Pattern Shifting or Knowledge Losing? A Forgetting Perspective for Understanding the Effect of Instruction Fine-Tuning

Chunkang Zhang[1,3]*, **Boxi Cao**[1,3], **Yaojie Lu**[1], **Hongyu Lin**[1]†, **Liu Cao**[4]
Ke Zeng[4], **Guanglu Wan**[4], **Xunliang Cai**[4], **Xianpei Han**[1,2], **Le Sun**[1,2]
[1]Chinese Information Processing Laboratory [2]State Key Laboratory of Computer Science
Institute of Software, Chinese Academy of Sciences, Beijing, China
[3]University of Chinese Academy of Sciences, Beijing, China
[4]Meituan, Beijing, China
{zhangchunkang2021,boxi2020,luyaojie,hongyu,xianpei,sunle}@iscas.ac.cn
{liucao,zengke02,wanguanglu,caixunliang}@meituan.com

Abstract

Instruction Fine-Tuning(IFT) emerges as an essential step of training large language models to robustly carry out tasks of interest. However, there lacks a systematic investigation about the underlying mechanisms of instruction fine-tuning, particularly on the forgetting phenomenon after IFT, known as alignment tax. Therefore, to understand the mechanism of IFT from the forgetting perspective, we investigate the alternation of the text pattern and knowledge within models throughout the entire IFT process. Specifically, we *restore* fine-tuned models to their base version by training them on the data sharing a similar distribution with the pre-training corpus and compare their results Our experiment indicates that there is a stage transition of forgetting during IFT process: (1) **Pseudo Forgetting**: in this stage, models mainly shift their familiar text pattern away from pre-training data format while the world knowledge is preserved. Consequently, models will recover to their original performance when they are restored to the base version. (2) **Actual Forgetting**: in this stage, models forget the acquired knowledge as well. Therefore, they fail to reach the original performance even if they are restored to the base version.

1 Introduction

Instruction Fine-Tuning (IFT) has emerged as an indispensable process during the development of Large Language Models (LLMs) (Touvron et al., 2023a; Touvron et al., 2023b; OpenAI et al., 2023). By training LLMs in formatted input-output pairs, IFT enables them to demonstrate extraordinary capabilities on unseen tasks even in zero-shot settings(Wu et al., 2022). Since IFT requires fewer data compared with the pre-training stage and delivers superior outcomes(Wu et al., 2022), it has become an essential approach to enhance LLMs.

However, previous studies(Ouyang et al., 2022; Bai et al., 2022) find that fine-tuned models fail to solve the problems which their base versions could solve, resulting in performance degradation in specific tasks. Such a seemingly catastrophic forgetting(McCloskey and Cohen, 1989) phenomenon after IFT is reported as alignment tax(Ouyang et al., 2022).

Unfortunately, the underlying mechanism of such forgetting remains an open problem and receive distinct opinions. On the one hand, some researchers (Kotha et al., 2023; Jain et al., 2023;Kung and Peng, 2023) indicate that LLMs merely learn the superficial pattern of IFT data and forgetting originates from suppressing LLMs performance in areas out of distribution with the IFT data. On the other hand, other researchers(Burns et al., 2023;Yin et al., 2023;Lin et al., 2023) suggest that IFT enables LLMs to express the knowledge gained during pre-training in the format of instruction-output pairs, where forgetting originates from trading base models‘ generality for such speciality. The absence of such a

*Work was done during the internship at Meituan.
†Corresponding Author.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1381-1394, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

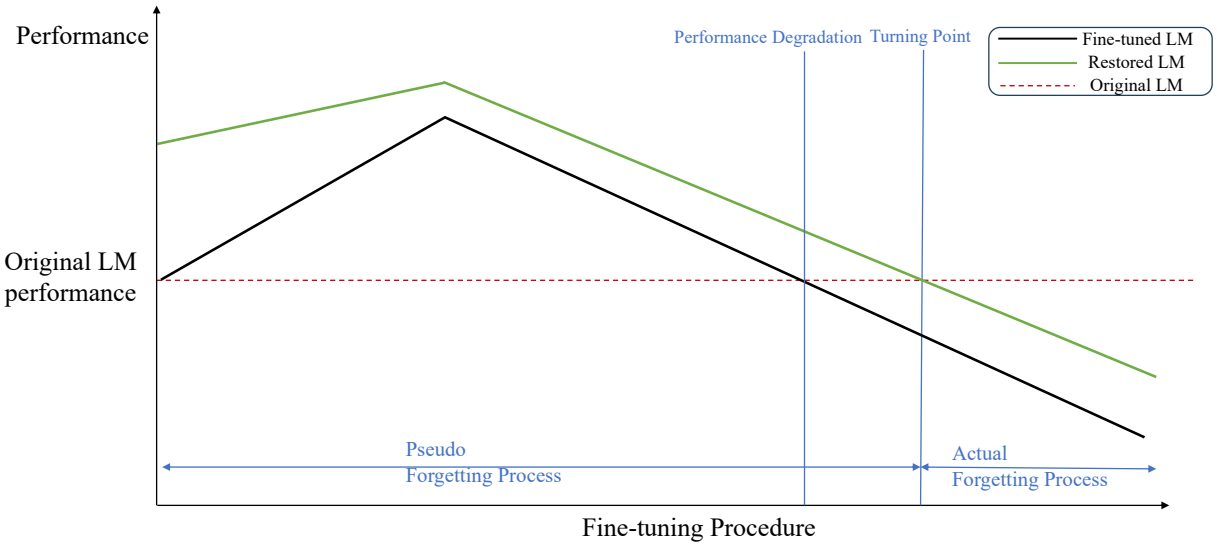


Figure 1: The illustration of theoretical instruction tuning process, during the initial stage, fine-tuned LLM is better than the original. Subsequently, LLM begins to fall behind the original base LLM while restoring it to the original text pattern can bring back the performance. In the late stage, even if LLM is restored to its original pattern, the performance is still worse than the original

systematic and unanimous analysis on the alignment tax limits our potential to develop a better IFT training strategy or thoroughly evaluate LLMs, which consequently delays improvements and further application of LLMs. Therefore, a comprehensive analysis of the underlying core factors behind forgetting after IFT is significant.

