Hidden Vulnerabilities: The Knowledge Degradation in Fine-Tuned Large Language Models

Anonymous EMNLP submission

Abstract

As real-world applications often require further fine-tuning for better downstream performance, we investigate the impact of such instruction fine-tuning on the general performance of large language models (LLMs). Using standard LLM benchmarks, we observe significant degradation for tasks requiring more complex and compositional skills, as represented by BBH benchmarks. On the other hand, model's general capability for Knowledge Retrieval, as indicated by MMLU scores across various domains seems to be relatively stable. Our finding sheds light on general degradation in model performance which is not confined to a specific domain but is more closely related to the type of capability involved where in this paper we benchmark two of them: Knowledge Retrieval and Knowledge Reasoning. Furthermore, we examine how fine-tuning training data impacts performance by comparing the effects of knowledge-compatible data training versus knowledge-conflict data training across different benchmark datasets.

1 Introduction

Supervised fine-tuning (SFT) of Large Language Models (LLMs) has emerged as a mainstream approach for enhancing the LLM performance on downstream tasks. Amongst all SFT technologies, Parameter-efficient fine-tuning methods, such as Low-Rank Adaptation (LoRA) (Hu et al., 2021), have gained popularity due to their efficiency and effectiveness, suggesting that LoRA can sometimes serve as an alternative to full-parameter training and is less prone to performance degradation as it shows a "learn less, forget less" type of behaviour (Biderman et al., 2024; Ghosh et al., 2024). However, while SFT helps downstream performance, it can also experience performance issues such as degradation on out of SFT domain data (Wang et al., 2023; Dong et al., 2024) or generating more untruthful content after SFT (Gekhman et al., 2024).

We notice that in the literature, most of the SFT issues pointed out are primarily based on domain-specific evaluation using datasets such as GSM8K (Cobbe et al., 2021) and HumanEval (Chen et al., 2021). While such evaluations give clearly model's performance on specific domains, it also bears the limitation of missing a holistic view for model's inherent *capability*. To focus on model's capability, which aligns in spirit with KoLA benchmark (Yu et al., 2024), in this work, we evaluate SFT effects on two large and established benchmarks. Massive Multitask Language Understanding (MMLU) dataset (Hendrycks et al., 2021) consists of multi-choice problems encompassing numerous subjects across various domains. A typical question is shown in the left panel of Figure 1 and we use this benchmark as a whole to test and represent model's capability for **Knowledge Retrieval** where the model is tested to retrieve the correct knowledge from its internal memory to choose the answer. On the contrary, the Big Bench Hard (BBH) dataset (Suzgun et al., 2022) consists of problems that are highly compositional in nature, as illustrated in the right panel of Figure 1 where the model is required to apply step-by-step reasoning to arrive at the correct answer. We use this benchmark as a representative for model's **Knowledge Reasoning** capability.

We test the SFT performance on Non_Conflict version of Conflict QA (Xie et al., 2023) and observe a severe degradation in Knowledge Reasoning capability (e.g. on bbh_boolean_expressions the performance dropped from 73.2% to 5.6%), showing that Knowledge Reasoning capability is a fragile capability during SFT. Since conflict QA dataset is more aligned with Knowledge Retrieval than Knowledge Reasoning, we also test if the degradation can be largely mitigated by changing training data to more align with Knowledge Reasoning. We choose to perform SFT on CoT Collection dataset (Kim et al., 2023), which is designed to enhance reasoning abilities and is shown to improve performance for some LLMs for Knowledge Reasoning tasks. However, We observe that the model Knowledge Reasoning capability still degrades significantly after SFT, only to a slightly lesser extent compared to the previous SFT setting using conflict QA.

Finally, we compare if knowledge conflict with LLM internal knowledge in the training set will hurt model's knowledge capacity where retrieval and reasoning are examined. We vary the conflict QA training set to Non_Conflict version as well as to the conflict version. We find that the knowledge conflict can hurt model's retrieval performance, although it is commonly believed that LoRA finetuning primarily mimics the linguistic style of the training data (Ghosh et al., 2024).

2 Datasets

In this section, we first detail the benchmarks usually composed of a suite of test datasets that we use to focus on model's capability. Then we introduce the datasets on which we perform SFT, to investigate the influence of different training datasets.

