Analyzing the Effects of Supervised Fine-Tuning on Model Knowledge from Token and Parameter Levels

Junjie Ye[1*], **Yuming Yang**[1*], **Yang Nan**[1], **Shuo Li**[1], **Qi Zhang**[1,3], **Tao Gui**[1,3,4†], **Xuanjing Huang**[1,3], **Peng Wang**[2], **Zhongchao Shi**[2], **Jianping Fan**[2]

[1]Fudan University [2]Lenovo Research, Beijing, China
[3]Shanghai Key Lab of Intelligent Information Processing
[4]Shanghai Innovation Institute

jjye23@m.fudan.edu.cn, tgui@fudan.edu.cn

Abstract

Large language models (LLMs) acquire substantial world knowledge during pre-training, which is further shaped by post-training techniques such as supervised fine-tuning (SFT). However, the impact of SFT on a model's knowledge remains underexplored, limiting our ability to control knowledge change behavior in fine-tuned models. To address this gap, we evaluate closed-book question answering (CBQA) performance across five LLMs from the LLaMA-2 and LLaMA-3 families. Surprisingly, models fine-tuned on 1,920 samples perform up to 14% worse than those fine-tuned on only 240 samples. Furthermore, varying the level of knowledge mastery in the fine-tuning data leads to performance fluctuations of over 12%. To investigate these effects, we analyze model behavior at both the token and parameter levels. Our analysis reveals that up to 90% of parameter updates during SFT do not contribute to knowledge enhancement. Restoring these updates can improve performance on the CBQA task, depending on the characteristics of the fine-tuning data. These insights offer practical guidance for developing fine-tuning strategies that more effectively strengthen model knowledge.

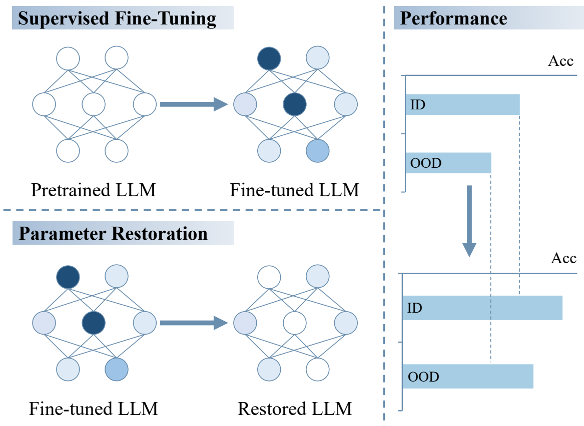


Figure 1: Illustration of parameter restoration. We find that SFT introduces many unnecessary parameter updates, and model performance can be significantly improved by restoring some of the most updated parameters in the fine-tuned model to their original values in the pre-trained model.

1 Introduction

Large language models (LLMs) (Bai et al., 2022b; OpenAI, 2023; Team, 2024; Yang et al., 2024a) acquire extensive world knowledge through pre-training on massive text corpora (Chen et al., 2023; Ye et al., 2023). This knowledge is subsequently shaped through post-training techniques such as supervised fine-tuning (SFT) (Yang et al., 2024b) and reinforcement learning (Bai et al., 2022a), enabling LLMs to perform diverse downstream tasks, including reading comprehension (Samuel et al., 2024), code generation (Rozière et al., 2023), and tool use (Ye et al., 2024a,b).

Recent research has explored how model knowledge evolves during training. For instance, pre-training has been shown to encode knowledge modularly (Wang et al., 2024), with each parameter storing up to 2 bits of information (Allen-Zhu and Li, 2025). Conversely, instruction fine-tuning may increase hallucinations (Gekhman et al., 2024; Ghosh et al., 2024). Empirical evidence suggests that preserving the distribution of internal representations is crucial to maintaining performance (Ren et al., 2024), and models with richer knowledge can be easier to fine-tune for enhanced reasoning ability (Ye et al., 2025).

Despite these insights, the specific impact of SFT on model knowledge remains insufficiently understood. Key open questions include how model knowledge changes with different categories and scales of fine-tuning data, the mechanisms behind these changes, and strategies to mitigate

*Equal Contribution.
†Corresponding Author.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 471–513
November 4-9, 2025

undesirable effects. This gap limits our ability to predict and control knowledge change behavior in fine-tuned models.

To address this, we evaluate five LLMs from the LLaMA-2 and LLaMA-3 families on the closed-book question answering (CBQA) task. We categorize fine-tuning data into five groups based on the knowledge mastery level and systematically examine how performance varies across these categories and different data scales. Surprisingly, models fine-tuned with 1,920 samples perform up to 14% worse than those fine-tuned with only 240 samples. Moreover, performance fluctuates by over 12% depending on the data category used.

To investigate these discrepancies, we conduct a token-level analysis by computing the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951) between token logits of fine-tuned and pre-trained models (Section 4). Our results show that as fine-tuning data size increases, KL divergence initially decreases, reflecting reduced deviation from the pre-trained model. However, beyond a threshold, KL divergence sharply rises, especially when fine-tuning on poorly mastered data, correlating with performance degradation.

Building on these findings, we perform a parameter-level analysis (Section 5) by selectively restoring parameters that changed most during SFT back to their pre-trained values (Figure 1). We observe that restoring up to 90% of parameter updates does not harm and can even improve performance on training and test sets, with improvements exceeding 10% in some cases. This indicates that many SFT-induced updates are unnecessary for knowledge enhancement, suggesting new directions for optimizing fine-tuning.

In summary, our contributions are as follows:

- We conduct extensive experiments on the CBQA task and reveal surprising effects of fine-tuning data category and scale on model knowledge.
- Through token-level and parameter-level analyses, we find that 90% of the parameter updates from fine-tuning do not contribute to knowledge enhancement.
- We demonstrate that restoring these parameters improves model performance, offering practical guidance for more effective fine-tuning strategies.

2 Related Work

CBQA and Model Knowledge The CBQA task evaluates an LLM's ability to answer factual questions using its internal knowledge, without relying on external reference materials (Zhang et al., 2024; Wen et al., 2024; Sticha et al., 2024). This makes CBQA a rigorous test of the model's knowledge accuracy and completeness. One significant challenge in CBQA is addressing hallucinations-instances where the model generates incorrect or fabricated answers (Huang et al., 2023; Kandpal et al., 2023; Kang and Choi, 2023). To mitigate hallucinations and enhance performance, several strategies have been proposed. For instance, Ren et al. (2024) investigate the impact of fine-tuning on the consistency of a model's pre-existing knowledge, emphasizing the need for stable knowledge retention during fine-tuning. Similarly, Gekhman et al. (2024) identify overfitting to fine-tuning data as a major source of hallucinations, noting that fine-tuning with data unfamiliar to the model exacerbates this issue. Additionally, Ye et al. (2024c) examine how variations in dataset size and quality influence CBQA outcomes, highlighting the trade-offs between data volume and model performance. Despite these advances, prior studies primarily focus on dataset characteristics and overlook the fine-tuning process's internal dynamics. In contrast, our work provides a detailed analysis at both the token and parameter levels, identifying unnecessary parameter updates during fine-tuning as a key factor contributing to performance degradation on CBQA.

Data Quality and Scale of SFT SFT plays a pivotal role in adapting LLMs to labeled data, enabling strong performance on downstream tasks. Consequently, constructing high-quality fine-tuning datasets is critical for maximizing SFT's effectiveness (Muennighoff et al., 2023; Lin et al., 2024; Ma et al., 2024). Recent research highlights the effectiveness of SFT with small, high-quality datasets, achieving performance on par with larger datasets (Zhou et al., 2023; Yang et al., 2025b). High-quality data is typically characterized as accurate, diverse, and complex (Huang et al., 2024; Liu et al., 2024; Ye et al., 2024d; Yang et al., 2025a), prompting efforts to synthesize such datasets automatically (Xu et al., 2023, 2024; Zhu et al., 2024). Concurrently, studies show that scaling the quantity of fine-tuning data, while maintaining quality, can yield further performance improvements (Kaplan et al., 2020; Chung et al., 2022; Wei et al., 2022; Dong et al., 2024). While prior work has explored dataset quality and size, few studies have examined how a model's prior knowledge of fine-tuning data influences performance or how different data quantities affect the model's knowledge. Our study differs by investigating SFT performance on the CBQA task, focusing on how mastery levels and data scale impact model knowledge.

3 Experiments

To explore how SFT affects the factual knowledge of LLMs in the CBQA setting, we conduct a series of controlled experiments. In this section, we outline the datasets used (Section 3.1), the models tested (Section 3.2), and the experimental setup (Section 3.3), followed by a presentation of the results and a summary of our findings (Section 3.4).

3.1 Dataset

Following Gekhman et al. (2024) and Ye et al. (2024c), we use the ENTITYQUESTIONS (Sciavolino et al., 2021) to construct the training and testing datasets for our experiments, which is a CBQA-specific dataset containing knowledge across 24 topics extracted from Wikipedia.

Training Data Our training dataset, denoted as $\mathcal{D}_{train}$, consists of data on 10 location-related topics extracted from the original training corpus. Following the method of Ye et al. (2024c), we classify the training samples based on the pre-trained model $\mathcal{M}$'s mastery level on the knowledge associated with each data point k. Specifically, we enhance the multi-template completion mechanism of Ye et al. (2024c) to allow $\mathcal{M}$ to complete each data point k using multiple templates. The training data is then divided into five categories according to the proportion $R_k^{\mathcal{M}}$ of knowledge points correctly completed.[1] Formally:

$$\mathcal{D}^{\mathcal{M}}_{train-i} = \begin{cases} \{k \in \mathcal{D}_{train} \mid R_k^{\mathcal{M}} = 0\}, \\ \qquad\qquad i = 0, \\ \{k \in \mathcal{D}_{train} \mid R_k^{\mathcal{M}} \in (\frac{i-1}{4}, \frac{i}{4}]\}, \\ \qquad\qquad i \in \{1,2,3,4\}. \end{cases}$$

Testing Data For the in-domain testing dataset $\mathcal{D}_{test}$, we select data from the same 10 location-related topics in the original test set. Data from the remaining 14 topics are used as the out-of-domain testing dataset $\mathcal{D}_{testood}$. Similar to the training data, both $\mathcal{D}_{test}$ and $\mathcal{D}_{testood}$ are categorized as:

$$\mathcal{D}_{test} = \bigcup_{i=0}^{4} \mathcal{D}^{\mathcal{M}}_{test-i},\ \mathcal{D}_{testood} = \bigcup_{i=0}^{4} \mathcal{D}^{\mathcal{M}}_{testood-i}$$

An example of data distribution is listed in Table 1.[2]

$\mathcal{D}_{train}$	$\mathcal{D}^{\mathcal{M}}_{train-0}$	$\mathcal{D}^{\mathcal{M}}_{train-1}$	$\mathcal{D}^{\mathcal{M}}_{train-2}$	$\mathcal{D}^{\mathcal{M}}_{train-3}$	$\mathcal{D}^{\mathcal{M}}_{train-4}$
Number	18456	29571	11558	8923	7436
$\mathcal{D}_{test}$	$\mathcal{D}^{\mathcal{M}}_{test-0}$	$\mathcal{D}^{\mathcal{M}}_{test-1}$	$\mathcal{D}^{\mathcal{M}}_{test-2}$	$\mathcal{D}^{\mathcal{M}}_{test-3}$	$\mathcal{D}^{\mathcal{M}}_{test-4}$
Number	2383	3664	1484	1109	915
$\mathcal{D}_{testood}$	$\mathcal{D}^{\mathcal{M}}_{testood-0}$	$\mathcal{D}^{\mathcal{M}}_{testood-1}$	$\mathcal{D}^{\mathcal{M}}_{testood-2}$	$\mathcal{D}^{\mathcal{M}}_{testood-3}$	$\mathcal{D}^{\mathcal{M}}_{testood-4}$
Number	4127	4539	1271	1120	556

Table 1: An example of data distribution, where $\mathcal{M}$ refers to LLaMA-3-8B.