To this end, this paper aims to reveal the mechanism of IFT from the forgetting perspective. Specifically, this paper will explore two specific research questions:

- **RQ1: Does the forgetting effect of IFT follow specific patterns?**
- **RQ2: What is the characteristic of these forgetting patterns?**

For RQ1, we propose to restore fine-tuned LLMs to base LLMs in order to investigate whether the world knowledge(Touvron et al., 2023a) is retained in LLMs after IFT. Specifically, we collect data from the corpus base LLMs are pre-trained on, which shares a similar data distribution with the pre-training data. We then sample them to form the *restoring data* for training, which aims in restoring the LLMs familiar text pattern from instruction text to plain text, and *probing data* to measure whether the familiar text pattern is recovered to plain text. Subsequently, we apply several benchmarks to compare the performance between base, IFT and restored LLMs. We find that during the early IFT process, LLMs forget the original text pattern and is more familiar with the text pattern of fine-tuning data, we call such an alternation of text pattern as **pattern shifting**. After restoring the text pattern to the original plain text format, their recovered performance is comparable with base models. Such results indicates that the world knowledge still preserves and the drop originates from the forget of the original text pattern, which we refer to as *Pseudo Forgetting*. As for the late training process, even if the familiar text pattern is restored, the clear gap between recovered models and base models reveals that both text pattern and knowledge are forgotten, which we refer to as *Actual Forgetting*.

As for RQ2, we further investigate the probability distribution of the answer to the benchmark to further reveal the characteristic of forgetting pattern. For the pseudo forgetting stage, we analysis the probability and entropy of the answer across the whole vocabulary and find that the overall desired output generation probability across the whole vocabulary of LLMs begins to converge at high outcome during the early process of pseudo forgetting. We call such phenomenons as *Format Alignment*. After reaching a high generation probability, the entropy of the next generated tokens keeps altering, indicating that models are aligning their preserved knowledge to the desired output format, refered to as *Knowledge*

Alignment. The entropy keeps declining as the training process continues, revealing that models are becoming overconfident, which may cause the actual forgetting. As for the actual forgetting stage, we conduct experiment to measure the memorization of the instruction fine-tuning data and confirm that over memorizing training data is another reason behind actual forgetting.

By summarizing and analyzing results from the performance of different models and instruction data in knowledge benchmark and probing data, we find that there exists a stage of transition in forgetting and shares specific characteristic as followed:

- LLMs are undergoing **Pseudo Forgetting** process during the early stage of IFT. We find that LLMs mainly forget the text pattern of the pre-training data. Specifically, once the fine-tuned LLMs are restored, their performances exceeds the base LMMs.
- LLMs are undergoing **Actual Forgetting** process during the late stage of IFT. We find that LLMs not only forget the text pattern, but also lose their world knowledge. Specifically, we observe a notable gap from base LLMs in performance, even if the performance increase when these LLMs are restored.
- LLMs finish **Format Alignment** early and begins **Knowledge Alignment** during the pseudo forgetting. We conclude that models initially align their output pattern with the questions, which we refer to as format alignment, and continue altering their answers after the completion of format alignment, indicating LLMs are undergoing knowledge alignment.

2 Related Work

Mechanism of Instruction Fine-tuning Instruction fine-tuning has already become an essential part of developing large language models. However, there is not a uniform opinion towards the mechanism of instruction fine-tuning. On the one hand, some researchers believe that LLMs merely learn the superficial pattern of instruction(Kung and Peng, 2023;Yin et al., 2023;Gupta et al., 2022), On the other hand, other researchers believe that instruction fine-tuning unlocks the potential LLMs acquire during pre-training stage(Burns et al., 2023;Wang et al., 2023). Such disagreement towards the mechanism of IFT restricts the thorough understanding of its theoretical value.

Distributional Shift the pre-training and fine-tuning framework has already realized significant success across a variety of applications (Devlin et al., 2019; Radford et al., 2021). Nonetheless, the deployment of pre-trained models into actual real-world settings and their subsequent fine-tuning often leads to a prevalent dilemma: the models encounter novel instances from a target distribution that diverges from the one used during fine-tuning (Andersen and Maalej, 2022; Goyal et al., 2022; Zhang and Ré, 2022). To tackle this problem, a range of solutions have been suggested. For example, Cha et al. (2021b) advocates for the utilization of a weight ensemble that combines the pre-trained and fine-tuned models to improve performance on out-of-distribution (OOD) data. Another method, introduced in (Kumar et al., 2022), is the LP-FT approach, which entails commencing with a pre-trained feature extractor that is paired with a reasonably competent classifier. This initial step is crucial, especially when the classifier begins with a random initialization, as the pre-trained features could quickly become skewed to fit the random classifier during fine-tuning, thereby intensifying the catastrophic forgetting issue.

Catastrophic Forgetting Traditional neural network is prone to forgetting the knowledge from a previously learned capabilities upon learning a new task(McCloskey and Cohen, 1989; McClelland et al., 1995; French, 1999). To address such issues, numerous strategies have been developed, such as parameter penalization(Kirkpatrick et al., 2017), continue learning(Cha et al., 2021a; Peng and Risteski, 2022), knowledge distillation(Rebuffi et al., 2017), lifelong learning(Silver et al., 2013;Fischer, 2000) and so on. With the arrival of Large Language Models, research towards catastrophic forgetting mainly focuses on revealing the mechanism of catastrophic forgetting behind continual fine-tuning(Luo et al., 2023;Scialom et al., 2022;Zeng et al., 2023) or continual pre-training(Xia et al., 2023;Tirumala et al., 2022). Therefore, the general forgetting of the pre-train knowledge during instruction tuning remains an open problem,

Kotha et al. (2023) indicates catastrophic forgetting originates from suppressing LLMs performance on data out of distribution with instruction fine-tuning data while Lin et al. (2023) proposes trade-off between speciality of instruction tuning data and generality of pre-traing is the factor.