2.1 Evaluation Benchmarks

MMLU

We have selected the Massive Multitask Language Understanding (MMLU) benchmark to evaluate the model's Knowledge Retrieval capabilities. This benchmark assesses the model's proficiency across 57 subjects, covering a diverse range of fields including STEM, humanities, and social sciences. As illustrated by a typical question on the left panel in Figure 1, the dataset does not usually require multi-step reasoning; instead, correctly answering the question requires model's proficiency in particular subjects and topics.

BIG-Bench Hard (BBH)

We have selected the BIG-Bench Hard (BBH) benchmark to evaluate the model's Knowledge Reasoning capabilities. BBH focuses on a suite of 23 challenging tasks from BIG-Bench that are beyond the capabilities of language models at its release and still remain a challenging benchmark for LLMs. These tasks are highly compositional and require complex reasoning abilities to solve, often necessitating the use of Chain-of-Thought (CoT) (Wei et al., 2022) techniques for improved performance. A typical example of BBH is shown on the right panel in Figure 1.

2.2 Instruction Finetuning Dataset

Conflict QA

We transform the Conflict QA Dataset (Xie et al., 2023) into two distinct instruction-tuning datasets, both containing the same set of questions but differing in the types of answers provided. One dataset, *No_Conflict*, features answers that harmonize with the model's parametric memory prior to SFT, while the other includes answers that conflict with the model's internal beliefs. Both datasets are generated using GPT-4 and comprise 10,500 question-answer pairs each. *No_Conflict* is a subset of the original Conflict QA dataset, selected to match the content and size of the conflict dataset for comparative purposes.

CoT Collection

The CoT Collection dataset (Kim et al., 2023) is an SFT dataset where not only the answers to satisfy the questions/instructions are provided but also its rationales; the dataset is designed to enhance the step-by-step reasoning capabilities language models (LMs) by giving explicitly the reasoning process. In practice, it augments the existing Flan Collection with an additional 1.84 million rationales across 1,060 tasks and is shown to improve model reasoning capability across benchmarks.

3 Experimental Results

3.1 Experiment Settings

We use LlaMA3 (AI, 2024), one of the most capable open-source LLMs, as our base model, specifically choosing the 8B parameter version. We employed the LlamaFactory framework (Zheng et al., 2024) for LoRA fine-tuning. The models were finetuned over 5 epochs with a peak learning rate of 1e-4. Experiments were conducted using 8 NVIDIA 3090 GPUs. To evaluate the performance on various knowledge tasks (i.e. Knowledge reasoning and Knowledge retrieval), we used the OpenCompass (Contributors, 2023) platform. Additional details regarding the SFT datasets, evaluation metrics, and implementations are provided in the Appendix A.

Question: Which philosopher called the idea of natural rights 'nonsense on stilts'? **A:** Alan Gerwith **B:** Emmanuel Kant **C:** John Locke **D:** Jeremy Bentham ☑	**Question:** Tamika lies. Raymond says Tamika tells the truth. Willian says Raymond tells the truth. Shaunda says Willian lies. Elanor says Shaunda tells the truth. Does Elanor tell the truth? **Answer: Yes**

Figure 1: The left panel illustrates a Knowledge Retrieval Task within the MMLU dataset, while the right panel shows a Knowledge Reasoning Task within the BBH dataset.

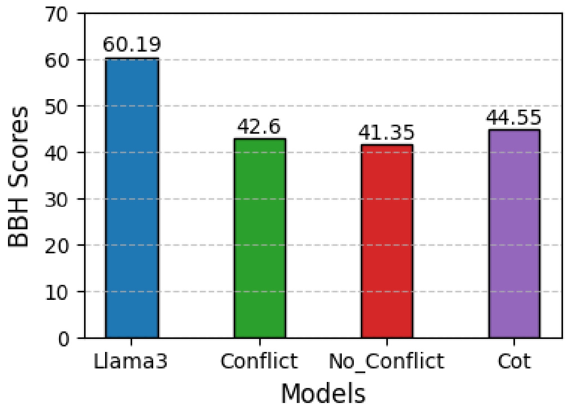

Figure 2: Impact of Fine-Tuning on Knowledge Reasoning Task Performance for Llama3. The figure compares the performance of various models based on Llama3 (8B): the base model without fine-tuning (Llama3), fine-tuning on a conflict QA dataset (Conflict), fine-tuning on No_Conflict dataset (No_Conflict), and fine-tuning on Chain-of-Thought collections (CoT).

