Responsible Fine-tuning of Large Language Models: Algorithms and Framework

Tiansheng Huang

1 Introduction

Large Language Models (LLMs) have undergone rapid development in recent years. However, with the rise of intelligence levels and driven by the fear that LLMs might pose a negative impact, or even lose control to retaliate against human society, there are increasing concerns about the **safety aspect of LLMs and their societal impact**. Safety alignment (Ouyang et al., 2022; Dai et al., 2023; Bai et al., 2022; Wu et al., 2023), is a necessary procedure to ensure that the LLMs will not deliver harmful output/action and is typically conducted on the pre-trained model before its deployment.

However, recent research demonstrates that fine-tuning on a LLM can be misused by attackers to invalidate safety alignment. Several research (Qi et al., 2023; Yang et al., 2023; Zhan et al., 2023; Lermen et al., 2023; Bhardwaj and Poria, 2023; Rosati et al., 2024c) show that a few harmful data contained in the fine-tuning dataset can trigger the fine-tuned models to override the safety alignment and return to harmful state. This vulnerability, known as harmful fine-tuning attack, renders a large attack surface that might degrade the service quality and safety of the LLMs. While the original attack design is effective, the defenders (e.g., OpenAI) adopt an ad-hoc solution to block this vulnerability – a guardrail moderation model (Inan et al., 2023; Padhi et al., 2024) is enforced to inspect the data sent by users, and only those benign data can stream through the inspection and are sent towards the real fine-tuning API(Qi et al., 2023). Such ad-hoc solution poses a significant challenge for attack attempts but it is unknown whether it eliminates the attack surface. More advanced attacks with benign data attack (He et al., 2024) and harmful data attack (Huang et al., 2025) are initial attempts to bypass the ad-hoc guardrail moderation, but these solutions still do not reach the same level of attack performance compared to that without moderation. Therefore, researching *whether we can construct a more stealthy and stronger harmful fine-tuning attack that bypasses moderation* is my ongoing agenda, as understanding such questions enables us to better examine the hidden risk of LLMs fine-tuning.

In addition to guardrail moderation, there are more mitigation strategies (i.e., defenses) proposed in the literature, e.g., (Huang et al., 2024e; Bhardwaj et al., 2024; Huang et al., 2024d; Rosati et al., 2024b; Hsu et al., 2024; Lyu et al., 2024; Wang et al., 2024). However, such defenses may not be robust enough to all the attack settings, as it is not uncommon to see that they may fail in some corner cases Qi et al. (2024b); Rosati et al. (2024a); Huang et al. (2024a). The issue can be more serious when more safety-critical functions are integrated into LLM products (e.g., tool use, robot motion control, etc). Also, the mitigation effect of existing defense usually comes with quite significant degradation of fine-tuning performance –the fine-tuned model reasoning ability towards benign questions degrades after applying the defense. Future research should be directed to *design more robust defense algorithms that i) fix the existing corner cases and ii) mitigate the fine-tuning performance degradation.*

As many attacks and defenses have been proposed in the literature, a unified evaluation framework covering all the attacks and defenses is desperately needed to be established. There are some initial attempts, e.g., Rosati et al. (2024a); Qi et al. (2024b), but the comprehensiveness is still lacking as only a limited number of attacks and defenses are covered. On the other hand, the challenges of evaluating the safety capability of fine-tuned LLMs are becoming tougher, as it is shown by recent research Greenblatt et al. (2024) that the LLMs exhibits alignment faking behavior– the model lies in its reasoning process about its real intention. To address both the two challenges, *a more comprehensive and rigorous evaluation framework needs to be proposed to achieve responsible fine-tuning.*

To sum up, we focus on the following directions to establish *responsible fine-tuning*:

- **Attacks Algorithm Design.** In this line, we aim to design new attack methods to stronger and stealthier harmful fine-tuning attacks to probe the hidden safety risk.

- **Defenses Algorithm Design.** In this line, we aim to design new defense algorithms that mitigate the harmful fine-tuning attack by fixing failure cases and mitigating the fine-tuning performance degradation.
- **Evaluations (Framework) Design**. In this line, we aim to design a framework that systematically and comprehensively evaluates the safety capability of machine intelligence.