3 Preliminaries

This section mainly introduces the process of the experiment testbed and corresponding setting to help better understand effect of the finding. Specifically, this section introduces the techniques and details of restoring stage in section 3.1, followed by the introduction of the procedure about testbed in section 3.2. the specific experiment settings are presented in Section 3.3

3.1 Restoring the Fine-tuned Model

We propose to restore fine-tuned models to its base versions with data that shares a similar distribution pattern with their pre-training corpus to determine if the knowledge is retained within LLMs. After fine-tuning LLMs on instruction format data, they shift their familiar text distribution patterns to instruction and become less familiar with the pre-training text. Thus the benchmark performance does not thoroughly reflect the related world knowledge within LLMs, it is likely to be related with other phenomenons, such as whether LLMs understand the question of specific task or learn the correct output format. Therefore, We train the fine-tuned LLMs with text resembling their original pre-training data to restore their familiar text pattern from instruction data to the original plain text to ensure the existence of certain knowledge in LLMs.

To be more specific, we control the overall tokens of the restoring training data to be the same as those of IFT data in one epoch, since the main purpose of the restoring stage is to recover the familiar text pattern of the LLMs from instruction to the pre-training data instead of injecting knowledge as the pre-training stage does. Controlling the amount of tokens to be the same as IFT data can reflect whether the restoring process is possible based on the comparison with the base LLMs on probing data as well as a more convinced confirmation of forgetting stages.

In order to further control the distributional shift during restoring stage, we rank the data based on perplexity of llama from low to high and select the top ones as the restoring data. By doing so, the selected data is more possible to be pre-trained before, thus no more new knowledge of text pattern will be introduced, assuring the restoring process just restore the text pattern.

3.2 Testbed Procedure

This paper leverages world knowledge probing benchmark as a testbed in order to investigate the forgetting mechanism of LLMs, which focuses on determining whether the world knowledge exists after IFT. As illustrated in Figure **??**, to evaluate how well LLMs preserves their learned knowledge after IFT, we fine-tune these LLMs with instruction data, probe them with knowledge benchmarks and check their answers through out the whole IFT process. To comprehensively determine the existence of specific knowledge, we propose to restore the fine-tuned LLMs to be close to the base version, which controls the effect of pattern shifting of IFT on knowledge probing result.

3.3 Experiment Setting

Language Model We conduct experiments on transformer-based architecture such as LLaMa2-7b and LLaMa2-13b(Touvron et al., 2023b), since numerous LLMs are derived from these models. We believe findings based on these models will be more general.

Training Data The construction of our testbed requires both instruction fine-tuning process and restoring process. For the instruction tuning process, we utilize alpaca(Taori et al., 2023) and vicuna(Zheng et al., 2023) as the instruction training dataset. As for the pattern restoring process, we randomly sample data from Common Crawl-MAIN-2023-40 corpus(Touvron et al., 2023b), which shares a similar distribution with the pre-training corpus that LLaMa2 is pre-trained on to avoid another distribution shifting. Subsequently, we sample about top forty thousand passages which shares the same amount of

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1381-1394, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers

tokens as the instruction tuning data with a low perplexity and ten thousand passages as the probing data, indicating whether fine-tuned models are restored to their original pattern.

Benchmark and Evaluation Metric We construct our experiment testbed based on ARC(Min, 2023) as well as MMLU(Hendrycks et al., 2021) for probing world knowledge. ARC is designed to evaluate the common sense knowledge while MMLU is a comprehensive benchmark containing variety of domain knowledge. These benchmarks test distinct world knowledge type within LLMs, leading to a more comprehensive finding. As for the evaluation metric, apart from the accuracy of MMLU and ARC, we apply perplexity on formally sampled probing data to measure whether the pattern is restored following the setting of Xia et al. (2023).

Training Details During the instruction tuning process, we only calculate loss on outputs, setting epoch to 10, learning rate to $2e^{-5}$ and batch size to 256. As for the restoring process, we set epoch to 1, learning rate to $3e^{-5}$ and batch size to 256. We use FSDP(Zhao et al., 2023) for all the training stage and all experiments are implemented on Nvidia A100-80GB GPUs.

4 Stage Transition of Forgetting during Instruction Fine-Tuning

This section proposes results and findings of the forgetting phenomenon. Experiment results lead to following conclusions that forgetting in instruction tuning has two stages, pseudo forgetting and actual forgetting Further analysis based on these conclusions are presented in section 4.2.

4.1 Result and Findings

Finding 1. *LLMs suffer from pseudo forgetting during the early stage of instruction tuning, where world knowledge is retained in LLMs*

To show this, we firstly investigate whether the restoring process capable of restoring IFT models' famliar text pattern to their pre-training data version. Figure 3a is the trajectory of perplexity on probing data of the ten-epoch fine-tuned model, where we find the perplexity declines as the training process continues and conveys to the original base model perplexity. Such a result indicate that the text pattern is restored to the pre-training data style.