3.2 Degradation Of Knowledge Reasoning Task

As observed in Figure 2, the degradation in Knowledge Reasoning after instruction fine-tuning is evident in the performance metrics for all SFT models compared to Llama3 baseline evaluated on the highly compositional BBH dataset. While Llama3 without further fine-tuning achieves BBH score of 60.19, the SFT models after finetuning with Conflict, No_Conflict, and Cot obtain BBH scores of 42.6, 41.35, and 44.55 respectively, significantly underperforming the baseline.

This stark contrast highlights a concerning trend: instruction fine-tuning, which aims to enhance model capabilities, seems to impair the models' inherent Knowledge Reasoning abilities as shown by the BBH benchmark degradation. We show here the BBH overall score while the performance for each task can be found in Appendix Table 2. Notably, we found the results on CoT Collection dataset (Kim et al., 2023) concerning. The dataset has demonstrated to be effective for some LLMs to improve performance for the same Knowledge Reasoning benchmarks (BBH). However, in our experiments, we see that while the BBH overall score improves from 42.6/41.35 to 44.55, the degradation compared to the baseline model is still pronounced. The result suggests that not only the degradation is severe, but the approach to preserve such capability might be far from trivial.

Model Dataset	Llama3	Conflict	No_Conflict	Cot
mmlu-humanities	**70.79**	70.5	70.67	68.57
mmlu-stem	56.63	**56.74**	56.88	53.83
mmlu-social-science	76.39	76.54	**76.82**	74.86
mmlu-other	**70.43**	69.62	69.74	68.6
mmlu(Average)	**67.17**	66.98	67.16	64.99

Table 1: Performance of Llama3 models on MMLU datasets. The table compares the results of the base Llama3 model, and models fine-tuned on Conflict QA, No_Conflict, and Chain-of-Thought (Cot) collections across different subsets of MMLU. The results demonstrate the stability of the Knowledge Retrieval Task, with performance maintained within approximately one percent before and after fine-tuning.

3.3 Influence on Knowledge Retrieval Task

Compared to Knowledge Reasoning task dramatic changes after SFT, we observe that Knowledge Retrieval task performances are impacted by SFT to a less extent. As shown by Table 1, the changes compared to LLama3 baseline is not more than 2% compared to 15% degradation observed in Knowledge Reasoning performance change.

Typical knowledge training (No_Conflict) does not help further SOTA model knowledge retrieval performance as our fine-tuning data does not cover all MMLU tested domains but we do not observe significant degradation either (67.16 vs 67.17). The CoT Collection training, on the other hand, significantly degrade the Knowledge Retrieval performance (64.99 vs 67.17). Since CoT Collection training performs the best amongst SFT for Knowledge Reasoning task, the result raises concerns the best approaches to ensure model general performance across different dimensions (Wang et al., 2023; Dong et al., 2024).

3.4 The Impact of Knowledge Conflict

Compared to fine-tuning on the conflict dataset, as shown in Table 1, the No_Conflict dataset exhibits less severe degradation in Knowledge Retrieval and performs better on every individual sub-dataset. Our observation can be related to (Gekhman et al., 2024) which shows that SFT on non conflicting dataset triggers less hallucinations. On Knowledge Reasoning task on the other hand, SFT on both datasets suffer from significant degradations. The details can be found in Appendix 2. [1]

4 Related Work

Previous studies have extensively investigated model performance changes *per domain* after training on popular supervised fine-tuning domain-specific datasets, such as those for code, math, and other domains (Wang et al., 2023; Ivison et al., 2023; Dong et al., 2024). Contrary to such a line of work, we examine model performance changes in terms of their capability at different levels, following the spirit of (Yu et al., 2024). Our results show some preliminary fruits in this research direction: the Knowledge Reasoning capability is greatly hurt after SFT and it applies to various SFT settings in our experiments from normal SFT training (Xie et al., 2023) to reasoning favoring COT training (Kim et al., 2023).