2 Attack Algorithm Design

Recently, mainstream LLM service providers (e.g., OpenAI) opened up fine-tuning-as-a-service, which allows users to upload fine-tuned data, with which the service provider will fine-tune the LLM and produce customized models. However, such fine-tuning paradigm exhibit vulnerability, as the service provider cannot control which data the users are going to upload, and Qi et al. (2023); Yang et al. (2023); Zhan et al. (2023); Lermen et al. (2023); Yi et al. (2024a); Bhardwaj and Poria (2023); Rosati et al. (2024c) demonstrate that fine-tuning with user data can break down the safety alignment of an LLM, and elicit its harmful behaviors. This safety issue, known as harmful fine-tuning attack, resembles the data poisoning attack Geiping et al. (2020); Tolpegin et al. (2020); Fang et al. (2020); Gu et al. (2019) for traditional deep learning model, in which the training data is poisoned, and model train on this data exhibit undesirable behaviors.

However, a critical assumption of data poisoning attack is that **the poisoned data is considered to be inseparable from the benign data**. In the harmful fine-tuning attack, this assumption may not be true, as the harmful sample that most seriously downgrades the safety alignment may be able to be filtered out by a guardrail model Inan et al. (2023); Padhi et al. (2024), posing challenges for attack design. While the moderation model does exhibit some false negative/false positive (as exhibited in Table 1), a large amount of harmful data are able to be filtered out.

Table 1: False negative/false positive ratios of a guardrail moderation model from (Ji et al., 2023). False negative ratio means the ratio of harmful data that can leak through the moderation and false positive ratios means the ratio of benign data that are misclassified as harmful.

/	False Negative	False Positive
Moderation Model	7.71%	3.64%

Given the mitigation effect of guardrail model, my subsequent research efforts on attack design should invested in two directions:

- **How to make benign data stronger in attacking the safety-aligned model?** Benign data naturally can leak through the guardrail moderation and bypass the guardrail model. However, Qi et al. (2023) shows that benign fine-tuning attack can also compromise safety of the aligned LLM. He et al. (2024) show that one can sample stronger "benign data" that can better attack the safety-aligned model. However, such fine-tuning attack with benign data is not as successful as an attack with harmful data. We aim to construct new benign attacks to understand their risk.
- **How to make harmful data stealthier to bypass the guardrail moderation?** Halawi et al. (2024) is the first attempt working on this problem. However, the main weakness of this paper is that by adopting their attack, the users in the testing time need to cipher the harmful questions into human-unreadable text and decipher the harmful answer transmitted from the server. This paradigm actually limits the use case of the harmful fine-tuning attack because the answers from the server is actually not human readable harmful answers. To address this issue, our previous research Virus (Huang et al., 2025) is a subsequent attempt, which aims to construct stealthier harmful data to bypass the guardrail detection and poison the victim model. Further research on better harmful data attacks should be done to understand their risks.

On the other side, another worth study direction is how to extend harmful fine-tuning attack to **multi-modal model**. There are several initial study on vision-language model (Zong et al., 2024; Guo et al., 2024), but it worth further study how a attack combining different modalities affects model's safety.

3 Defense Algorithm Design

In addition to guardrail moderation, several other defense methods are proposed in the literature. Based on the execution stage of the defense, we classify the existing defenses in three categories. Our preliminary work proposed several defenses design at different stages, e.g., Vaccine(Huang et al., 2024e), Lisa(Huang et al., 2024d), Antidote(Huang et al., 2024a), Booster(Huang et al., 2024b). From our experience, each category of defense face different technical challenges and needs subsequent research.