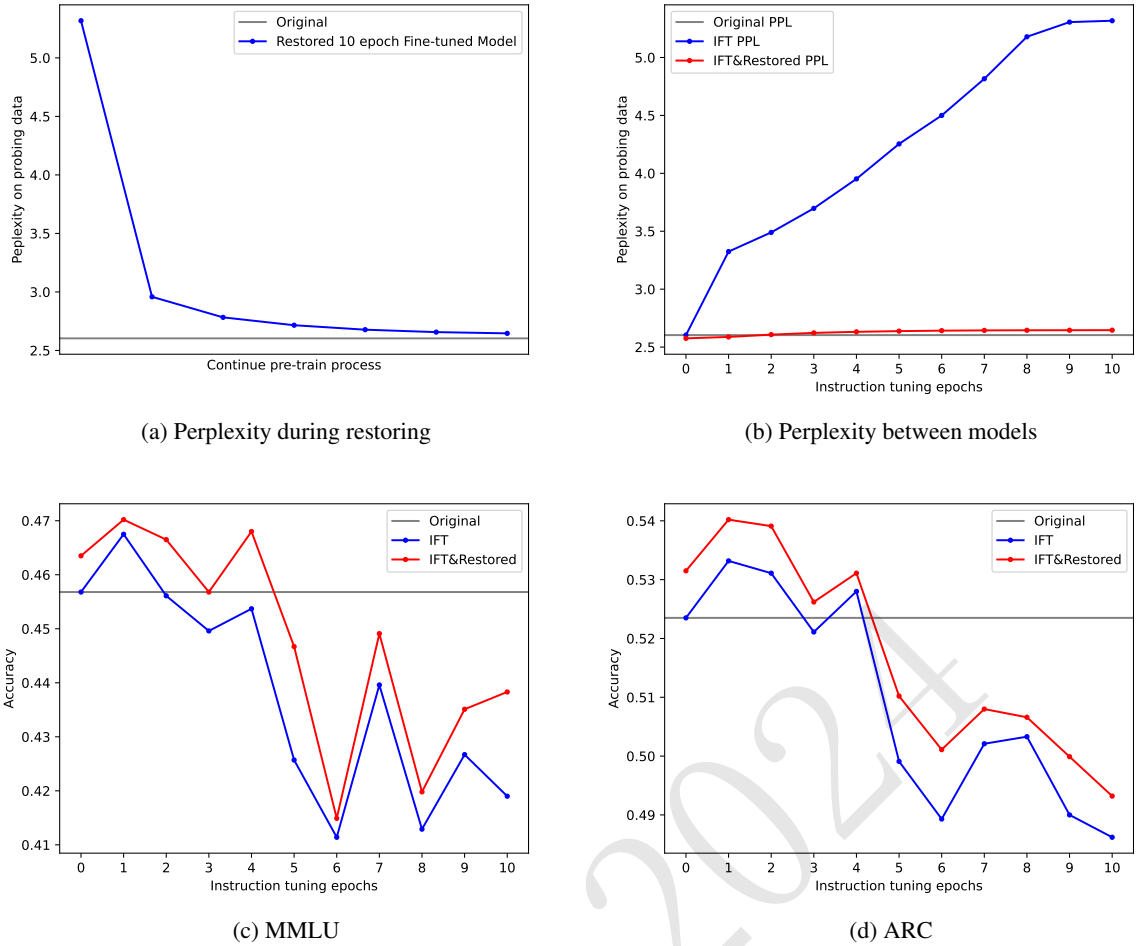
From Figure 3b, we find that LLMs' perplexity against pre-training probing data reveals a continuing upward trend along with the instruction tuning process, indicating that the text pattern LLMs learn during the pre-training process is forgotten. Subsequently, the perplexity of different fine-tuned LLMs have all returned to convergence with the base LLM no matter how large the perplexity is. Both of these results reveal the effectiveness of the restoring process on recovering text pattern.

Subsequently, to further determine whether the world knowledge is retained, we conduct experiments on world knowledge benchmark. Figure 3c and Figure 3d demonstrate the performance of IFT model trained on alpaca as well as the restored LLMs, which share almost identical text pattern with base LLMs as revealed in Figure 3b.

From Figure 3c and Figure 3d, we can see that during the initial training process, knowledge is retained in LLMs: 1) despite the majority of the fine-tuned outcomes falling short of the established original baseline, there is a marked enhancement in performance after the restoring stage. 2)In the early training process, the performance of LLMs exceeds original base LLMs when their familiar text patterns have been restored. Both phenomenons indicate that initial degradation after IFT originates from pattern shifting. When the pattern is restored to its original setting, performance is restored near or above the baseline, confirming that the world knowledge within LLMs is retained.

Finding 2. *LLMs suffer from actual forgetting during the latter stage of instruction tuning, where LLMs not only shift text pattern away but also lose learned world knowledge*

Figure 3c and Figure 3d illustrate a notable distinction during the late IFT process, indicating the lost of knowledge we refer to as actual forgetting, where performances of LLMs on both benchmarks are notably below the original LLMs even though text patterns are restored. Such phenomenon suggests that IFT induces a more profound and potentially deleterious effect on the models' knowledge base. Together



(a) Perplexity during restoring

(b) Perplexity between models

(c) MMLU

(d) ARC

Figure 3: Figure 3a illustrates the perplexity on probing data of fine-tuned llama2-7b along with the restoring process, and Figure 3b reveals perpexity on probing data of fine-tuned and restored models. Figure 3c and Figure 3d show the performance of original, sft and restored models on both MMLU and ARC.