3.2 Models

Given the dominance of decoder-only architectures in current LLMs, our analysis focuses exclusively on models of this type. We examine five LLMs from two model families: **LLaMA-2-7B**, **LLaMA-2-13B**, and **LLaMA-2-70B** from the LLaMA-2 family (Touvron et al., 2023), and **LLaMA-3-8B** and **LLaMA-3-70B** from the LLaMA-3 family (Dubey et al., 2024).[3]

3.3 Experimental Setup

Our experiment involves data categorization, training, and testing, aimed at evaluating model performance under diverse settings.

Data Categorization To balance the stability and diversity of the generated output, we design 21 mapping templates tailored to each topic's data. The sampling temperature is set to 0.7 to introduce controlled randomness, and each prompt is sampled 10 times to enhance robustness. The output's maximum token length is limited to 32.

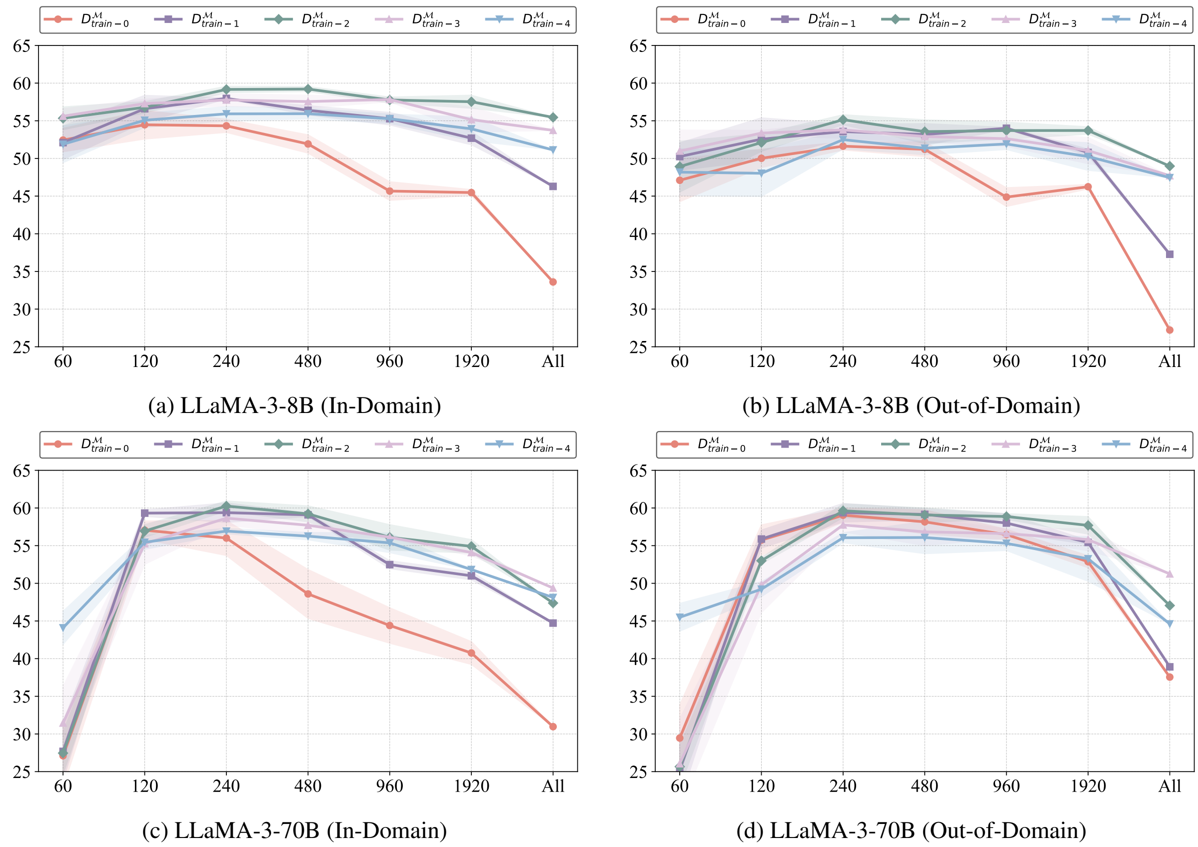
Training Training is conducted using a batch size of 8 over 1 epoch, employing the AdamW (Loshchilov and Hutter, 2019) optimizer with cosine learning rate scheduling for stable and efficient convergence. The learning rate is set to 1×10^{-5}.[4]

[1] For additional details on data processing, see Appendix F.

[2] Data distribution of other LLMs can be found in Appendix D.

[3] Details of models can be found in Appendix B.1.

[4] To ensure a fair comparison, we use uniform prompt templates during training, as detailed in Appendix A.



(a) LLaMA-3-8B (In-Domain)

(b) LLaMA-3-8B (Out-of-Domain)

(c) LLaMA-3-70B (In-Domain)

(d) LLaMA-3-70B (Out-of-Domain)

Figure 2: In-domain ($\mathbf{Acc}^{\mathcal{M}}_{test}$) and out-of-domain ($\mathbf{Acc}^{\mathcal{M}}_{testood}$) performance of the LLaMA-3 family models fine-tuned with varying data scales, where 'All' indicates the use of the entire dataset listed in Appendix D.

Testing For testing, we utilize a greedy decoding strategy with a maximum output length of 16, maintaining consistency with the prompt templates used during training. To mitigate bias from the training data selection, we generate five distinct training datasets by random sampling. Each experiment is repeated using these datasets, and the final results are reported as the mean and variance across the five runs. Evaluation metrics include accuracy, categorized by different knowledge mastery levels, with the mean accuracy across all test sets serving as the final metric:

$$\mathbf{Acc}^{\mathcal{M}}_{test} = \sum_{i=0}^{4} \mathbf{Acc}^{\mathcal{M}}_{test-i}/5$$

$$\mathbf{Acc}^{\mathcal{M}}_{testood} = \sum_{i=0}^{4} \mathbf{Acc}^{\mathcal{M}}_{testood-i}/5$$

3.4 Main Results

We fine-tune each of the five selected LLMs using datasets with five different mastery levels. To conduct a more detailed analysis, we compare changes in model performance across varying data scales. To enhance robustness, we ensure a balanced data distribution across topics and repeat each experiment three times. Figure 2 presents the in-domain and out-of-domain test results for the LLaMA-3 family of models.[5] From the results, we observe two unexpected phenomena.

Phenomenon 1 *Regardless of the type of training data used, LLMs achieve their optimal performance with just 240 data points. Adding more training data beyond this point risks degrading model performance.*

Our analysis reveals that model performance improves as the amount of fine-tuned data increases from 60 to 240 entries, aligning with the general expectation that more data enhances performance. However, performance peaks at **only 240 entries**, and adding additional fine-tuned data not only fails to yield further improvements but often leads to a significant decline. For instance, when fine-tuned with barely mastered data (i.e., $\mathcal{D}^{\mathcal{M}}_{train-0}$), LLaMA-3-8B achieves an $\mathbf{Acc}^{\mathcal{M}}_{test}$ score that is 8.86% lower

[5]Test results for the LLaMA-2 family of models can be found in Appendix C.1.

Source	In-Domain						Out-of-Domain					
	$\mathbf{Acc}^{\mathcal{M}}_{test-0}$	$\mathbf{Acc}^{\mathcal{M}}_{test-1}$	$\mathbf{Acc}^{\mathcal{M}}_{test-2}$	$\mathbf{Acc}^{\mathcal{M}}_{test-3}$	$\mathbf{Acc}^{\mathcal{M}}_{test-4}$	$\mathbf{Acc}^{\mathcal{M}}_{test}$	$\mathbf{Acc}^{\mathcal{M}}_{testood-0}$	$\mathbf{Acc}^{\mathcal{M}}_{testood-1}$	$\mathbf{Acc}^{\mathcal{M}}_{testood-2}$	$\mathbf{Acc}^{\mathcal{M}}_{testood-3}$	$\mathbf{Acc}^{\mathcal{M}}_{testood-4}$	$\mathbf{Acc}^{\mathcal{M}}_{testood}$
$\mathcal{M}$ = *LLaMA-3-8B*												
$\mathcal{D}^{\mathcal{M}}_{train-0}$	$\mathbf{1.75}_{0.17}$	$16.07_{0.67}$	$55.03_{1.39}$	$71.06_{1.09}$	$83.46_{1.23}$	$45.47_{0.40}$	$\mathbf{1.91}_{0.33}$	$15.89_{1.20}$	$59.01_{0.51}$	$74.08_{0.63}$	$80.33_{0.98}$	$46.24_{0.29}$
$\mathcal{D}^{\mathcal{M}}_{train-1}$	$0.98_{0.14}$	$\mathbf{40.12}_{0.74}$	$63.93_{0.55}$	$74.19_{0.73}$	$84.22_{3.96}$	$52.69_{0.88}$	$1.66_{0.09}$	$23.88_{0.45}$	$65.03_{0.77}$	$79.63_{0.63}$	$83.84_{0.55}$	$50.80_{0.45}$
$\mathcal{D}^{\mathcal{M}}_{train-2}$	$0.78_{0.03}$	$36.56_{0.53}$	$\mathbf{75.61}_{1.18}$	$83.98_{1.37}$	$90.71_{1.31}$	$\mathbf{57.53}_{0.86}$	$1.45_{0.35}$	$\mathbf{25.02}_{0.30}$	$\mathbf{70.52}_{1.59}$	$\mathbf{83.66}_{0.67}$	$87.89_{0.45}$	$\mathbf{53.71}_{0.49}$
$\mathcal{D}^{\mathcal{M}}_{train-3}$	$0.64_{0.15}$	$27.20_{3.69}$	$70.33_{1.73}$	$\mathbf{85.90}_{1.47}$	$91.66_{1.57}$	$55.15_{1.64}$	$1.39_{0.34}$	$21.66_{3.13}$	$63.91_{2.70}$	$81.34_{0.93}$	$86.87_{1.85}$	$51.04_{1.73}$
$\mathcal{D}^{\mathcal{M}}_{train-4}$	$0.64_{0.06}$	$24.26_{3.38}$	$68.28_{2.00}$	$83.29_{1.23}$	$\mathbf{93.19}_{1.91}$	$53.93_{1.56}$	$0.93_{0.11}$	$17.72_{1.33}$	$63.64_{4.39}$	$80.55_{2.05}$	$\mathbf{88.43}_{1.47}$	$50.25_{1.83}$
$\mathcal{M}$ = *LLaMA-3-70B*												
$\mathcal{D}^{\mathcal{M}}_{train-0}$	$\mathbf{3.72}_{0.33}$	$22.68_{1.53}$	$47.28_{1.26}$	$57.97_{2.25}$	$72.08_{3.20}$	$40.75_{1.51}$	$\mathbf{3.08}_{0.39}$	$25.90_{1.59}$	$67.04_{1.63}$	$82.61_{0.95}$	$85.74_{1.30}$	$52.87_{0.79}$
$\mathcal{D}^{\mathcal{M}}_{train-1}$	$1.94_{0.11}$	$\mathbf{43.85}_{0.29}$	$63.45_{1.47}$	$66.22_{1.66}$	$79.54_{0.65}$	$51.00_{0.53}$	$2.61_{0.45}$	$31.01_{0.79}$	$72.63_{0.16}$	$84.69_{0.30}$	$86.22_{0.69}$	$55.43_{0.26}$
$\mathcal{D}^{\mathcal{M}}_{train-2}$	$1.23_{0.07}$	$38.17_{1.78}$	$\mathbf{71.68}_{0.82}$	$77.58_{1.27}$	$85.89_{1.44}$	$\mathbf{54.91}_{0.89}$	$2.06_{0.50}$	$\mathbf{31.26}_{2.10}$	$\mathbf{74.51}_{1.27}$	$88.63_{0.97}$	$92.01_{1.19}$	$\mathbf{57.69}_{1.16}$
$\mathcal{D}^{\mathcal{M}}_{train-3}$	$1.00_{0.11}$	$31.52_{0.61}$	$68.32_{0.30}$	$\mathbf{81.11}_{0.73}$	$88.49_{1.60}$	$54.09_{0.45}$	$1.91_{0.79}$	$26.70_{1.71}$	$69.60_{2.77}$	$\mathbf{89.61}_{1.44}$	$91.22_{1.39}$	$55.81_{1.47}$
$\mathcal{D}^{\mathcal{M}}_{train-4}$	$0.90_{0.05}$	$26.16_{1.45}$	$64.27_{0.75}$	$78.00_{0.43}$	$\mathbf{89.83}_{0.77}$	$51.83_{0.05}$	$0.81_{0.35}$	$21.80_{3.65}$	$66.52_{5.65}$	$84.85_{2.57}$	$\mathbf{92.29}_{2.63}$	$53.25_{2.97}$

Table 2: Performance of the fine-tuned LLaMA-3 family models on in-domain and out-of-domain test sets, using 1920 data points with varying levels of mastery.

