilities back to pretraining data, 2025. URL https://arxiv.org/abs/2407.14985.

Alexander Xiong, Xuandong Zhao, Aneesh Pappu, and Dawn Song. The landscape of memorization in llms: Mechanisms, measurement, and mitigation, 2025. URL https://arxiv.org/abs/2507.05578.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for llms: A survey, 2024. URL https://arxiv.org/abs/2403.08319.

Yuming Yang, Yang Nan, Junjie Ye, Shihan Dou, Xiao Wang, Shuo Li, Huijie Lv, Mingqi Wu, Tao Gui, Qi Zhang, and Xuanjing Huang. Measuring data diversity for instruction tuning: A systematic analysis and a reliable metric, 2025. URL https://arxiv.org/abs/2502.17184.

Dora Zhao, Jerone T. A. Andrews, Orestis Papakyriakopoulos, and Alice Xiang. Position: Measure dataset diversity, don't just claim it, 2024. URL https://arxiv.org/abs/2407.08188.

Zhenhong Zhou, Jiuyang Xiang, Chaomeng Chen, and Sen Su. Quantifying and analyzing entity-level memorization in large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19741–19749, Mar. 2024. doi: 10.1609/aaai.v38i17.29948. URL https://ojs.aaai.org/index.php/AAAI/article/view/29948.

A Appendix

You may include other additional sections here.