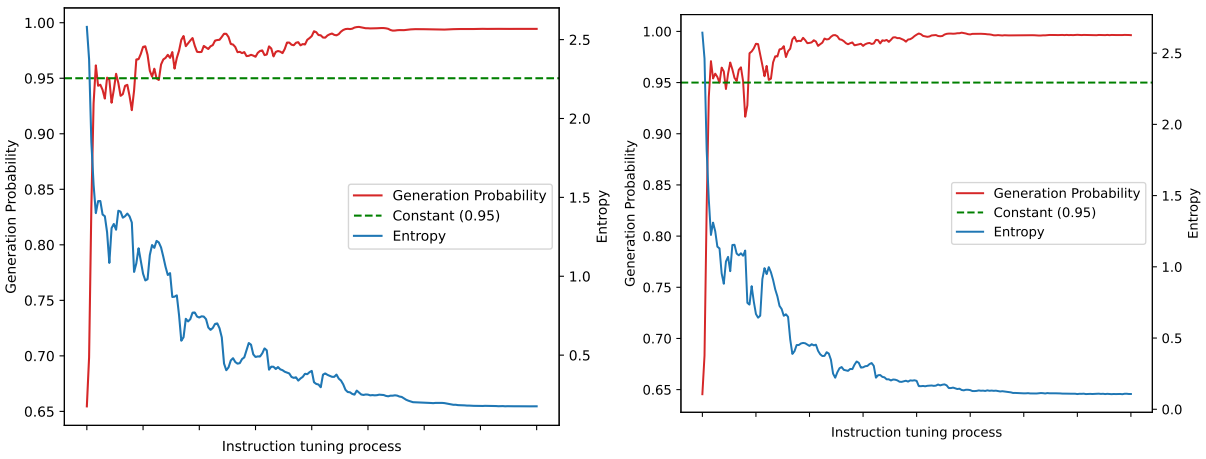
these results imply that during the late process of IFT, LLMs begin to forget specific knowledge that were previously acquired during pre-training stage.

4.2 Analysis

During the instruction tuning phase, the empirical evidence suggests that the performance of LLMs experiences an initial surge, surpassing the baseline, before subsequently diminishing and forming a substantial deficit relative to the baseline. Such a performance fluctuation implies that the knowledge embedded within the LLMs is initially more accessible as a result of IFT. However, this activation does not have sustained durability, as indicated by the eventual performance drop. Consequently, when the text pattern of the LLMs is reverted to its original through restoring stage, the performance of the LLMs not only recovers but also exceeds the original baseline, which is possibly the outcome of knowledge activation as Burns et al. (2023) indicates. Thus we move on to investigate the specific characteristic of these forgetting stages.

5 Characteristic of Forgetting Stages

In this section, we verify the characteristic of stages during the pseudo forgetting process, we focus more on the trajectory of actual generation probability as well as entropy over the whole vocabulary distribution. We introduce the preliminaries in section 5.1 and the corresponding finding in section 5.2. Our result



(a) MMLU (b) ARC

Figure 4: Entropy and generation probability across different selection in the vocabulary, where we find the format alignment finish during the pseudo forgetting, while the entropy descends as the training process goes on.

indicates that format alignment completes during the pseudo forgetting stage and begins knowledge alignment. Models begin to be over confident as the training proceeds which might cause the actual forgetting.

5.1 Preliminary

Exact Memorization Derived from the definition in Tirumala et al. (2022), Let V denote the vocabulary size. Let C denote a set of contexts, which can be thought of as a list of tuples (s, y) where s is an input context (instruction text) and y is the index of the ground truth token in the dataset that completes the block of text. Let S denote the set of input contexts, and let $f : S \rightarrow R^V$ denote a language model. A instruction context $c = (s, y) \in C$ is memorized if $\text{argmax}(f(s)) = y$. We refer to this as exact memorization, thus we define the overall exact memorization score EM as the proportion of $\text{argmax}(f(s)) = y$ through out the whole instruction dataset.

5.2 Result and Findings

Finding 1. *format alignment is hard to interfere and completes early during pseudo forgetting, while knowledge alignment takes over the pseudo forgetting phase after the completion of format alignment*

To show this, we dive into the generation process of LLMs during IFT. Figure 4a and Figure 4b illustrate the average generation probability of LLMs on benchmark questions as well as the entropy of the possible generation throughout the whole vocabulary.

From Figure 4a and Figure 4b we can see that the generation probability of correct format answer rise up in a few steps and converges. According to the principle of significance in statistics(Yaddanapudi, 2016), p-value less than 0.05 suggests that the model's output was not due to random chance. As depicted in Figure 4a and Figure 4b, models exhibit format alignment early in the training phase, maintaining a probability of generating the anticipated response above 95%. This high probability persists throughout the entire training process demonstrating the model's format alignment is not readily susceptible to disruption.

To further confirm the knowledge alignment, we can conclude from Figure 4a and Figure 4b that after the completion of format alignment, LLMs' internal entropy undergoes a period of fluctuation. This observation suggests a dynamic alteration in the responses provided by the model throughout the training phase. Despite these variations, the probability of the model producing correctly formatted responses remains stable. Such a fluctuating trend means that the distribution of the answer is changing, proving the existence of the model's knowledge alignment.