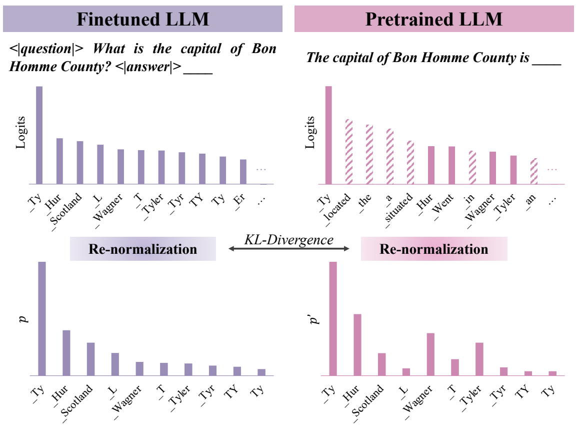


Figure 3: Illustration of logits re-normalization. Since the pre-trained LLM tends to assign high probabilities to common dummy words, we identify the ten highest logits in the fine-tuned LLM and extract the corresponding values from the pre-trained LLM. After re-normalization, we compute the KL divergence to quantify the distributional difference.

when trained with 1,920 entries compared to 240 entries. A decline of 13.69% is even observed when comparing 240 entries from $\mathcal{D}^{\mathcal{M}}_{train-2}$. Notably, when LLMs are trained with the full dataset for each data category, their performance on the CBQA task is nearly at its lowest across all data categories. This striking finding suggests that increasing the volume of fine-tuned data does not necessarily enhance model knowledge and may impair it.

Phenomenon 2 *When the amount of fine-tuned data reaches a certain threshold (e.g., 1,920 entries), model performance varies significantly based on the knowledge mastery level of the training data.*

While model performance generally declines when the fine-tuned data exceeds 240 entries, the rate of decline differs depending on the knowledge mastery level of the training data. Notably, models fine-tuned with data from $\mathcal{D}^{\mathcal{M}}_{train-0}$ exhibit a steeper performance drop compared to those trained on other data types. For instance, when fine-tuned with 1,920 entries, the $\mathbf{Acc}^{\mathcal{M}}_{test}$ difference between LLaMA-3-8B models trained on $\mathcal{D}^{\mathcal{M}}_{train-0}$ and $\mathcal{D}^{\mathcal{M}}_{train-2}$ reaches 12.06%, which is 1.50 times the difference observed with only 240 training entries. Table 2 illustrates the performance of LLaMA-3 family models across various test sets when fine-tuned with 1,920 entries from different categories. The results show that models trained on $\mathcal{D}^{\mathcal{M}}_{train-0}$ experience substantial performance degradation on test sets other than $\mathcal{D}^{\mathcal{M}}_{test-0}$. More generally, training on low-mastery data significantly impairs performance on high-mastery test data. Conversely, training on high-mastery data (e.g., $\mathcal{D}^{\mathcal{M}}_{train-4}$) leads to suboptimal performance on low-mastery test data. Training with mid-level mastery data, such as $\mathcal{D}^{\mathcal{M}}_{train-2}$, strikes a better balance, yielding superior overall performance.

4 Token-Level Analysis

To explain the performance variation observed across fine-tuned LLMs, we analyze how fine-tuning alters token-level output distributions compared to the pre-trained model. Specifically, we compute the divergence in predicted token distributions between fine-tuned and pre-trained models using KL divergence (Section 4.1). This token-level analysis reveals some interesting findings (Section 4.2).

4.1 KL Divergence Computation

Given the performance degradation observed in Section 3.4, we investigate the underlying token

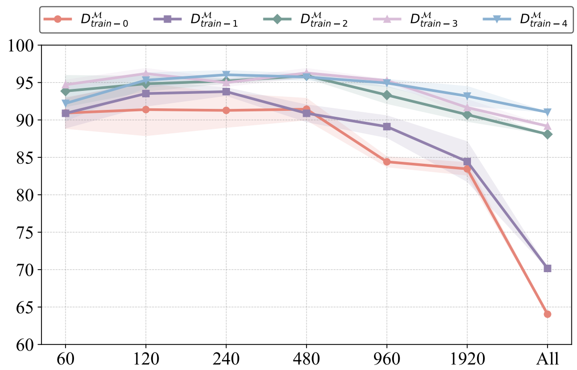


Figure 4: Performance on $\mathcal{D}^{\mathcal{M}}_{test-4}$ ($\mathbf{Acc}^{\mathcal{M}}_{test-4}$) of LLMs fine-tuned on LLaMA-3-8B.

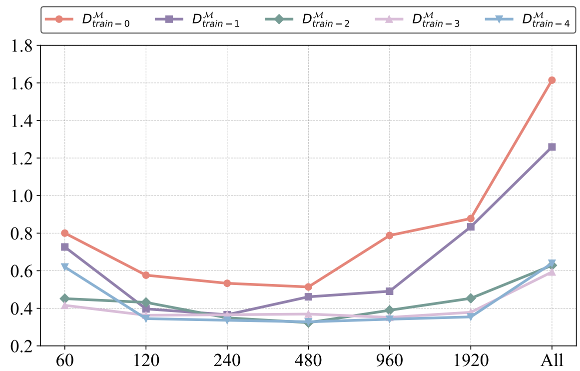


Figure 5: KL divergence of logits distribution between LLaMA-3-8B fine-tuned with different datasets and the pre-trained one.

distribution shifts caused by SFT. Specifically, we use KL divergence to quantify the differences in token probabilities between fine-tuned and pre-trained models. A higher KL divergence suggests a more significant shift in the model's token probability distribution.

Data Selection Given that the pre-trained model is used to complement the prior text, the quality of its completions depends on both the input prompt and the structure of the mapping template, as outlined in Section 3.3. The selection of appropriate data is critical to ensuring the robustness of the results. For $\mathcal{D}^{\mathcal{M}}_{test-4}$, we observe that the pre-trained model's completion success rate exceeds 75% across multiple samples and templates, suggesting that this dataset is relatively insensitive to variations in the mapping template. In contrast, other datasets are more sensitive to such variations, so our comparison of different LLMs in this section is limited to $\mathcal{D}^{\mathcal{M}}_{test-4}$. For each topic, we select the mapping template yielding the highest success rate across samples and focus our analysis on tokens in completions where the answers appear near the beginning of the generated text.

Logits Re-normalization Our goal is to compute the KL divergence between the logits distributions for the first token predicted by both the fine-tuned and pre-trained LLMs. However, as shown in Figure 3, the pre-trained model tends to assign higher probabilities to common dummy words (e.g., 'the', 'a', etc.), whereas fine-tuned models typically reduce the likelihood of these words in favor of more relevant tokens. If we directly compute the KL divergence on the raw logits, these dummy words could distort the results and obscure meaningful differences between the models. To mitigate this issue, we introduce a logits re-normalization procedure. Specifically, we sort the logits predicted by the fine-tuned model and extract the top 10 values, denoted as $l_0, l_1, \ldots, l_9$. We then identify the corresponding logits, $l'_0, l'_1, \ldots, l'_9$, from the pre-trained model's completions. Moreover, we apply the softmax function to these logits to derive their normalized probabilities, respectively:

$$p_i = \mathrm{Softmax}(l_i),\ p'_i = \mathrm{Softmax}(l'_i).$$

After completing the logits re-normalization, we compute the KL divergence between the probability distributions p and p' for the fine-tuned and pre-trained models as follows:

$$s_{\mathrm{KL}}(p \parallel p') = -\sum_i p_i \log \frac{p'_i}{p_i}.$$

4.2 Results Analysis

We analyze the performance of individual LLMs fine-tuned based on LLaMA-3-8B, presenting their results on $\mathcal{D}^{\mathcal{M}}_{test-4}$ in Figure 4 and their KL divergence relative to the pre-trained model's distribution in Figure 5. From these results, we derive two key findings.

Finding 1 *Regardless of the category of fine-tuning data, the difference in predicted logits distributions between the fine-tuned and pre-trained models initially decreases and then increases as the amount of data grows.*

Figure 5 illustrates how the predicted logits distributions of fine-tuned model diverge from the pre-trained model as training data increases. When fine-tuning with a small dataset (e.g., 60 samples), the logits distribution shifts significantly due to insufficient data, leading to unstable training. As the dataset grows (e.g., 240 samples), this discrepancy decreases, indicating improved stability. However, with further increases, the difference in logits distributions grows again, particularly for models trained on $\mathcal{D}^{\mathcal{M}}_{train-0}$ and $\mathcal{D}^{\mathcal{M}}_{train-1}$. This suggests that as training data increases, the model deviates further from its pre-trained knowledge. The effect is more pronounced when fine-tuning on low-mastery data, making the model more susceptible to knowledge shifts.

Finding 2 *As the difference in the predicted logits distribution between the fine-tuned model and the pre-trained model increases, model performance declines, indicating a negative impact of excessive knowledge shifts.*

Figure 4 and Figure 5 reveal a strong correlation between performance degradation on $\mathcal{D}^{\mathcal{M}}_{test-4}$ and increasing divergence in logits distributions. Since $\mathcal{D}^{\mathcal{M}}_{test-4}$ contains samples well mastered by the pre-trained model, substantial shifts in learned knowledge during fine-tuning can lead to catastrophic forgetting, where previously acquired knowledge is lost, thereby degrading performance. This effect is particularly evident when training with large datasets. For instance, the model fine-tuned on $\mathcal{D}^{\mathcal{M}}_{train-0}$ experiences the most significant knowledge shift and performs the worst among all fine-tuned models. Since changes in logits distribution reflect underlying modifications to model parameters, we hypothesize that **excessive parameter updates during fine-tuning, especially when using large or low-mastery datasets, lead to overall performance decline**.

5 Parameter-Level Analysis

The observations and analyses in Section 4 indicate that excessive parameter updates can degrade model performance. To further investigate this, we analyze the impact at the parameter level by progressively restoring the updated parameters and examining the resulting performance changes (Section 5.1). Our findings indicate that a significant proportion of parameter updates during SFT do not contribute to performance improvement and may even be detrimental (Section 5.2).