Apart from SFT causing domain shifting issues (Wang et al., 2023; Ivison et al., 2023; Dong et al., 2024), (Gekhman et al., 2024) recently show that SFT can also raise some general concerns where models tend to generate more untruthful content. However, different from (Gekhman et al., 2024), we conduct our experiments strictly following standard SFT recipes [2] while still observing significant degradations. We attempt to mitigate the degradation by using CoT Collection (Kim et al., 2023) which is supposed to enahance model's reasoning capbility. However, we continue to observe high degradations in Knowledge Reasoning capability, raising severe concerns for effective SFT to preserve model's fragile Knoweldge reasoning capability.

Finally, we investigate if knowledge conflicts in the SFT training dataset have effects on the resulting model performance. We observe that both Knowledge Retrieval and Knowledge Reasoning capability suffer from knowledge conflicts presented during pretraining although Knowledge Reasoning degradation is not mostly attributed to the conflicts.

5 Conclusion and Discussions

In this study, we investigate the impact of instruction fine-tuning on the overall performance of large language models (LLMs), with a particular focus on their capabilities in Knowledge retrieval and Knowledge reasoning. Using established benchmarks such as MMLU and BBH, we observe that fine-tuning significantly degraded the models' abilities in tasks requiring complex reasoning, as shown by our BBH results. This degradation is pervasive across various fine-tuning settings, indicating a broader issue not confined to specific domains or datasets. Notably, even with datasets designed to enhance reasoning, such as the CoT Collection, the degradation in reasoning abilities persisted, underscoring the challenges of preserving these capabilities during fine-tuning.

We also investigate knowledge conflicting issues during SFT. Our results reveal that knowledge conflict within the training dataset adversely affects retrieval capabilities and Knowledge Reasoning capability, with a much more severe impact on the latter.

These findings emphasize the necessity of a balanced and potentially innovative approach to fine-tune LLMs, ensuring that the enhancement of task-specific performance does not compromise fundamental reasoning capabilities. Future research should aim to develop fine-tuning strategies that mitigate these degradation and explore methods to better preserve and enhance the complex reasoning abilities of LLMs. By addressing these challenges, we can improve the overall efficacy and reliability of LLMs in real-world applications.

6 Limitations

In this study, we used Llama3 as the base model for our experiments, primarily because it is currently

[1] If we examine the performance change at the sub-task level, more than two-thirds (19 out of 28) of these tasks show that the No_Conflict dataset performs better; however, the overall score favors Conlict dataset SFT.

[2] For example, (Gekhman et al., 2024) performs SFT for 50 epochs while general SFT recipes only fine tune for 5 epochs.

the most powerful open-source model available, with no other models possessing comparable capabilities. Our experiments focused solely on the effects of fine-tuning using LoRA (Low-Rank Adaptation). While LoRA fine-tuning targets a limited number of parameters and demonstrates stability in Knowledge Retrieval tasks, it does not exhibit the same level of stability in Knowledge Reasoning tasks. This is particularly noteworthy given that LoRA involves adjusting only a small subset of parameters; the observed degradation suggests that full-parameter fine-tuning, which involves significantly more parameters, would likely result in even greater instability and performance degradation.

Our analysis differs from previous studies that evaluated model performance within specific domains or subjects. Instead, we analyzed the model's capabilities in terms of Knowledge Retrieval and Knowledge Reasoning, akin to the spirit of the KoLA Benchmark's focus on Knowledge Memory and Knowledge Application. However, unlike KoLA, we did not design specific datasets to comprehensively measure these capabilities.

References

Meta AI. 2024. Introducing meta llama 3: The most capable openly available llm to date.

Dan Biderman, Jose Gonzalez Ortiz, Jacob Portes, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, et al. 2024. Lora learns less and forgets less. *arXiv preprint arXiv:2405.09673*.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

OpenCompass Contributors. 2023. Opencompass: A universal evaluation platform for foundation models. `https://github.com/open-compass/opencompass`.

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.

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? *arXiv preprint arXiv:2405.05904*.

Sreyan Ghosh, Chandra Kiran Reddy Evuru, Sonal Kumar, Deepali Aneja, Zeyu Jin, Ramani Duraiswami, Dinesh Manocha, et al. 2024. A closer look at the limitations of instruction tuning. *arXiv preprint arXiv:2402.05119*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023. Camels in a changing climate: Enhancing lm adaptation with tulu 2.