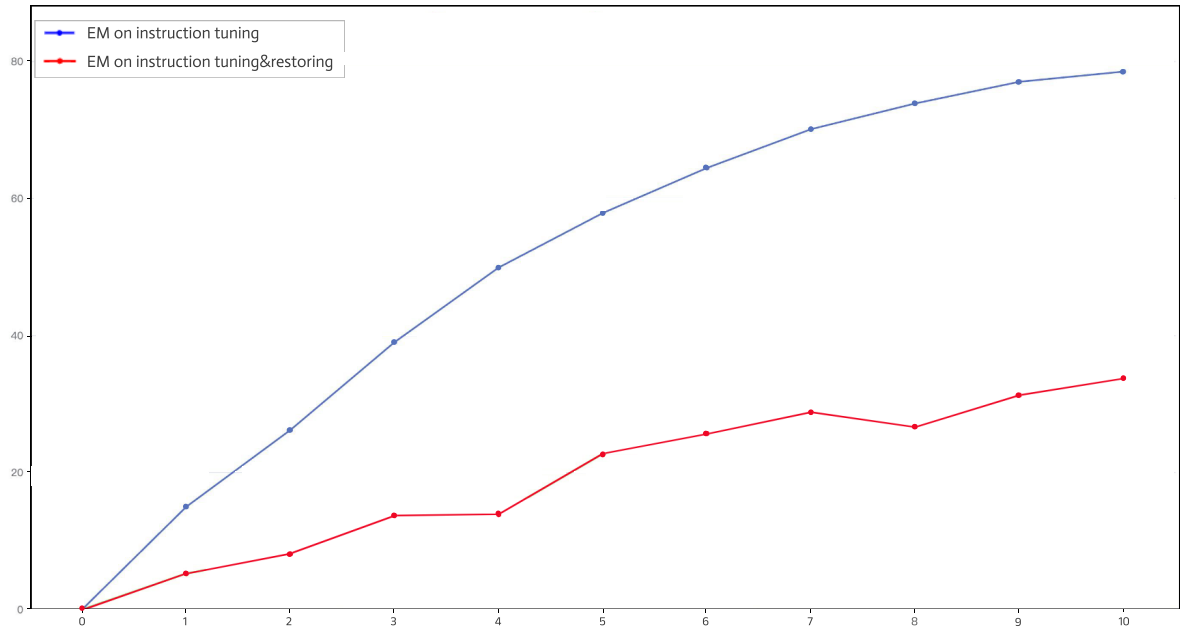


Figure 5: The illustration of alpaca EM along with the training epochs

Finding 2. *LLMs are becoming over-confident during the later training process, likely to be the cause of actual forgetting*

Figure 4a and Figure 4b provides a visual representation of the entropy and accuracy trajectories over the instruction tuning process. The data manifests a consistent decrement in overall entropy, indicative of the model's escalating certainty in its predictions. Focusing on the late training process, we find that the entropy on both benchmark no longer fluctuate and undergoes a decline trend. Such results suggest that as the model's confidence in its answers intensifies, the precision of its responses to these questions concomitantly deteriorates. This phenomenon points towards overconfidence as a contributing factor to the emergence of actual forgetting.

Finding3. *LLMs are memorizing data throughout the instruction tuning phase, these data remains even after restoring the original pattern*

We can conclude from Figure 5 that as the instruction tuning process proceeds, LLMs start to memorize more instruction data and form a ascending trend along with the training process. Figure 3b shows that after the pattern is restored to its original pattern, a distinct gap is still observed from the base model, indicating the presence of actual forgetting. We can initially conclude from the ascending trend in Figure 5 and its comparison with 3b that when the data restored in LLMs exceed a certain threshold, the original knowledge distribution is affected, which cause the actual forgetting phenomenon.

6 Forgetting across models and data

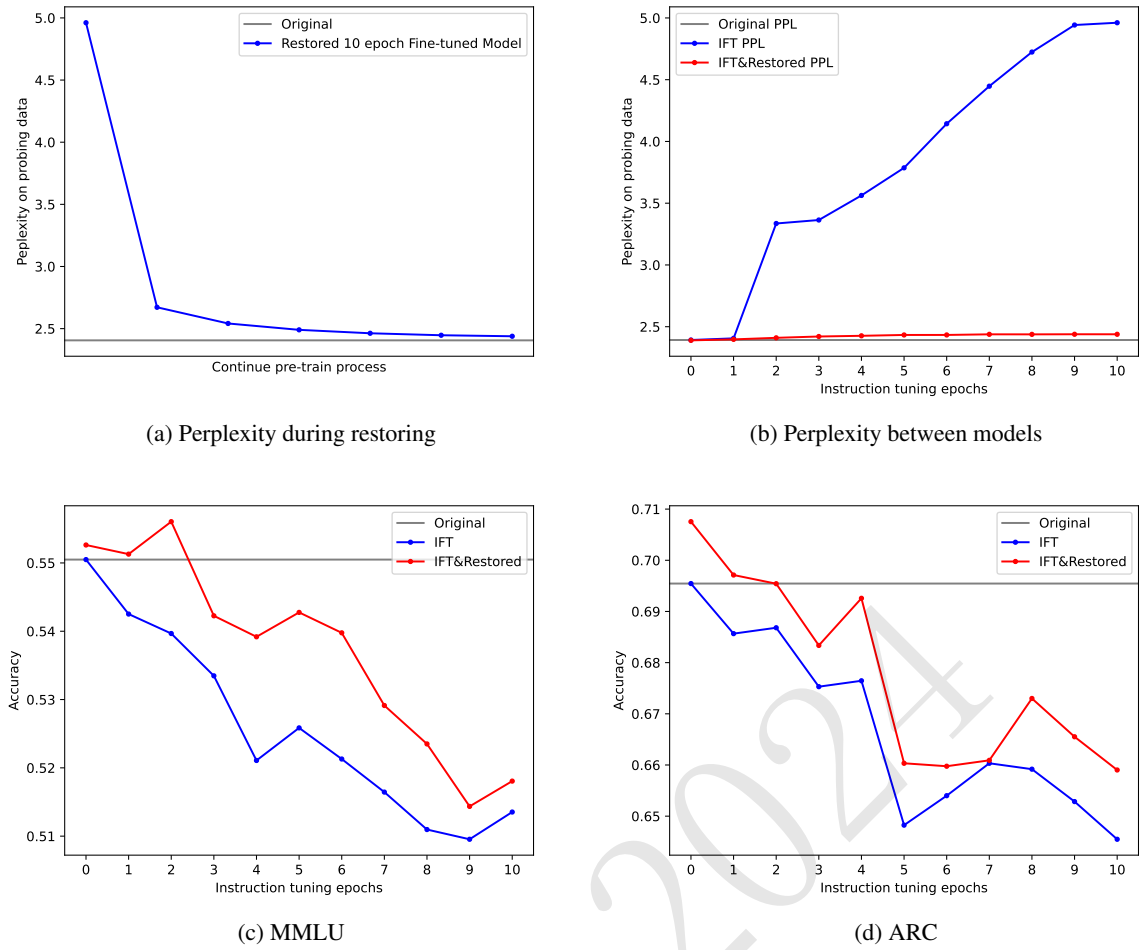
This section expands the forgetting phenomenon to a larger LLMs and other IFT data to investigate whether such a pattern of forgetting preserves across models. Experiment results implies a similar stage transition of forgetting, which helps to settle down the finding of this paper. The comparison with the smaller model reveals that larger model are more likely to face actual forgetting.