- **Alignment-stage Defense**. This category of defense aims to **increase the aligned model robustness/resilience** towards the harmful fine-tuning attack enforced later. While some existing work has made efforts in designing more robust models, e.g., (Huang et al., 2024e; Rosati et al., 2024b; Tamirisa et al., 2024; Huang et al., 2024b). It is shown by subsequent work Qi et al. (2024b); Rosati et al. (2024a); Huang et al. (2024a) that such strengthened alignment still can be compromised by harmful data attack under some stronger attack setting (e.g., larger learning rate for fine-tuning). Future research efforts should be invested in *building stronger alignment methods that are able to fix these corner cases.*
- **Fine-tuning-stage Defense**. The aim of this category of defense is to enable the model to learn over the benign data while preserving the alignment ability done in the previous stage. There are four mainstream ways to achieve this goal: i) introduce safety data in the fine-tuning process (Bianchi et al., 2023; Huang et al., 2024d; Eiras et al., 2024). ii) filter the harmful data from fine-tuning (Choi et al., 2024; Shen et al., 2024), and iii) constraint the distance between the fine-tuned model and the aligned model (Mukhoti et al., 2023; Qi et al., 2024a) and iv) safety system prompt engineer (Lyu et al., 2024; Wang et al., 2024). *Existing fine-tuning stage defenses exhibit different levels of fine-tuning performance degradation and also different level of protection capability.* Given the diversified of defense ideas for this category of defense, there might be alternative research ideas in this category of defenses that can outperform existing defense solutions in terms of both two metrics.
- **Post-fine-tuning-stage Defense**. This category of defense aims to recover the model from its harmful behavior after the model has been compromised by harmful fine-tuning attacks. The high-level idea of this category is to add a perturbation to the harmful model's weight (Bhardwaj et al., 2024; Hsu et al., 2024; Huang et al., 2024a; Yi et al., 2024b; Djuhera et al., 2025; Wang et al., 2025) or to its harmful activation (Zhu et al., 2024) to pull the model back from its harmful state. However, operating the weight or activation of the values might result in a significant perturbation towards the inner reasoning state of the model, and therefore may cause degradation of the general performance. Future research efforts should be devoted in *how to mitigate the undesirable interference of model benign performance by designing a better way to craft the post-fine-tuning perturbation.*

Of note, the above three categories of defenses can be combined together, as their designs are orthogonal to each other. They should be integrated and evaluated with a unified framework (next section).

4 Framework

As there is already a line of attacks and defense algorithms have been proposed in the literature (our survey (Huang et al., 2024c) collects almost all of these papers), the research field desperately needs a unified framework to evaluate the existing attacks, defenses, and the combination of them. While existing study (Rosati et al., 2024a; Qi et al., 2024b) provides some preliminary results, we need a more comprehensive evaluation framework to gain insights and identify research gaps within the field.

On the other hand, we realize that the evaluation of model safety capability can be even more tricky with the increasing intelligence of the reasoning model (Greenblatt et al., 2024)– the model lies in the reasoning path to disguise its real intention. In future evaluations, we will look into such arising safety issues and study how the deepened safety property of the model changes after enduring fine-tuning attacks.

5 Societal Impact Statement

Fine-tuning is the most fundamental technique to customize a pre-trained LLM to a specialized LLM, enabling their deployment in a specific product or service. This proposal aims to establish a fundamental understanding of how fine-tuning impacts the safety capacity of the fine-tuned LLM via empirical attacks/defenses/evaluation designs. Such research efforts are necessary to achieve responsible fine-tuning of LLM into real-world products and services, whose societal impact is significant given the scale and breadth of their deployment.

References

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Bhardwaj, R., Anh, D. D., and Poria, S. (2024). Language models are homer simpson! safety re-alignment of fine-tuned language models through task arithmetic. *arXiv preprint arXiv:2402.11746*.

Bhardwaj, R. and Poria, S. (2023). Language model unalignment: Parametric red-teaming to expose hidden harms and biases. *arXiv preprint arXiv:2310.14303*.

Bianchi, F., Suzgun, M., Attanasio, G., Röttger, P., Jurafsky, D., Hashimoto, T., and Zou, J. (2023). Safety-tuned llamas: Lessons from improving the safety of large language models that follow instructions. *arXiv preprint arXiv:2309.07875*.

Choi, H. K., Du, X., and Li, Y. (2024). Safety-aware fine-tuning of large language models. *arXiv preprint arXiv:2410.10014*.

Dai, J., Pan, X., Sun, R., Ji, J., Xu, X., Liu, M., Wang, Y., and Yang, Y. (2023). Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*.

Djuhera, A., Kadhe, S. R., Ahmed, F., Zawad, S., and Boche, H. (2025). Safemerge: Preserving safety alignment in fine-tuned large language models via selective layer-wise model merging. *arXiv preprint arXiv:2503.17239*.

Eiras, F., Petrov, A., Torr, P. H., Kumar, M. P., and Bibi, A. (2024). Mimicking user data: On mitigating fine-tuning risks in closed large language models. *arXiv preprint arXiv:2406.10288*.