5.1 Parameter Restoration

To examine the impact of excessive parameter updates on model performance, we design an experimental framework for parameter restoration.

Proportion	1%	3%	5%	10%	20%	40%	60%
Number of Training Data: 240							
$\mathcal{D}^{\mathcal{M}}_{train-0}$	70.59%	78.82%	82.35%	87.06%	91.76%	96.47%	99.12%
$\mathcal{D}^{\mathcal{M}}_{train-1}$	71.01%	79.29%	82.84%	87.57%	92.31%	97.04%	99.11%
$\mathcal{D}^{\mathcal{M}}_{train-2}$	71.13%	79.17%	82.74%	87.50%	92.26%	96.43%	99.12%
$\mathcal{D}^{\mathcal{M}}_{train-3}$	70.72%	78.97%	82.51%	87.22%	91.93%	96.65%	99.09%
$\mathcal{D}^{\mathcal{M}}_{train-4}$	70.98%	78.74%	82.18%	87.36%	91.95%	96.55%	99.04%
Number of Training Data: 1920							
$\mathcal{D}^{\mathcal{M}}_{train-0}$	70.56%	78.50%	82.24%	86.92%	92.06%	96.26%	98.69%
$\mathcal{D}^{\mathcal{M}}_{train-1}$	70.89%	78.87%	82.63%	87.32%	92.02%	96.71%	98.69%
$\mathcal{D}^{\mathcal{M}}_{train-2}$	70.75%	78.77%	82.08%	87.26%	91.98%	96.70%	98.70%
$\mathcal{D}^{\mathcal{M}}_{train-3}$	70.74%	78.70%	81.98%	87.13%	91.82%	96.50%	98.70%
$\mathcal{D}^{\mathcal{M}}_{train-4}$	70.83%	78.70%	82.41%	87.04%	92.13%	96.30%	98.70%

Table 3: Percentage of total parameter updates concentrated in different proportions of the most highly updated parameters in various LLMs fine-tuned on LLaMA-3-8B.

Specifically, we compare the fine-tuned model with the pre-trained model, sorted by the rate of parameter change.[6] Table 3 reports the percentage of total parameter updates attributed to different proportions of the most highly updated parameters in LLMs fine-tuned on LLaMA-3-8B. The results indicate that parameter updates are heavily concentrated in a small subset of parameters. For instance, more than 70% of the total updates occur in fewer than 1% of the parameters. Following this, we progressively restore the most significantly updated parameters to their original values in the pre-trained model, starting with the largest updates and gradually including smaller ones, while monitoring the corresponding changes in model performance. This process is illustrated in Figure 1.

5.2 Results Analysis

We evaluate the performance of LLaMA-3-8B after restoring different proportions of parameters across various fine-tuning datasets. The results are summarized in Table 4. Our analysis of these results reveals several noteworthy findings.

Finding 1 *The majority of parameter updates introduced by SFT are unnecessary and can significantly degrade model knowledge.*[7]

Table 4 shows that restoring a portion of the model's parameters to their pre-trained values consistently improves performance, regardless of the fine-tuning dataset. For instance, when fine-tuning with 1,920 samples, restoring 20% of the parameters enhances the performance of all models. Specifically, the model fine-tuned with $\mathcal{D}^{\mathcal{M}}_{train-0}$ achieves a 9.85% performance gain. Table 3

[6] Specific calculation details can be found in Appendix B.2.
[7] More discussion can be found in Appendix E.

Restore	$\mathcal{D}^{\mathcal{M}}_{\text{train}-0}$	$\mathcal{D}^{\mathcal{M}}_{\text{train}-1}$	$\mathcal{D}^{\mathcal{M}}_{\text{train}-2}$	$\mathcal{D}^{\mathcal{M}}_{\text{train}-3}$	$\mathcal{D}^{\mathcal{M}}_{\text{train}-4}$
Number of Training Data: 240					
0	55.33	57.96	59.32	59.12	53.97
1%	55.76	58.17	59.62	59.24	54.30
3%	56.64	58.52	59.77	59.40	54.31
5%	57.22	58.68	59.89	59.63	54.44
10%	58.32	59.45	60.40	59.83	54.69
20%	59.07	59.81	59.88	59.91	46.45
40%	59.77	33.40	42.44	11.20	23.83
60%	1.68	2.20	3.65	2.56	1.65
Number of Training Data: 1920					
0	44.96	52.43	58.80	57.70	55.22
1%	46.73	53.72	59.85	58.68	55.88
3%	48.53	55.01	60.56	59.23	56.76
5%	49.85	55.96	61.10	59.65	57.34
10%	52.10	57.14	61.67	60.02	58.24
20%	54.81	58.33	62.21	58.93	58.66
40%	55.44	22.06	59.97	6.92	56.50
60%	1.48	1.12	1.62	0.51	0.60

(a) In-Domain ($\mathbf{Acc}^{\mathcal{M}}_{test}$)

Restore	$\mathcal{D}^{\mathcal{M}}_{\text{train}-0}$	$\mathcal{D}^{\mathcal{M}}_{\text{train}-1}$	$\mathcal{D}^{\mathcal{M}}_{\text{train}-2}$	$\mathcal{D}^{\mathcal{M}}_{\text{train}-3}$	$\mathcal{D}^{\mathcal{M}}_{\text{train}-4}$
Number of Training Data: 240					
0	52.37	51.70	55.35	55.23	50.69
1%	52.62	52.39	56.45	56.17	50.82
3%	53.03	52.82	56.47	56.41	50.74
5%	53.27	53.09	56.80	56.56	50.59
10%	53.44	53.87	56.46	56.72	49.71
20%	54.18	54.36	55.95	55.52	43.13
40%	53.79	20.77	45.49	17.56	31.19
60%	0.20	0.22	0.32	0.20	0.23
Number of Training Data: 1920					
0	49.40	52.38	54.04	53.79	51.70
1%	50.78	54.20	55.17	54.75	52.62
3%	52.03	55.12	56.00	55.52	53.35
5%	52.54	55.12	56.34	55.84	53.77
10%	53.42	55.08	56.68	55.54	54.32
20%	54.50	53.91	57.10	52.23	53.82
40%	53.64	20.51	53.84	9.67	50.17
60%	0.30	0.10	0.27	0.07	0.18

(b) Out-of-Domain ($\mathbf{Acc}^{\mathcal{M}}_{testood}$)

Table 4: Performance of LLaMA-3-8B after restoring different scales of parameters across various fine-tuning datasets. Improvements over the non-restored model are highlighted in green, while performance declines are shown in red, with darker shades indicating larger differences.

Restore	XSum			GSM8K
	ROUGE-1	ROUGE-2	ROUGE-L	ACC
0	42.57	19.50	34.55	57.69
1%	42.50	19.71	34.67	57.69
3%	**42.63**	**19.78**	**34.75**	57.75
5%	42.36	19.47	34.44	58.49
10%	42.57	19.40	34.60	**59.60**
20%	41.31	18.59	33.51	58.72
40%	15.59	4.15	12.09	0
60%	0	0	0	0

Table 5: Performance of LLaMA-3-8B after restoring different scales of parameters on XSum (Narayan et al., 2018) (Summarization) and GSM8K (Cobbe et al., 2021) (Math).

further reveals that over 90% of the total parameter variation is restored at this point. Importantly, the benefits of parameter restoration generalize across tasks, as shown in Table 5. However, the degree of improvement depends on the relevance of the task to the model's knowledge. Notably, performance on the training set also improves, suggesting that *many of the parameter updates introduced by SFT neither help fit the training data nor support generalization, and may impair previously learned knowledge*. Compared to other strategies, restoring redundant parameter updates is an effective and simple method for enhancing model performance, offering useful insights for designing more efficient fine-tuning approaches.[8]

[8]A comparison of different strategies is presented in Appendix C.3.

Finding 2 *Models fine-tuned with larger datasets or lower-mastery data are more adversely affected by unnecessary parameter changes during SFT.*

While SFT consistently introduces unnecessary parameter updates that degrade model performance, the extent of this effect depends on the scale and category of fine-tuning data. On one hand, models fine-tuned with larger datasets experience a greater impact. Specifically, models trained with 240 samples generally show performance degradation when more than 20% of the parameters are restored. In contrast, models fine-tuned with 1,920 samples continue to gain performance improvements even after restoring 40% of the parameters. This suggests that fine-tuning with 1,920 samples introduces a higher proportion of unnecessary updates. Additionally, the maximum performance gain achieved through parameter restoration is greater for models fine-tuned with 1,920 samples than for those fine-tuned with 240 samples. On the other hand, models fine-tuned with low-mastery data are also more affected. Regardless of dataset size, models fine-tuned with $\mathcal{D}^{\mathcal{M}}_{train-0}$ consistently allow more parameter restoration while achieving greater performance gains compared to other models. For instance, when using 1,920 samples, the model fine-tuned with $\mathcal{D}^{\mathcal{M}}_{train-0}$ can restore 40% of the parameters and achieve a 10.48% performance gain, whereas the model fine-tuned with $\mathcal{D}^{\mathcal{M}}_{train-4}$ achieves a maximum gain of only 3.44% after restoring 20%

of the parameters.

6 Conclusion

In this paper, we conduct an in-depth analysis of five LLMs across two families on the CBQA task, revealing that both the category and scale of fine-tuning data significantly influence performance in unexpected ways. Through token-level analysis, we find that large changes in token logits correlate with degraded model performance, suggesting that excessive parameter updates can harm model knowledge. At the parameter level, we show that up to 90% of the updates made during SFT are unnecessary or even detrimental for knowledge enhancement. By selectively restoring these updates, we improve model performance while preserving prior knowledge. Our findings challenge conventional fine-tuning practices and offer practical guidance for developing more efficient methods for LLMs.

Limitations

Although we conduct an in-depth analysis of anomalies arising from SFT, our work has certain limitations. On one hand, the study does not propose a more efficient fine-tuning strategy based on the findings. This is because the focus is on phenomenological analysis to uncover the underlying mechanisms of SFT on model knowledge. Future work should focus on designing adaptive fine-tuning strategies that minimize unnecessary updates while maximizing performance gains. On the other hand, due to resource constraints, the analysis is limited to the LLaMA-2 and LLaMA-3 model series. However, preliminary validation on other model families shows that the conclusions generalize, suggesting broader applicability.

Acknowledgments

The authors wish to thank the anonymous reviewers for their helpful comments. This work was partially funded by the Science and Technology Commission of Shanghai Municipality (No.24511103100), National Natural Science Foundation of China (No.62476061,62206057), Shanghai Rising-Star Program (23QA1400200), Natural Science Foundation of Shanghai (23ZR1403500).

References

Zeyuan Allen-Zhu and Yuanzhi Li. 2025. Physics of language models: Part 3.3, knowledge capacity scaling laws. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, Benjamin Mann, and Jared Kaplan. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *CoRR*, abs/2204.05862.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosiute, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemí Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022b. Constitutional AI: harmlessness from AI feedback. *CoRR*, abs/2212.08073.

Xuanting Chen, Junjie Ye, Can Zu, Nuo Xu, Rui Zheng, Minlong Peng, Jie Zhou, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. How robust is GPT-3.5 to predecessors? A comprehensive study on language understanding tasks. *CoRR*, abs/2303.00293.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168.

Guanting Dong, Hongyi Yuan, Keming Lu, Chengpeng Li, Mingfeng Xue, Dayiheng Liu, Wei Wang, Zheng Yuan, Chang Zhou, and Jingren Zhou. 2024. How abilities in large language models are affected by supervised fine-tuning data composition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 177–198. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Jonathan Frankle and Michael Carbin. 2019. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. 2024. Does fine-tuning llms on new knowledge encourage hallucinations? *CoRR*, abs/2405.05904.