6.1 Result and Findings

Finding 1. *Pseudo forgetting and actual forgetting are also discovered across models*

Figure 6 illustrates both the perplexity and performance of LLaMa2-13b during the IFT process on alpaca as well as its restored version, from Figure 6a and Figure 6b we can conclude that the text pattern is restored which confirms the effectiveness of restoring stage on larger models. Figure 6c and Figure 6d

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1381-1394, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers



(a) Perplexity during restoring

(b) Perplexity between models

(c) MMLU

(d) ARC

Figure 6: Perplexity trajectory of LLaMa2-13b as well as the performance on knowledge benchmarks.

both indicate the existence of pseudo forgetting and actual forgetting with the gap between the original performance. We can conclude that the transition of forgetting keeps its consistency across the different LLMs, while the pattern shifting after IFT could also be restored to its original pattern.

Finding 2. *Larger Models are more likely to face actual forgetting*

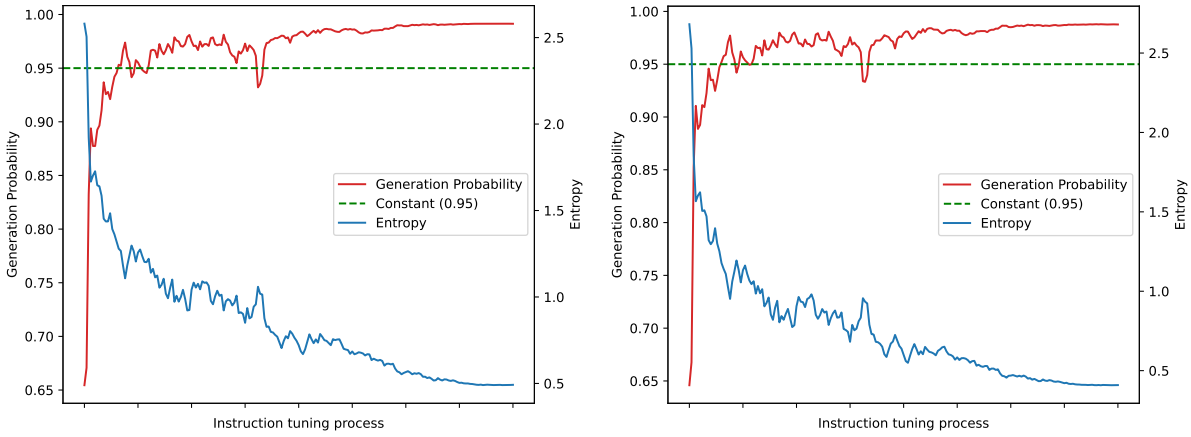
Figure 6c and Figure 6d where we find that after two epochs of instruction fine-tuning, model fails to reach its original performance even if the text pattern is restored, While the actual forgetting in Figure 3c and Figure 3d begins after four epochs of instruction fine-tuning. Such difference indicate that larger models are more likely to lose their knowledge during instruction fine-tuning and face actual forgetting.

Finding 3. *format alignment and knowledge alignment persists across IFT data*

we rerun the testbed on LLaMa2-7b fine-tuned with vicuna and calculate the entropy and generation probability on output. From figure 7a and 7b we can also find that the generation probability of correct format answer also conveys at a high point, indicating the format alignment,while the fluctuation of entropy reveals the existence of knowledge alignment indicating that the specific characteristic also exist in models trained with other IFT data.

7 Conclusion

In our research, we have delved into the phenomenon of catastrophic forgetting within the context of Iterative Fine-Tuning (IFT) with the objective of elucidating its underlying mechanisms. Through our investigations, we have identified two distinct stages of forgetting: Pseudo Forgetting and Actual



(a) MMLU (b) ARC

Figure 7: Entropy and generation probability across different selection in the vocabulary, where we find the format alignment finish during the pseudo forgetting, while the entropy descends as the training process goes on.

Forgetting. During the initial phase of IFT, we observe what we term as Pseudo Forgetting, where Large Language Models (LLMs) like LLaMa undergo a pattern shift rather than actual knowledge loss. The models adapt their existing patterns to new instructions while retaining the knowledge acquired during pre-training. This stage is characterized by a performance trajectory that initially oscillates before trending downward. In contrast, Actual Forgetting emerges in the later stages of IFT, signifying a genuine loss of knowledge within the LLMs. Our experiments demonstrate that models saved at later checkpoints exhibit a significant performance disparity when compared to their original pre-trained state, indicating that knowledge has indeed been forgotten during the instruction tuning process.

Our study also indicates the characteristic of these forgetting stages, LLMs are undergoing format alignment during the initial training process since the generation probability of correct format answer reaches a high point in just a few steps. After the completion of format alignment, the generation probability persist while the entropy across these correct format answer fluctuate, indicating the knowledge alignment. During the late stage of training, the entropy declines continuously, implying that models are over confident at their answer, which may cause the actual forgetting. Finally, by calculating the EM of LLMs on instruction data, we find that LLMs are memorizing IFT data throughout the instruction tuning process and may cause the actual forgetting.