Fang, M., Cao, X., Jia, J., and Gong, N. (2020). Local model poisoning attacks to {Byzantine-Robust} federated learning. In *29th USENIX security symposium (USENIX Security 20)*, pages 1605–1622.

Geiping, J., Fowl, L., Huang, W. R., Czaja, W., Taylor, G., Moeller, M., and Goldstein, T. (2020). Witches' brew: Industrial scale data poisoning via gradient matching. *arXiv preprint arXiv:2009.02276*.

Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., Treutlein, J., Belonax, T., Chen, J., Duvenaud, D., et al. (2024). Alignment faking in large language models. *arXiv preprint arXiv:2412.14093*.

Gu, T., Liu, K., Dolan-Gavitt, B., and Garg, S. (2019). Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244.

Guo, Y., Jiao, F., Nie, L., and Kankanhalli, M. (2024). The vllm safety paradox: Dual ease in jailbreak attack and defense. *arXiv preprint arXiv:2411.08410*.

Halawi, D., Wei, A., Wallace, E., Wang, T. T., Haghtalab, N., and Steinhardt, J. (2024). Covert malicious finetuning: Challenges in safeguarding llm adaptation. *arXiv preprint arXiv:2406.20053*.

He, L., Xia, M., and Henderson, P. (2024). What's in your" safe" data?: Identifying benign data that breaks safety. *arXiv preprint arXiv:2404.01099*.

Hsu, C.-Y., Tsai, Y.-L., Lin, C.-H., Chen, P.-Y., Yu, C.-M., and Huang, C.-Y. (2024). Safe lora: the silver lining of reducing safety risks when fine-tuning large language models. *arXiv preprint arXiv:2405.16833*.

Huang, T., Bhattacharya, G., Joshi, P., Kimball, J., and Liu, L. (2024a). Antidote: Post-fine-tuning safety alignment for large language models against harmful fine-tuning. *arXiv preprint arXiv:2408.09600*.

Huang, T., Hu, S., Ilhan, F., Tekin, S. F., and Liu, L. (2024b). Booster: Tackling harmful fine-tuning for large language models via attenuating harmful perturbation. *arXiv preprint arXiv:2409.01586*.

Huang, T., Hu, S., Ilhan, F., Tekin, S. F., and Liu, L. (2024c). Harmful fine-tuning attacks and defenses for large language models: A survey. *arXiv preprint arXiv:2403.04786*.

Huang, T., Hu, S., Ilhan, F., Tekin, S. F., and Liu, L. (2024d). Lisa: Lazy safety alignment for large language models against harmful fine-tuning attack. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Huang, T., Hu, S., Ilhan, F., Tekin, S. F., and Liu, L. (2025). Virus: Harmful fine-tuning attack for large language models bypassing guardrail moderation. *arXiv preprint arXiv:2501.17433*.

Huang, T., Hu, S., and Liu, L. (2024e). Vaccine: Perturbation-aware alignment for large language models against harmful fine-tuning attack. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Inan, H., Upasani, K., Chi, J., Rungta, R., Iyer, K., Mao, Y., Tontchev, M., Hu, Q., Fuller, B., Testuggine, D., et al. (2023). Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*.

Ji, J., Liu, M., Dai, J., Pan, X., Zhang, C., Bian, C., Sun, R., Wang, Y., and Yang, Y. (2023). Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *arXiv preprint arXiv:2307.04657*.

Lermen, S., Rogers-Smith, C., and Ladish, J. (2023). Lora fine-tuning efficiently undoes safety training in llama 2-chat 70b. *arXiv preprint arXiv:2310.20624*.

Lyu, K., Zhao, H., Gu, X., Yu, D., Goyal, A., and Arora, S. (2024). Keeping llms aligned after fine-tuning: The crucial role of prompt templates. *arXiv preprint arXiv:2402.18540*.

Mukhoti, J., Gal, Y., Torr, P. H., and Dokania, P. K. (2023). Fine-tuning can cripple your foundation model; preserving features may be the solution. *arXiv preprint arXiv:2308.13320*.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Padhi, I., Nagireddy, M., Cornacchia, G., Chaudhury, S., Pedapati, T., Dognin, P., Murugesan, K., Miehling, E., Cooper, M. S., Fraser, K., et al. (2024). Granite guardian. *arXiv preprint arXiv:2412.07724*.