Sreyan Ghosh, Chandra Kiran Reddy Evuru, Sonal Kumar, Ramaneswaran S., Deepali Aneja, Zeyu Jin, Ramani Duraiswami, and Dinesh Manocha. 2024. A closer look at the limitations of instruction tuning. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

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 *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Hui Huang, Bing Xu, Xinnian Liang, Kehai Chen, Muyun Yang, Tiejun Zhao, and Conghui Zhu. 2024. Multi-view fusion for instruction mining of large language model. *Inf. Fusion*, 110:102480.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *CoRR*, abs/2311.05232.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 15696–15707. PMLR.

Cheongwoong Kang and Jaesik Choi. 2023. Impact of co-occurrence on factual knowledge of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 7721–7735. Association for Computational Linguistics.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *CoRR*, abs/2001.08361.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Jianzhe Lin, Maurice Diesendruck, Liang Du, and Robin Abraham. 2024. Batchprompt: Accomplish more with less. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. 2024. What makes good data for alignment? A comprehensive study of automatic data selection in instruction tuning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Yingwei Ma, Yue Liu, Yue Yu, Yuanliang Zhang, Yu Jiang, Changjian Wang, and Shanshan Li. 2024.

At which training stage does code data help llms reasoning? In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M. Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 15991–16111. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 1797–1807. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Mengjie Ren, Boxi Cao, Hongyu Lin, Cao Liu, Xianpei Han, Ke Zeng, Guanglu Wan, Xunliang Cai, and Le Sun. 2024. Learning or self-aligning? rethinking instruction fine-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 6090–6105. Association for Computational Linguistics.

Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton-Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre Défossez, Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier, Thomas Scialom, and Gabriel Synnaeve. 2023. Code llama: Open foundation models for code. *CoRR*, abs/2308.12950.

Vinay Samuel, Houda Aynaou, Arijit Ghosh Chowdhury, Karthik Venkat Ramanan, and Aman Chadha. 2024. Can llms augment low-resource reading comprehension datasets? opportunities and challenges. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024 - Student Research Workshop, Bangkok, Thailand, August 11-16, 2024*, pages 411–421. Association for Computational Linguistics.

Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. 2021. Simple entity-centric questions challenge dense retrievers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6138–6148. Association for Computational Linguistics.

Abigail Sticha, Norbert Braunschweiler, Rama Sanand Doddipatla, and Kate M. Knill. 2024. Advancing faithfulness of large language models in goal-oriented dialogue question answering. In *ACM Conversational User Interfaces 2024, CUI 2024, Luxembourg, July 8-10, 2024*, page 32. ACM.

Meta Team. 2024. Introducing llama 3.1: Our most capable models to date.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Mengru Wang, Yunzhi Yao, Ziwen Xu, Shuofei Qiao, Shumin Deng, Peng Wang, Xiang Chen, Jia-Chen Gu, Yong Jiang, Pengjun Xie, Fei Huang, Huajun Chen, and Ningyu Zhang. 2024. Knowledge mechanisms in large language models: A survey and perspective. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 7097–7135. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Trans. Mach. Learn. Res.*, 2022.

Zhihua Wen, Zhiliang Tian, Zexin Jian, Zhen Huang, Pei Ke, Yifu Gao, Minlie Huang, and Dongsheng Li. 2024. Perception of knowledge boundary for large language models through semi-open-ended question answering. *CoRR*, abs/2405.14383.

Canwen Xu, Daya Guo, Nan Duan, and Julian J. McAuley. 2023. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 6268–6278. Association for Computational Linguistics.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. 2024. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing. *CoRR*, abs/2406.08464.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024a. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Yuming Yang, Yang Nan, Junjie Ye, Shihan Dou, Xiao Wang, Shuo Li, Huijie Lv, Tao Gui, Qi Zhang, and Xuanjing Huang. 2025a. Measuring data diversity for instruction tuning: A systematic analysis and A reliable metric. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 18530–18549. Association for Computational Linguistics.

Yuming Yang, Wantong Zhao, Caishuang Huang, Junjie Ye, Xiao Wang, Huiyuan Zheng, Yang Nan, Yuran Wang, Xueying Xu, Kaixin Huang, Yunke Zhang, Tao Gui, Qi Zhang, and Xuanjing Huang. 2025b. Beyond boundaries: Learning a universal entity taxonomy across datasets and languages for open named entity recognition. In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pages 10902–10923. Association for Computational Linguistics.

Zhaorui Yang, Tianyu Pang, Haozhe Feng, Han Wang, Wei Chen, Minfeng Zhu, and Qian Liu. 2024b. Self-distillation bridges distribution gap in language model fine-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 1028–1043. Association for Computational Linguistics.

Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, Jie Zhou, Siming Chen, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. A comprehensive capability analysis of GPT-3 and GPT-3.5 series models. *CoRR*, abs/2303.10420.

Junjie Ye, Yilong Wu, Songyang Gao, Caishuang Huang, Sixian Li, Guanyu Li, Xiaoran Fan, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024a. Rotbench: A multi-level benchmark for evaluating the robustness of large language models in tool learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 313–333. Association for Computational Linguistics.

Junjie Ye, Yilong Wu, Sixian Li, Yuming Yang, Tao Gui, Qi Zhang, Xuanjing Huang, Peng Wang, Zhongchao Shi, Jianping Fan, and Zhengyin Du. 2024b. Tl-training: A task-feature-based framework for training large language models in tool use. *CoRR*, abs/2412.15495.

Junjie Ye, Yuming Yang, Qi Zhang, Tao Gui, Xuanjing Huang, Peng Wang, Zhongchao Shi, and Jianping Fan. 2024c. 60 data points are sufficient to fine-tune llms for question-answering. *CoRR*, abs/2409.15825.

Tian Ye, Zicheng Xu, Yuanzhi Li, and Zeyuan Allen-Zhu. 2024d. Physics of language models: Part 2.2, how to learn from mistakes on grade-school math problems. *CoRR*, abs/2408.16293.

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. 2025. LIMO: less is more for reasoning. *CoRR*, abs/2502.03387.

Liang Zhang, Katherine Jijo, Spurthi Setty, Eden Chung, Fatima Javid, Natan Vidra, and Tommy Clifford. 2024. Enhancing large language model performance to answer questions and extract information more accurately. *CoRR*, abs/2402.01722.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. LIMA: less is more for alignment. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2024. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

A Prompt for SFT

To ensure a fair comparison, we use uniform prompt templates during training.

```
{% if messages[0]['from'] == 'system' %}
    {% set system_message = '<<SYS>>\n' + messages[0]['value'] | trim +
    ↪  '\n<</SYS>>\n\n' %}
    {% set messages = messages[1:] %}
{% else %}
    {% set system_message = '' %}
{% endif %}
{% for message in messages %}
    {% if (message['from'] == 'user') != (loop.index0 % 2 == 0) %}
       {{ raise_exception('Conversation roles must alternate user/assistant...') }}
    {% endif %}
    {% if loop.index0 == 0 %}
        {% set content = system_message + message['value'] %}
    {% else %}
        {% set content = message['value'] %}
    {% endif %}
    {% if message['from'] == 'user' %}
        {{ bos_token + '<|question|> ' + content | trim + ' <|answer|>' }}
    {% elif message['from'] == 'assistant' %}
        {{ ' ' + content | trim + ' ' + eos_token }}
    {% endif %}
{% endfor %}
```

B More Details of Experiments

B.1 Details of Models

To ensure generalizable results, we analyze five LLMs from two different families.

LLaMA-2 Family The LLaMA-2 family includes three open-source LLMs developed by Meta. These models are pre-trained on over 2 trillion tokens, equipping them with extensive world knowledge and strong semantic representations. For this study, we select **LLaMA-2-7B**, **LLaMA-2-13B**, and **LLaMA-2-70B**.

LLaMA-3 Family The LLaMA-3 family builds upon the LLaMA-2 architecture with significant advancements, such as improved parameter efficiency and task generalization. We analyze **LLaMA-3-8B** and **LLaMA-3-70B**.

B.2 Details of Parameter Restoration

To assess how excessive parameter updates affect model performance, we compare the fine-tuned model with the pre-trained model by ranking parameters according to their relative change.

For each parameter i, let p_i denote its value before fine-tuning and s_i its value afterward. The relative change is defined as:

$$r_i = \frac{|s_i - p_i|}{|p_i|}$$

We sort all parameters in descending order of r_i to obtain the set I.

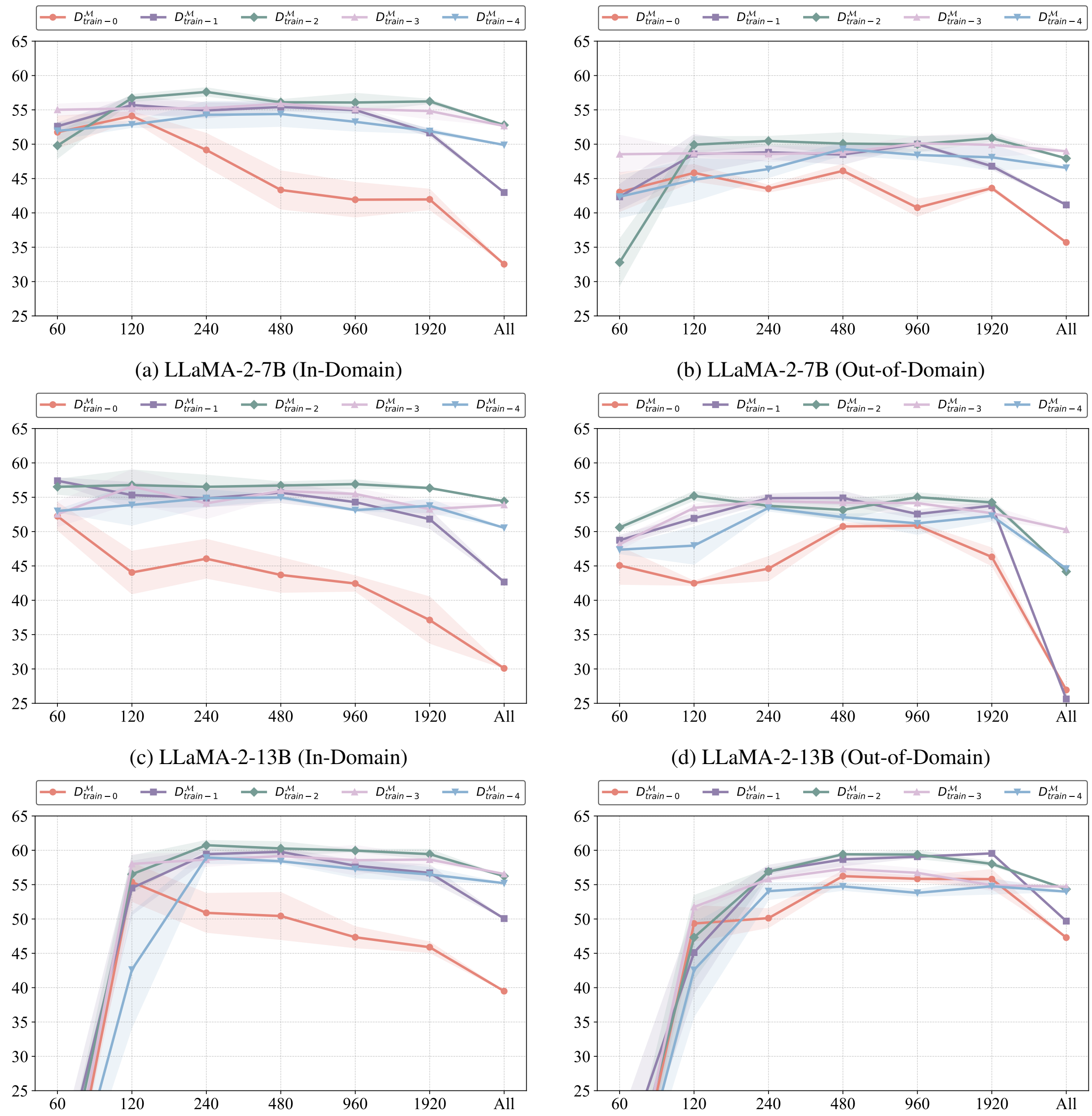
To measure the concentration of parameter updates, we compute the cumulative sum of r_i for the top percentage of parameters in I, divided by the total sum of all r_i. For instance, Table 3 shows that the top 1% of parameters contribute 70.59% of the total relative change.

