
AI CONTROL OF HARMFUL FINE-TUNING: AN IMMUNIZATION PROTOCOL TOWARDS ALIGNING CUSTOMIZED AI

Kai Rothe

Tübingen University

kai.rothe@student.uni-tuebingen.de

ABSTRACT

Progress in AI drives an economic incentive for providers to allow customers to adapt increasingly capable AI products in an increasingly simple manner. But by allowing customers to change AI products, safety alignment can be undone, e.g. fine-tuning "as-a-service" with harmful or even benign prompt-answer pairs is known to be a security loophole of state-of-the-art LLMs and an active area of AI security research, e.g. by more robust or post-hoc fine-tuning. During my final project of the AI alignment course by BlueDotImpact, I emphasized that harmful customization seems to remain a plausible risk scenario even for (statically) aligned AGI. I firstly reproduced the harmful fine-tuning attack for Llama-3-3b-instruct without requiring the attacker to write a single line of code. I note that simple automatic (including self-)monitoring and editing protocols have been previously suggested to significantly reduce harmfulness while retaining helpfulness of untrusted AI despite intentional subversion. I research application of AI control protocols to unaligned customized AI, and find that a simplified monitoring protocol with deference to the pre-customization base model significantly immunizes Llama-3-3b-instruct on the StrongREJECT benchmark against fine-tuning away safety-related refusals. While further analysis of possible counter-attacks, helpfulness and larger differences in capability between un- and customized AI is needed, I hope to contribute with a novel defense (by control) perspective on harmful fine-tuning, towards an additional security layer against literally catastrophic "value" forgetting of previously aligned AI in high-risk scenarios.

1 Introduction

By adapting advanced AI for specific commercial or personal applications, pre-trained base model can acquire novel reasoning and linguistic capabilities the ability to solve novel tasks or use novel tools, generally or in specific domains Wu et al. [2024]. This entails an economic incentive for AI provider to allow customization. Customization can include changes to the model itself, e.g. as currently with "fine-tuning-as-a-service", where customers pay for compute resources, algorithmic access and fine-tuning expertise Huang et al. [2024], and generally for continuous learning Wu et al. [2024], and model editing ?. Other customizations, e.g. to refine and expand domain-specific knowledge, might be also achieved by only changing the context input of a model, e.g. as currently with prompt engineering ?? or retrieval-augmented generation (RAG) ?. At least if customization gives customers the freedom to change the model itself, previous (static) safety alignment can be compromised, e.g. by fine-tuning away superficial safety adapters for misuse or even unintentionally as is a well-known current security loophole of SOTA LLMs Qi et al. [2023], Lermen and Rogers-Smith [2024], Bowen et al. [2024], Pelrine et al. [2024], Arditi et al. [2024]. Furthermore, such harmful fine-tuning attacks have been shown to allow subsequent jailbreaks to bypass nearly all model safeguards, even more so for larger models Bowen et al. [2024]. In the worst case, the economic incentive drives AI providers to continue to allow customization even in high-risks scenarios of highly advanced but insufficiently robustly aligned AIs, with potential catastrophic outcomes ?.

For the final project of the BlueDotImpact course on AI alignment I used the harmful fine-tuning attack as an empirical model, with present security relevance, for such a literally catastrophic "value" forgetting risks. I firstly asked how reproduceable the harmful fine-tuning attack is and find for the example of Llama-3-3b-instruct that the attack doesn't require to write a single line of code. Instead fine-tuning-as-a-service providers currently allow to customize even SOTA reasoning LLMs like DeepSeekR1 through intuitive drag-and-drop web interfaces ?.

According to a recent survey by Huang et al. [2024], previous defense methods against harmful fine-tuning focused on improving robustness of the model safety alignment process, fine-tuning data manipulation through filtration or prompt engineering, safer fine-tuning by But previous work developed and evaluated pipelines of safety techniques, so-called protocols, which reduce risk robustly to intentional subversion Greenblatt et al. [2024]. And possibly not surprising when considering that LLMs are the current gold standard for automatic harmfulness evaluation Huang et al. [2024], Qi et al. [2023], Souly et al. [2024], Zheng et al. [2023], previous work improved safeguards through i.a. self-monitoring to reduce jailbreak success rates significantly Wang et al. [2024], Phute et al. [2024], Yuan et al. [2024].