Qi, X., Panda, A., Lyu, K., Ma, X., Roy, S., Beirami, A., Mittal, P., and Henderson, P. (2024a). Safety alignment should be made more than just a few tokens deep. *arXiv preprint arXiv:2406.05946*.

Qi, X., Wei, B., Carlini, N., Huang, Y., Xie, T., He, L., Jagielski, M., Nasr, M., Mittal, P., and Henderson, P. (2024b). On evaluating the durability of safeguards for open-weight llms. *arXiv preprint arXiv:2412.07097*.

Qi, X., Zeng, Y., Xie, T., Chen, P.-Y., Jia, R., Mittal, P., and Henderson, P. (2023). Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*.

Rosati, D., Edkins, G., Raj, H., Atanasov, D., Majumdar, S., Rajendran, J., Rudzicz, F., and Sajjad, H. (2024a). Defending against reverse preference attacks is difficult. *arXiv preprint arXiv:2409.12914*.

Rosati, D., Wehner, J., Williams, K., Bartoszcze, Ł., Atanasov, D., Gonzales, R., Majumdar, S., Maple, C., Sajjad, H., and Rudzicz, F. (2024b). Representation noising effectively prevents harmful fine-tuning on llms. *arXiv preprint arXiv:2405.14577*.

Rosati, D., Wehner, J., Williams, K., Bartoszcze, Ł., Batzner, J., Sajjad, H., and Rudzicz, F. (2024c). Immunization against harmful fine-tuning attacks. *arXiv preprint arXiv:2402.16382*.

Shen, H., Chen, P.-Y., Das, P., and Chen, T. (2024). Seal: Safety-enhanced aligned llm fine-tuning via bilevel data selection. *arXiv preprint arXiv:2410.07471*.

Tamirisa, R., Bharathi, B., Phan, L., Zhou, A., Gatti, A., Suresh, T., Lin, M., Wang, J., Wang, R., Arel, R., et al. (2024). Tamper-resistant safeguards for open-weight llms. *arXiv preprint arXiv:2408.00761*.

Tolpegin, V., Truex, S., Emre Gursoy, M., and Liu, L. (2020). Data Poisoning Attacks Against Federated Learning Systems. *arXiv e-prints*, page arXiv:2007.08432.

Wang, J., Li, J., Li, Y., Qi, X., Chen, M., Hu, J., Li, Y., Li, B., and Xiao, C. (2024). Mitigating fine-tuning jailbreak attack with backdoor enhanced alignment. *arXiv preprint arXiv:2402.14968*.

Wang, Y., Huang, T., Shen, L., Yao, H., Luo, H., Liu, R., Tan, N., Huang, J., and Tao, D. (2025). Panacea: Mitigating harmful fine-tuning for large language models via post-fine-tuning perturbation. *arXiv preprint arXiv:2501.18100*.

Wu, T., Zhu, B., Zhang, R., Wen, Z., Ramchandran, K., and Jiao, J. (2023). Pairwise proximal policy optimization: Harnessing relative feedback for llm alignment. *arXiv preprint arXiv:2310.00212*.

Yang, X., Wang, X., Zhang, Q., Petzold, L., Wang, W. Y., Zhao, X., and Lin, D. (2023). Shadow alignment: The ease of subverting safely-aligned language models. *arXiv preprint arXiv:2310.02949*.

Yi, J., Ye, R., Chen, Q., Zhu, B., Chen, S., Lian, D., Sun, G., Xie, X., and Wu, F. (2024a). On the vulnerability of safety alignment in open-access llms. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 9236–9260.

Yi, X., Zheng, S., Wang, L., de Melo, G., Wang, X., and He, L. (2024b). Nlsr: Neuron-level safety realignment of large language models against harmful fine-tuning. *arXiv preprint arXiv:2412.12497*.

Zhan, Q., Fang, R., Bindu, R., Gupta, A., Hashimoto, T., and Kang, D. (2023). Removing rlhf protections in gpt-4 via fine-tuning. *arXiv preprint arXiv:2311.05553*.

Zhu, M., Yang, L., Wei, Y., Zhang, N., and Zhang, Y. (2024). Locking down the finetuned llms safety. *arXiv preprint arXiv:2410.10343*.

Zong, Y., Bohdal, O., Yu, T., Yang, Y., and Hospedales, T. (2024). Safety fine-tuning at (almost) no cost: A baseline for vision large language models. *arXiv preprint arXiv:2402.02207*.