C More Results

In this section, we present additional experimental results that are not included in the main body of the paper due to the limitation of space.

C.1 Test Results for the LLaMA-2 Family Models

We fine-tune five LLMs using datasets with five different levels of mastery. The results for the LLaMA-3 family models are presented in Section 3.4, while the results for the LLaMA-2 family are shown in Figure 6. Notably, although the peak performance occurs at different data sizes depending on the base model and hyperparameters, the trend of performance degradation beyond a certain point (size) remains consistent.



(a) LLaMA-2-7B (In-Domain)

(b) LLaMA-2-7B (Out-of-Domain)

(c) LLaMA-2-13B (In-Domain)

(d) LLaMA-2-13B (Out-of-Domain)

(e) LLaMA-2-70B (In-Domain)

(f) LLaMA-2-70B (Out-of-Domain)

Figure 6: In-domain ($\mathbf{Acc}^{\mathcal{M}}_{test}$) and out-of-domain ($\mathbf{Acc}^{\mathcal{M}}_{testood}$) performance of the LLaMA-3 family models fine-tuned with varying data scales, where ‘All’ indicates the use of the entire dataset listed in Appendix D.

C.2 Performance on the Training Set

We compare the performance of different LLMs fin-tuned from LLaMA-3-8B on their respective training sets when restoring different proportions of parameters. The results in Table 6 show that parameter reduction improves model performance on the training set, further supporting the idea that SFT introduces a significant number of unnecessary or even detrimental parameter updates.

Restore	$\mathcal{D}^{\mathcal{M}}_{\mathbf{train-0}}$	$\mathcal{D}^{\mathcal{M}}_{\mathbf{train-1}}$	$\mathcal{D}^{\mathcal{M}}_{\mathbf{train-2}}$	$\mathcal{D}^{\mathcal{M}}_{\mathbf{train-3}}$	$\mathcal{D}^{\mathcal{M}}_{\mathbf{train-4}}$
Number of Training Data: 240					
0	12.08	61.25	84.58	90.00	92.92
5%	12.50	62.92	85.00	90.83	93.75
20%	11.25	62.08	83.75	92.5	82.92
Number of Training Data: 1920					
0	16.56	62.81	83.44	89.48	93.39
5%	15.68	64.74	85.52	90.47	94.22
20%	15.16	65.00	89.06	90.57	94.90

Table 6: Performance of LLaMA-3-8B on the **training set** after restoring different scales of parameters across various fine-tuning datasets. Improvements over the non-restored model are highlighted in green, while performance declines are shown in red, with darker shades indicating larger differences.

C.3 Comparison of Results Across Different Strategies

We compare the performance of LLaMA-3-8B trained using four different strategies:

- **LLaMA-3-8B-Instruct**: A chat-optimized version fine-tuned by Meta, demonstrating strong performance across various benchmarks.
- **SFT (Mixed)**: Fine-tuning LLaMA-3-8B using a randomly mixed dataset. Results are tested across different data volumes, with the best outcomes reported.
- **SFT (Divided)**: Fine-tuning LLaMA-3-8B with data divided based on the model's mastery level. The best results are reported when fine-tuning with 1,920 samples.
- **LoRA**: Fine-tuning LLaMA-3-8B using a randomly mixed dataset with LoRA (Hu et al., 2022).
- **Parameter Restore**: Fine-tuning LLaMA-3-8B using the divided dataset, followed by a parameter restoration process. The best results are reported when fine-tuning with 1,920 samples.

The results in Table 7 indicate that data division and parameter restoration strategies significantly enhance model performance, offering valuable insights for optimizing data selection and fine-tuning approaches.

Strategies	**LLaMA-3-8B-Instruct**	**SFT (Mixed)**	**SFT (Divided)**	**LoRA**	**Parameter Restoration**
$\mathbf{Acc}^{\mathcal{M}}_{test}$	53.83	58.67	58.80	57.82	**62.21**
$\mathbf{Acc}^{\mathcal{M}}_{testood}$	54.14	53.88	54.04	51.52	**57.10**

Table 7: Performance of different LLMs fine-tuned using various strategies. The best results are highlighted in **bold**.

D Data Distribution of Different LLMs

Since data division is based on the model's mastery of the data, we analyze the data distributions corresponding to different pre-trained LLMs. The results for LLaMA-3-8B are presented in Section 3.1, while the distributions for other models are shown in Table 8.

$\mathcal{D}_{train}$	$\mathcal{D}^{\mathcal{M}}_{train-0}$	$\mathcal{D}^{\mathcal{M}}_{train-1}$	$\mathcal{D}^{\mathcal{M}}_{train-2}$	$\mathcal{D}^{\mathcal{M}}_{train-3}$	$\mathcal{D}^{\mathcal{M}}_{train-4}$
Number	12530	26805	14961	11542	10106
$\mathcal{D}_{test}$	$\mathcal{D}^{\mathcal{M}}_{test-0}$	$\mathcal{D}^{\mathcal{M}}_{test-1}$	$\mathcal{D}^{\mathcal{M}}_{test-2}$	$\mathcal{D}^{\mathcal{M}}_{test-3}$	$\mathcal{D}^{\mathcal{M}}_{test-4}$
Number	1595	3374	1876	1491	1219
$\mathcal{D}_{testood}$	$\mathcal{D}^{\mathcal{M}}_{testood-0}$	$\mathcal{D}^{\mathcal{M}}_{testood-1}$	$\mathcal{D}^{\mathcal{M}}_{testood-2}$	$\mathcal{D}^{\mathcal{M}}_{testood-3}$	$\mathcal{D}^{\mathcal{M}}_{testood-4}$
Number	2795	4517	1704	1542	1055

(a) LLaMA-3-70B

$\mathcal{D}_{train}$	$\mathcal{D}^{\mathcal{M}}_{train-0}$	$\mathcal{D}^{\mathcal{M}}_{train-1}$	$\mathcal{D}^{\mathcal{M}}_{train-2}$	$\mathcal{D}^{\mathcal{M}}_{train-3}$	$\mathcal{D}^{\mathcal{M}}_{train-4}$
Number	22725	30566	9336	7508	5809
$\mathcal{D}_{test}$	$\mathcal{D}^{\mathcal{M}}_{test-0}$	$\mathcal{D}^{\mathcal{M}}_{test-1}$	$\mathcal{D}^{\mathcal{M}}_{test-2}$	$\mathcal{D}^{\mathcal{M}}_{test-3}$	$\mathcal{D}^{\mathcal{M}}_{test-4}$
Number	2941	3805	1162	958	689
$\mathcal{D}_{testood}$	$\mathcal{D}^{\mathcal{M}}_{testood-0}$	$\mathcal{D}^{\mathcal{M}}_{testood-1}$	$\mathcal{D}^{\mathcal{M}}_{testood-2}$	$\mathcal{D}^{\mathcal{M}}_{testood-3}$	$\mathcal{D}^{\mathcal{M}}_{testood-4}$
Number	5201	4181	1030	786	415

(b) LLaMA-2-7B

$\mathcal{D}_{train}$	$\mathcal{D}^{\mathcal{M}}_{train-0}$	$\mathcal{D}^{\mathcal{M}}_{train-1}$	$\mathcal{D}^{\mathcal{M}}_{train-2}$	$\mathcal{D}^{\mathcal{M}}_{train-3}$	$\mathcal{D}^{\mathcal{M}}_{train-4}$
Number	20899	30562	9798	7996	6689
$\mathcal{D}_{test}$	$\mathcal{D}^{\mathcal{M}}_{test-0}$	$\mathcal{D}^{\mathcal{M}}_{test-1}$	$\mathcal{D}^{\mathcal{M}}_{test-2}$	$\mathcal{D}^{\mathcal{M}}_{test-3}$	$\mathcal{D}^{\mathcal{M}}_{test-4}$
Number	2675	3791	1275	1006	808
$\mathcal{D}_{testood}$	$\mathcal{D}^{\mathcal{M}}_{testood-0}$	$\mathcal{D}^{\mathcal{M}}_{testood-1}$	$\mathcal{D}^{\mathcal{M}}_{testood-2}$	$\mathcal{D}^{\mathcal{M}}_{testood-3}$	$\mathcal{D}^{\mathcal{M}}_{testood-4}$
Number	4671	4242	1233	981	486

(c) LLaMA-2-13B

$\mathcal{D}_{train}$	$\mathcal{D}^{\mathcal{M}}_{train-0}$	$\mathcal{D}^{\mathcal{M}}_{train-1}$	$\mathcal{D}^{\mathcal{M}}_{train-2}$	$\mathcal{D}^{\mathcal{M}}_{train-3}$	$\mathcal{D}^{\mathcal{M}}_{train-4}$
Number	15378	29468	13385	9344	8369
$\mathcal{D}_{test}$	$\mathcal{D}^{\mathcal{M}}_{test-0}$	$\mathcal{D}^{\mathcal{M}}_{test-1}$	$\mathcal{D}^{\mathcal{M}}_{test-2}$	$\mathcal{D}^{\mathcal{M}}_{test-3}$	$\mathcal{D}^{\mathcal{M}}_{test-4}$
Number	1956	3669	1719	1199	1012
$\mathcal{D}_{testood}$	$\mathcal{D}^{\mathcal{M}}_{testood-0}$	$\mathcal{D}^{\mathcal{M}}_{testood-1}$	$\mathcal{D}^{\mathcal{M}}_{testood-2}$	$\mathcal{D}^{\mathcal{M}}_{testood-3}$	$\mathcal{D}^{\mathcal{M}}_{testood-4}$
Number	3339	4537	1511	1338	888

(d) LLaMA-2-70B

Table 8: Data distribution for different LLMs.

E Discussion of Redundant Parameter Updates

E.1 Distribution of Redundant Parameter Updates

To investigate why SFT leads to redundant parameter updates, we analyze the distribution of redundant parameters in LLaMA-3-8B. As shown in Table 9 and Table 10, these parameters are spread across all layers of the model, with a higher concentration in the initial layers (i.e, 0–2), fewer in the final layers (i.e., 30–31), and a more uniform distribution in the middle layers (i.e., 3–29). This pattern may be due to the initial layers primarily encoding semantic information that is already well-learned during pretraining, resulting in greater parameter redundancy. In contrast, the final layers, which focus on output formatting, exhibit less redundancy. Furthermore, we observe that most redundant parameters are concentrated in the FFN layers, suggesting that their high parameter count presents a potential target for optimization. We also acknowledge that the emergence of redundant parameters may be linked to the lottery ticket hypothesis (Frankle and Carbin, 2019).

Layer	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Percentage	3.77	3.26	3.25	3.15	3.17	3.20	3.19	3.20	3.23	3.21	3.22	3.18	3.10	3.13	3.15	3.12
Layer	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**
Percentage	3.16	3.11	3.08	3.07	3.10	3.06	3.06	3.03	3.01	2.99	3.00	3.03	2.98	2.95	2.99	2.84

Table 9: Distribution of redundant parameter updates across layers in LLaMA-3-8B.