In conclusion, our research provides significant insights into the stages of catastrophic forgetting in LLMs during IFT and establishes a foundation for future work aimed at mitigating such forgetting in Large Language Models.

Limitations

Due to the constraint of computational resources, our major experiment was conducted on LLaMa, which may not cover all the forgetting mechanism throughout other model families, in the future we will extend our experiments to investigate the forgetting of instruction fine-tuning on other model families as well. Besides, owing to the lack of fair and intuitive metric, we apply multiple choice question as the probing methods for knowledge instead of generation question, we will continue to design a metric to investigate the forgetting phenomenon under the circumstances of generation.

Acknowledgement

We sincerely thank the reviewers for their insightful comments and valuable suggestions. This research work is supported by the National Natural Science Foundation of China under Grants no. 62122077, 62106251 and 62306303.

References

Jakob Smedegaard Andersen and Walid Maalej. 2022. Efficient, uncertainty-based moderation of neural networks text classifiers. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1536–1546, Dublin, Ireland. Association for Computational Linguistics.

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 Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeff Wu. 2023. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision.

Hyuntak Cha, Jaeho Lee, and Jinwoo Shin. 2021a. Co2l: Contrastive continual learning. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 9516–9525.

Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. 2021b. Swad: Domain generalization by seeking flat minima.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Gerhard Fischer. 2000. Lifelong learning—more than training. *Journal of Interactive Learning Research*, 11(3):265–294.

Robert M French. 1999. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135.

Sachin Goyal, Ananya Kumar, Sankalp Garg, Zico Kolter, and Aditi Raghunathan. 2022. Finetune like you pretrain: Improved finetuning of zero-shot vision models.

Prakhar Gupta, Cathy Jiao, Yi-Ting Yeh, Shikib Mehri, Maxine Eskenazi, and Jeffrey P. Bigham. 2022. Instructdial: Improving zero and few-shot generalization in dialogue through instruction tuning.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding.

Samyak Jain, Robert Kirk, Ekdeep Singh Lubana, Robert P. Dick, Hidenori Tanaka, Edward Grefenstette, Tim Rocktäschel, and David Scott Krueger. 2023. Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Suhas Kotha, Jacob Mitchell Springer, and Aditi Raghunathan. 2023. Understanding Catastrophic Forgetting in Language Models via Implicit Inference. ArXiv:2309.10105 [cs].

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. 2022. Fine-tuning can distort pretrained features and underperform out-of-distribution.

Po-Nien Kung and Nanyun Peng. 2023. Do Models Really Learn to Follow Instructions? An Empirical Study of Instruction Tuning. ArXiv:2305.11383 [cs].

Yong Lin, Lu Tan, Hangyu Lin, Zeming Zheng, Renjie Pi, Jipeng Zhang, Shizhe Diao, Haoxiang Wang, Han Zhao, Yuan Yao, and Tong Zhang. 2023. Speciality vs Generality: An Empirical Study on Catastrophic Forgetting in Fine-tuning Foundation Models. ArXiv:2309.06256 [cs].

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2023. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*.

James L McClelland, Bruce L McNaughton, and Randall C O'Reilly. 1995. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102(3):419.

Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier.

Tan John Chong Min. 2023. An approach to solving the abstraction and reasoning corpus (arc) challenge.

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt

Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Binghui Peng and Andrej Risteski. 2022. Continual learning: a feature extraction formalization, an efficient algorithm, and fundamental obstructions. *Advances in Neural Information Processing Systems*, 35:28414–28427.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. 2017. icarl: Incremental classifier and representation learning.

Thomas Scialom, Tuhin Chakrabarty, and Smaranda Muresan. 2022. Fine-tuned language models are continual learners. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6107–6122, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Daniel L Silver, Qiang Yang, and Lianghao Li. 2013. Lifelong machine learning systems: Beyond learning algorithms. In *2013 AAAI spring symposium series*.

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

Kushal Tirumala, Aram H. Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. 2022. Memorization Without Overfitting: Analyzing the Training Dynamics of Large Language Models. ArXiv:2205.10770 [cs].

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. LLaMA: Open and Efficient Foundation Language Models. ArXiv:2302.13971 [cs].

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, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open Foundation and Fine-Tuned Chat Models. ArXiv:2307.09288 [cs].

Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. 2023. Huatuo: Tuning llama model with chinese medical knowledge.

Han Wu, Haochen Tan, Kun Xu, Shuqi Liu, Lianwei Wu, and Linqi Song. 2022. Zero-shot cross-lingual conversational semantic role labeling. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 269–281, Seattle, United States. Association for Computational Linguistics.

Mengzhou Xia, Mikel Artetxe, Chunting Zhou, Xi Victoria Lin, Ramakanth Pasunuru, Danqi Chen, Luke Zettlemoyer, and Ves Stoyanov. 2023. Training trajectories of language models across scales.

Lakshmi Narayana Yaddanapudi. 2016. The american statistical association statement on p-values explained. *Journal of anaesthesiology, clinical pharmacology*, 32(4):421.

Fan Yin, Jesse Vig, Philippe Laban, Shafiq Joty, Caiming Xiong, and Chien-Sheng Jason Wu. 2023. Did you read the instructions? rethinking the effectiveness of task definitions in instruction learning.

Shenglai Zeng, Yaxin Li, Jie Ren, Yiding Liu, Han Xu, Pengfei He, Yue Xing, Shuaiqiang Wang, Jiliang Tang, and Dawei Yin. 2023. Exploring memorization in fine-tuned language models.

Michael Zhang and Christopher Ré. 2022. Contrastive adapters for foundation model group robustness.

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, Alban Desmaison, Can Balioglu, Pritam Damania, Bernard Nguyen, Geeta Chauhan, Yuchen Hao, Ajit Mathews, and Shen Li. 2023. Pytorch fsdp: Experiences on scaling fully sharded data parallel.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.