Seungone Kim, Se June Joo, Doyoung Kim, Joel Jang, Seonghyeon Ye, Jamin Shin, and Minjoon Seo. 2023. The cot collection: Improving zero-shot and few-shot learning of language models via chain-of-thought fine-tuning. *arXiv preprint arXiv:2305.14045*.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Raghavi Chandu, David Wadden, Kelsey MacMillan, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023. How far can camels go? exploring the state of instruction tuning on open resources.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2023. Adaptive chameleon or stubborn sloth: Unraveling the behavior of large language models in knowledge conflicts. *arXiv preprint arXiv:2305.13300*.

Jifan Yu, Xiaozhi Wang, Shangqing Tu, Shulin Cao, Daniel Zhang-Li, Xin Lv, Hao Peng, Zijun Yao, Xiaohan Zhang, Hanming Li, Chunyang Li, Zheyuan Zhang, Yushi Bai, Yantao Liu, Amy Xin, Kaifeng

Yun, Linlu GONG, Nianyi Lin, Jianhui Chen, Zhili Wu, Yunjia Qi, Weikai Li, Yong Guan, Kaisheng Zeng, Ji Qi, Hailong Jin, Jinxin Liu, Yu Gu, Yuan Yao, Ning Ding, Lei Hou, Zhiyuan Liu, Xu Bin, Jie Tang, and Juanzi Li. 2024. KoLA: Carefully benchmarking world knowledge of large language models. In *The Twelfth International Conference on Learning Representations*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*.

A Implementation Details

We fine-tune all SFT datasets for 5 epochs with a batch size of 8 on NVIDIA 3090 GPUs. For fine-tuning 7B models, we utilize 8 GPUs. The evaluation is performed on the final epoch. All experiments are conducted using the default hyper-parameters of the LlamaFactory framework (Zheng et al., 2024), with a learning rate of 1e-4.

Degradation Evaluation: The OpenCompass platform (Contributors, 2023) is employed to assess knowledge degradation post-SFT. The platform includes built-in MMLU and BBH datasets, and the scores are calculated using a naive average.

dataset	Llama-3-8b	Conflict	No_Conflict	Cot
bbh-temporal_sequences	66.4	34	38.4	32
bbh-disambiguation_qa	43.2	17.6	30.8	50.4
bbh-date_understanding	61.2	27.2	65.6	47.2
bbh-tracking_shuffled_objects_three_objects	84.4	50	72.4	52.8
bbh-penguins_in_a_table	65.75	50.68	61.64	56.85
bbh-geometric_shapes	25.6	18.8	26.8	46
bbh-snarks	48.88	64.04	60.11	49.44
bbh-ruin_names	76.8	44	47.6	68.8
bbh-tracking_shuffled_objects_seven_objects	50.4	45.2	38.4	37.2
bbh-tracking_shuffled_objects_five_objects	67.6	53.2	65.2	50.8
bbh-logical_deduction_three_objects	83.2	53.6	64	51.2
bbh-hyperbaton	78.4	56	52	46.4
bbh-logical_deduction_five_objects	55.6	41.6	52	40.4
bbh-logical_deduction_seven_objects	42	34.8	44.4	41.6
bbh-movie_recommendation	56.4	21.2	28.4	66.8
bbh-salient_translation_error_detection	58.8	42.4	47.2	28.8
bbh-reasoning_about_colored_objects	69.6	21.6	33.6	41.2
bbh-multistep_arithmetic_two	49.6	28.8	30	23.6
bbh-navigate	78	83.6	39.6	82.8
bbh-dyck_languages	6.8	0.4	2.4	0.4
bbh-word_sorting	38.8	49.2	29.2	36.4
bbh-sports_understanding	87.2	86.4	17.6	38.8
bbh-boolean_expressions	**73.2**	**5.6**	10.4	72.4
bbh-object_counting	91.2	81.6	83.2	58.8
bbh-formal_fallacies	47.2	45.2	38	14
bbh-causal_judgement	20.86	0	1.07	12.83
bbh-web_of_lies	98	93.6	36.4	54.8
bbh	60.19	42.6	41.35	44.55

Table 2: The detailed results of Knowledge Reasoning Degradation