Module	**mlp.down**	**mlp.up**	**mlp.gate**	**attn.o**	**attn.q**	**attn.v**	**attn.k**
Percentage	28.91	28.26	23.37	9.40	6.41	2.55	1.09

Table 10: Distribution of redundant parameter updates across modules in LLaMA-3-8B.

E.2 Layers for Preserving Model Behavior

To identify which layers are most critical for preserving model behavior, we perform experiments that selectively restore parameters in different layers and evaluate the resulting performance. As shown in Table 11, our results reveal that the lower layers (0–3) are most crucial for maintaining model behavior, while the middle layers (4–27) have the least impact. This implies that SFT affects not only the predictive distributions in the upper layers but also substantially modifies the lower-layer parameters, thereby influencing overall model performance.

Restored Layers	**0-3**	**4-7**	**8-11**	**12-15**	**16-19**	**20-23**	**24-27**	**28-31**
$\mathcal{D}_{\text{train-0}}$	9.83	57.05	55.59	55.56	55.22	55.50	55.14	57.60
$\mathcal{D}_{\text{train-1}}$	21.59	58.22	57.83	58.02	58.03	57.99	57.96	57.99
$\mathcal{D}_{\text{train-2}}$	53.75	56.50	59.54	59.38	59.26	59.36	58.95	34.68
$\mathcal{D}_{\text{train-3}}$	2.80	59.43	59.12	59.17	59.15	59.18	59.15	50.76
$\mathcal{D}_{\text{train-4}}$	4.00	53.72	54.04	53.94	53.74	53.91	53.14	47.32

Table 11: Model performance when restoring different layers under different datasets.

F Details of Data Processing

In this section, we provide additional details on data processing.

F.1 Robust Multi-Template Complementation Mechanism

As described in Ye et al. (2024c), consider a knowledge fact k represented as a triple $(subject, relation, object)$, such as $(Painblanc, locatedin, France)$. Given a sentence $x = \text{map}(subject, relation)$ that maps the subject and relation (e.g., *Painblanc is located in*), an LLM $\mathcal{M}$ is considered to have memorized k if it can predict $y = \text{map}(object)$ by mapping the object (e.g., *France*) such that $y \subseteq \mathcal{M}(x)$.

Since $\mathcal{M}$ is a probabilistic model influenced by different mapping templates and sampling probabilities, we design $N_{\text{map}} = 21$ different mappings for each knowledge fact k. With the temperature set to 0.7, the model generates $N_{\text{sample}} = 10$ outputs for each mapping. The degree to which the LLM memorizes k is then calculated as:

$$R_k^{\mathcal{M}} = \frac{\sum_{i=1}^{N_{\text{map}}} \sum_{j=1}^{N_{\text{sample}}} \mathbb{I}(y_i \subseteq \mathcal{M}^j(x_i))}{N_{\text{map}} \times N_{\text{sample}}}$$

where x_i and y_i represent the results from the ith mapping, $\mathcal{M}^j$ denotes the jth sampled output, and $\mathbb{I}(\cdot)$ is the indicator function.

This approach effectively utilizes the characteristics of LLMs to evaluate their mastery of different data. However, as entities often have multiple aliases (e.g., *USA* and *United States*), the singular entity labeling in the original dataset may introduce biases. To enhance robustness, a synonym mapping table (Table 12) is constructed to expand the set of equivalent entity names, significantly improving result accuracy. This table is also used in judging the accuracy of LLMs' answers after SFT.

Object	**Synonyms**
United States of America	USA, United States, United States of America
New York City	New York, New York City
University of Michigan	UMich, University of Michigan
South Korea	South Korea, Republic of Korea, Korea
Saint Petersburg	Saint Petersburg, St. Petersburg
Buenos Aires	Baires, Buenos Aires
People's Republic of China	PRC, People's Republic of China, China
Ohio State University	Ohio State University, Ohio State
Bosnia and Herzegovina	Bosnia, Bosnia and Herzegovina, Bosna i Hercegovina
University of Texas at Austin	University of Texas at Austin, University of Texas, UT Austin
University of Cambridge	Cambridge University, Cambridge, University of Cambridge
United States Military Academy	United States Military Academy, West Point
Rio de Janeiro	Rio de, Rio de Janeiro
University of Edinburgh	Edinburgh University, University of Edinburgh
Museo del Prado	Prado Museum, Museo Nacional del Prado, Museo del Prado
Salt Lake City	Salt Lake, Salt Lake City
North Carolina State University	NC State, North Carolina State University
University of Durham	University of Durham, Durham University
Harvard Law School	Harvard University, Harvard Law School
University of Paris (1896-1968)	Université de Paris, University of Paris, Paris University
Newcastle upon Tyne	Newcastle upon Tyne, Newcastle
University of Oslo	University of Oslo, Oslo University
Hebrew University of Jerusalem	University of Jerusalem, Hebrew University, Hebrew University of Jerusalem
Carnegie Mellon University	Carnegie Mellon University, Carnegie Mellon
University of Oxford	Oxford University, University of Oxford
Autodromo Nazionale Monza	Monza, Autodromo Nazionale Monza
Indiana State House	Indiana State House, Indiana State
Imperial College London	Imperial College, Imperial College London
United Arab Emirates	UAE, United Arab Emirates, The Emirates

Table 12: Synonym mapping table for objects in the dataset.

F.2 Topics and Mapping Templates of Data

We categorize 10 location-related topics as in-domain data and another 14 unrelated topics as out-of-domain data, designing 21 mapping templates for each topic. The corresponding data details of in-domain data are listed from Table 13 to Table 22, while the corresponding data details of out-of-domain data are listed from Table 23 to Table 36.

Topic: P17 **Question Template:** Which country is {subject} located in?
Mapping Templates: {subject} is located in The location of {subject} is in {subject} finds its place within the borders of The {subject} is situated in the country, If you're seeking the {subject}, look no further than the nation of The land encompassing the {subject} is known as {subject} can be found in {subject} has its roots in The place {subject} calls home is {subject} is situated in The geographical location of {subject} is in {subject} can be discovered in the nation of The country where {subject} is found is {subject}'s location is in {subject} resides in The country of {subject} is {subject} belongs to {subject} exists in You can find {subject} in {subject} is a part of {subject} lies within the borders of

Table 13: Information and mapping templates for topic P17 (in-domain).

Topic: P19 **Question Template:** Where was {subject} born?
Mapping Templates: {subject} was born in The birthplace of {subject} was It is known that {subject} came into the world in {subject} entered the world in {subject} was born, and that location is {subject}'s life began in The location of {subject}'s birth is {subject}'s birth occurred in The place where {subject} was born is {subject} hailed from The answer to where {subject} was born lies in {subject} originated from {subject} came into this world in {subject} entered life in {subject} first drew breath in The origin of {subject} is in {subject} hails from The place of birth for {subject} is {subject}'s birth took place in When it comes to birth, {subject} was born in If you were to ask where {subject} was born, it would be

Table 14: Information and mapping templates for topic P19 (in-domain).

Topic: P20 **Question Template:** Where did {subject} die?
Mapping Templates: {subject} met their end at {subject} breathed their last at {subject}'s life came to a close at The place of {subject}'s death is The location of {subject}'s demise is The site of {subject}'s final rest is The place where {subject} passed away is {subject}'s mortal remains are in {subject} succumbed to death in The destination of {subject}'s last days was The story of {subject}'s life concluded in {subject} bid farewell to the world from within the confines of The final resting place of {subject} is {subject} took his final breath in {subject}'s life journey ended in {subject} died in The place where {subject} died is {subject}'s death occurred in {subject} took their last breath in When it comes to death, {subject} died in Looking at the end of {subject}'s life, they died in

Table 15: Information and mapping templates for topic P20 (in-domain).

Topic: P36 **Question Template:** What is the capital of {subject}?
Mapping Templates: The capital of {subject} is When considering the capital of {subject}, it is In {subject}, the city designated as the capital is {subject}'s capital city is The capital city of {subject} is located in {subject} is governed from The seat of government in {subject} is {subject}'s governmental hub is The administrative center of {subject} is The political heart of {subject} beats in One can find {subject}'s seat of power in the city of One would find {subject}'s governing institutions nestled within the boundaries of {subject}'s capital is The capital of the region {subject} is {subject}'s capital designation goes to The main city of {subject} is {subject} has its capital in The chief city of {subject} is Looking at {subject}, its capital is In terms of capital cities, {subject} has As the capital of {subject}, you'll find

Table 16: Information and mapping templates for topic P36 (in-domain).

Topic: P69
Question Template: Where was {subject} educated?
Mapping Templates:
{subject} received education at
{subject} completed the studies at
{subject} was schooled at
{subject} was educated in
{subject} graduated from
{subject} spent the formative years at
{subject}'s alma mater is
{subject} pursued the studies at
{subject} gained the knowledge at
The academic journey of {subject} took place in
The institution where {subject} studied is
Education for {subject} was pursued within the walls of
The educational institution attended by {subject} is
{subject} is an alumnus/alumna of
The academic background of {subject} includes
The place where {subject} was educated is
{subject} attended school in
The education of {subject} took place in
The place of {subject}'s education is
{subject} received their education from
In terms of education, {subject} was schooled in

Table 17: Information and mapping templates for topic P69 (in-domain).

Topic: P131	
Question Template:	Where is {subject} located?

Mapping Templates:
The location of {subject} is where you'll find
If you look where {subject} is, you'll see
Where {subject} resides, there also is
{subject} is located at
{subject} can be found in
{subject} is positioned at
{subject} is stationed at
{subject} is based at
{subject} is headquartered at
The current location of {subject} is
One would locate {subject} in the vicinity of
Currently, {subject} resides or occupies
{subject} is in
The geographical position of {subject} is
{subject} is placed in
You can find {subject} in
{subject} exists in
{subject} lies in
The location of {subject} is
{subject} is situated in
{subject} resides in

Table 18: Information and mapping templates for topic P131 (in-domain).

Topic: P159 **Question Template:** Where is the headquarter of {subject}?
Mapping Templates: The headquarter of {subject} is located in {subject} has its headquarter in You can find the headquarter of {subject} in {subject}'s central office is situated in The main hub of {subject} is {subject} is headquartered in The location of {subject}'s headquarter is {subject}'s headquarter can be found at The address of {subject}'s headquarter is {subject}'s headquarters are located at The central hub of operations for {subject} can be found in The administrative heart of {subject} resides at {subject}'s head office is located in {subject} has its main base in {subject}'s headquarters can be found in The headquarters of {subject} is located in {subject}'s headquarters is in The main office of {subject} is in {subject}'s headquarter is located in The headquarter of {subject} is situated in When it comes to headquarters, {subject}'s is in

Table 19: Information and mapping templates for topic P159 (in-domain).

Topic: P276
Question Template: Where is {subject} located?
Mapping Templates:
{subject} can be found at
The location of {subject} is
{subject} is situated at
{subject} has its base in
{subject} is headquartered in
{subject} operates out of
The place where {subject} is located is
{subject} is positioned at
The site of {subject} is
{subject} can be found in the location
The whereabouts of {subject} are at
{subject} is situated in the place called
{subject} is established in
The coordinates of {subject} point to
The address of {subject} leads to
{subject} is located in
{subject} resides in
You can find {subject} in
When it comes to location, {subject} is in
Looking at where {subject} is, it is in
In terms of location, {subject} is situated in

Table 20: Information and mapping templates for topic P276 (in-domain).

Topic: P495 **Question Template:** Which country was {subject} created in?
Mapping Templates: {subject} was created in The creation of {subject} took place in The origin of {subject} is traced back to {subject} was born in {subject} originated from {subject} was founded in {subject} was created in the country of The country of origin for {subject} is {subject} originated in the country of The birthplace of {subject} is none other than {subject}'s formation occurred in the borders of Historically, {subject} emerged in the country known as {subject} was conceptualized in The country credit for the creation of {subject} goes to The country that witnessed the creation of {subject} is The country where {subject} was created is {subject} was made in {subject} came into being in If you were to ask where {subject} was created, it would be Looking at the origin of {subject}, it was created in In terms of country of origin, {subject} was created in

Table 21: Information and mapping templates for topic P495 (in-domain).

Topic: P740
Question Template: Where was {subject} founded?
Mapping Templates: The founding of {subject} took place in {subject} was originally established in {subject}'s origin is traced back to {subject} was founded in {subject} originated in {subject} has its roots in The founding location of {subject} is {subject} has its origins in The birthplace of {subject} is One can trace {subject}'s beginnings to {subject} came into existence in The roots of {subject} dig deep into the soil of {subject} traces its beginnings back to The inception of {subject} took place in {subject} was brought into existence in The founding place of {subject} is The origin of {subject} is in The establishment of {subject} occurred in If you were to ask where {subject} was founded, it would be Looking at the origin of {subject}, it was founded in In terms of its founding location, {subject} was established in

Table 22: Information and mapping templates for topic P740 (in-domain).

Topic: P112 **Question Template:** Who founded {subject}?
Mapping Templates: The founder of {subject} is {subject} was founded by The establishment of {subject} was initiated by {subject} owes its existence to {subject} was brought into being by {subject} is a brainchild of {subject} was established by {subject} has its roots in The person who founded {subject} is The visionary behind {subject}'s establishment is The inception of {subject} can be traced back to The idea and realization of {subject} were the brainchild of {subject} was brought into existence by {subject}'s founder is known to be {subject} owes its inception to {subject} was created by The creation of {subject} is attributed to {subject} was started by {subject} originated with {subject}'s origins lie with {subject} can trace its roots back to

Table 23: Information and mapping templates for topic P112 (out-of-domain).

Topic: P127 **Question Template:** Who owns {subject}?
Mapping Templates: The owner of {subject} is {subject} is owned by Ownership of {subject} belongs to {subject} belongs to {subject} is in the possession of {subject} is a property of {subject} is possessed by {subject} is under the ownership of {subject} is held by The proprietor of {subject} is none other than Responsibility for {subject} falls under the jurisdiction of The property known as {subject} is under the stewardship of The rights to {subject} are held by The individual who owns {subject} is The rightful owner of {subject} is identified as Ownership of {subject} is held by The possession of {subject} is with The entity owning {subject} is {subject}'s owner is {subject} is in the hands of If you're looking for the owner of {subject}, it's

Table 24: Information and mapping templates for topic P127 (out-of-domain).

Topic: P170 **Question Template:** Who was {subject} created by?
Mapping Templates: {subject} was created by The creator of {subject} was The person who created {subject} is known as {subject} was founded by {subject} owes its creation to {subject} was developed by {subject}'s creator is {subject} was the creation of The person behind {subject} is {subject} was brought into existence by The originator of {subject} is The creative force behind {subject} is attributed to {subject} came into existence thanks to {subject} was brought to life by {subject} was conceptualized by The creation of {subject} is attributed to The entity responsible for creating {subject} is {subject} was made by {subject}'s creation is attributed to When it comes to creation, {subject} was created by Looking at the creation of {subject}, it was done by

Table 25: Information and mapping templates for topic P170 (out-of-domain).

Topic: P175 **Question Template:** Who performed {subject}?
Mapping Templates: The performer of {subject} was {subject} was performed by The one responsible for performing {subject} was {subject} was brought to life by {subject} was presented by {subject} was executed by The artist behind {subject} is The talent behind {subject} is The one who performed {subject} was The one who executed {subject} skillfully was The artist responsible for {subject}'s interpretation was The responsibility of performing {subject} fell upon {subject} was enacted by The act of {subject} was performed by {subject} was executed on stage by The execution of {subject} was done by {subject} was carried out by The realization of {subject} was by {subject} had its performance by The performance of {subject} was done by Looking at the performance of {subject}, it was done by

Table 26: Information and mapping templates for topic P175 (out-of-domain).

Topic: P176 **Question Template:** Which company is {subject} produced by?
Mapping Templates: {subject} is produced by The producer of {subject} is The production company behind {subject} is {subject} is created by {subject} is assembled by {subject} comes from {subject} is manufactured by The company responsible for {subject} is {subject} is a product of The production of {subject} falls under the umbrella of {subject} comes from the production house of The production of {subject} is handled by none other than The company behind the production of {subject} is The company that crafts {subject} is Every unit of {subject} bears the production mark of {subject} comes from the company The production of {subject} is handled by The company responsible for producing {subject} is The company that produces {subject} is When it comes to production, {subject} is produced by Looking at the production of {subject}, it is done by

Table 27: Information and mapping templates for topic P176 (out-of-domain).

Topic: P26 **Question Template:** Who is {subject} married to?
Mapping Templates: {subject}'s spouse is {subject} has been married to {subject} is in a marital union with The person {subject} is married to is {subject}'s partner in marriage is {subject}'s better half is {subject} is wed to {subject} exchanged vows with {subject} shares a life with {subject} shares a marital bond with Their love story culminated in a wedding, uniting {subject} and The answer to {subject}'s nuptials lies in the presence of {subject} is married to {subject} has tied the knot with {subject} shares a matrimonial life with The spouse of {subject} is {subject} is wedded to In marriage, {subject} is united with The one {subject} is married to is {subject}'s husband/wife is When it comes to marriage, {subject} is married to

Table 28: Information and mapping templates for topic P26 (out-of-domain).

Topic: P40 **Question Template:** Who is {subject}'s child?
Mapping Templates: The child of {subject} is known to be Belonging to {subject} as a child is As a child to {subject}, there is {subject}'s child is {subject} is the parent of {subject}'s offspring is {subject}'s youngster is {subject}'s family includes {subject}'s lineage includes {subject} has a child named The offspring of {subject} is identified as The child of {subject} is recognized as The offspring of {subject} includes {subject} is the biological parent of {subject} is the father/mother to The child of {subject} is The progeny of {subject} is The next generation of {subject} includes If you were to ask who {subject}'s child is, it's Looking at {subject}'s offspring, it's In terms of children, {subject} has

Table 29: Information and mapping templates for topic P40 (out-of-domain).

Topic: P413
Question Template: What position does {subject} play?
Mapping Templates:
{subject} plays
The position of {subject} is
In the team, {subject} holds the position of
{subject} plays the position of
{subject}'s role is
{subject} is a
The position played by {subject} is
{subject} holds the position of
{subject} is a player in the position of
In the game, {subject} assumes the role of
{subject} is known for the position as
When playing, {subject} takes up the position of
The role {subject} takes on is
{subject} is assigned to the position
The position that {subject} occupies is
{subject} occupies the position of
{subject} fulfills the role of
{subject} is positioned as a
The position that {subject} plays is
{subject}'s position is
If you were to ask what position {subject} plays, it's

Table 30: Information and mapping templates for topic P413 (out-of-domain).

Topic: P50
Question Template: Who is the author of {subject}?
Mapping Templates:
{subject} was authored by
The writer of {subject} is
The person who authored {subject} is
The author of {subject} is
{subject} was written by
{subject} is a work by
The creator of {subject} is
The person responsible for {subject} is
{subject} owes its existence to
The creative mind behind {subject} is none other than
{subject} was penned by the talented writer,
The work known as {subject} was brought to life by the author,
{subject} is a work authored by
The penname associated with {subject} is
The words of {subject} were put together by
The person who wrote {subject} is
{subject} was created by
{subject} was drafted by
If you were to ask who authored {subject}, it was
Looking at the authorship of {subject}, it was written by
{subject} is a creation of

Table 31: Information and mapping templates for topic P50 (out-of-domain).

Topic: P136 **Question Template:** What type of music does {subject} play?
Mapping Templates: The music played by {subject} is When {subject} plays music, it is The musical style of {subject} can be categorized as {subject}'s sound is characterized as {subject}'s musical talent lies in {subject} has a knack for {subject}'s genre of music is {subject} is known for playing {subject}'s music style is The genre that {subject} excels in is When it comes to music, {subject} is known for their proficiency in The tunes produced by {subject} belong to the category of {subject}'s music falls under the category of {subject} has a musical style that is categorized as The music played by {subject} can be described as The type of music {subject} plays is The genre of music {subject} plays is The style of music {subject} plays is {subject} plays the music type of Musically, {subject} is known to play In terms of musical style, {subject} plays

Table 32: Information and mapping templates for topic P136 (out-of-domain).

Topic: P106 **Question Template:** What kind of work does {subject} do?
Mapping Templates: {subject} is employed in {subject} earns a living by working as {subject}'s occupation is {subject} is engaged in {subject}'s profession is {subject} works as a {subject} makes a living as {subject} has a career in {subject} is involved in {subject} engages in the occupation of The work that {subject} undertakes is classified as The focus of {subject}'s employment lies in The type of work {subject} engages in is The work performed by {subject} falls under The work done by {subject} falls under the category of The kind of work {subject} does is {subject} operates in the field of The work {subject} performs is When it comes to work, {subject} does {subject} works in the field of The work done by {subject} is

Table 33: Information and mapping templates for topic P106 (out-of-domain).

Topic: P264 **Question Template:** What music label is {subject} represented by?
Mapping Templates: {subject} is represented by The music label representing {subject} is Regarding representation, {subject} is under {subject} has a record deal with {subject} has a musical partnership with {subject}'s music is released by {subject} is signed to {subject} is affiliated with {subject} has a contract with {subject} is represented by the music label The talented {subject} is associated with the music label {subject}'s discography is managed by the renowned label {subject} is under contract with the music label {subject} is affiliated with the music label The music label backing {subject} is {subject} is signed with the music label {subject} works with the music label {subject} is under the music label The music label that represents {subject} is {subject} has representation from If you were to ask what music label represents {subject}, it is

Table 34: Information and mapping templates for topic P264 (out-of-domain).

Topic: P407 **Question Template:** Which language was {subject} written in?
Mapping Templates: {subject} was originally written in The language used for writing {subject} was The original text of {subject} appeared in {subject} was penned in The language of {subject} is {subject} was composed in {subject} was created in {subject} is written in the language of The writing language of {subject} is {subject} was composed in the language known as The linguistic medium of {subject} is The choice of language for {subject} is {subject} was written in the language of The language used to write {subject} is The original language of {subject} is The writing of {subject} is in {subject} is composed in The text of {subject} is in {subject} was written in If you were to ask what language {subject} was written in, it's Looking at the language of {subject}, it's

Table 35: Information and mapping templates for topic P407 (out-of-domain).

Topic: P800
Question Template: What is {subject} famous for?
Mapping Templates:
{subject} is famous for
The fame of {subject} is due to
People recognize {subject} for
{subject} is renowned for
{subject}'s claim to fame is
{subject} is celebrated for
{subject} is known for
{subject} is distinguished by
{subject} is admired for
Fame comes to {subject} due to
Among its achievements, {subject} is celebrated for
{subject}'s popularity largely stems from
{subject}'s notable recognition comes from
{subject} is celebrated widely due to
The fame of {subject} is attributed to
The reason {subject} is famous is
{subject} is well-known for
{subject} gained fame for
If you were to ask what {subject} is famous for, it's
Looking at what made {subject} famous, it's
In terms of fame, {subject} is associated with

Table 36: Information and mapping templates for topic P800 (out-of-domain).