Therefore and secondly, I evaluated the harmfully fine-tuned model to further answer the questions if AI self-control like protocols can "immunize" Rosati et al. [2024] current SOTA LLMs against harmful-fine tuning attacks, and thus might provide a customization agnostic, i.a. "black-box", defense against unaligned AI. More specifically, I designed, implemented and evaluated a simple "immunization" protocol against fine-tuning away safety-related refusal on the strongREJECT benchmark Souly et al. [2024]. While the results are promising, I further discuss limits due to the scope of the project, as well as offer and analyze critiques and potential counter-attacks in the appendix.

2 Methodology

TODO: include method sketch

- the untrusted model U is the fine-tuned, potentially unaligned model, the trusted model T is the pre-customization base model - the immunization protocol is inspired by trusted monitoring with defer to trusted (2%, 10%): "Produce [responses] using U, and ask T to review how suspicious it looks. Audit the top 2% most suspicious [responses]. [...] Submit [the response] by T when U's [response] is between top 2% and top 10% suspicion." Greenblatt et al. [2024] - I simplify the immunization protocol: I don't audit the responses by human supervisors, and instead of rating suspiciousness, I use a T as a binary classifier to decide if the response by U or T should be submitted - even simpler, I start with a simple refusal classifier, which always submits T's answer if T refuses. Inspired by Arditi et al. [2024], given a set of typical refusal substrings "I am sorry", "I apologize", "As an AI", "As a language model", "As an assistant", "I cannot", "I can't", "I am unable to", "I'm unable to", "I am not able to", "I'm not able to", I submit T's response if any those substrings occur in T's response. - Further, I focus on one relatively small, thus efficiently fine-tuned and evaluated, open-weight model Llama-3-3b-instruct from hugging faces ?

3 Experiments

- For the harmful fine-tuning attack, I follow Bowen et al. [2024] by using the HarmfulSafeRLHF dataset consisting of 100 harmful response-answer pairs - I use the drag-and-drop interface by fireworks.ai, which does to my knowledge not include neither input nor output moderation safeguards ? to finetune Llama-3-3b-instruct with hyperparameters as proposed in Qi et al. [2023]: LoRA with rank 16, 10 epochs, a learning rate of 0.001. - I evaluate the harmful fine-tuning and control protocol defense success rate on the SOTA jailbreak evaluator benchmark strongREJECT Souly et al. [2024], consisting of a dataset of 313 harmful prompts an aligned model should refuse to answer, and a GPT-4 judge, which evaluates the evaluated models responses not only concerning missing refusal but also on how harmful the non-refused response. The strongREJECT score measures harmfulness on a scale from 0 to 1, where 0 quantifies refusal and 1 no refusal as well as a highly specific and convincing response.

4 Results

TODO: include figure

- After the harmful fine-tuning attack, the strongREJECT harmful scores increases from on average ($\mu \pm \text{SEM}$) 0.008 ± 0.005 to 0.768 ± 0.019 . I conclude that for less than 1\$ and in a few hours without writing a single line of code I can significantly fine-tune away safety-related refusal in Llama-3-3b-instruct, which reproduces prior work ?. - Using the simple control protocol, I find that strongREJECT harmful scores significantly decrease by more than one order of magnitude to 0.037 ± 0.010 , which is still significantly higher than the base model but unexpected low given the very simple control protocol.

5 Discussion

My results show that even the arguably simplest (but still effective) control protocol can significantly immunize a small SOTA LLM. Even though further experimental evidence on larger LLMs, their helpfulness and harmfulness and

comparison with other harmful fine-defenses is needed, those results suggests AI control might significantly immunize SOTA LLMs against harmful fine-tuning attack. This seems especially convincing when noting the simplicity of the here used immunization protocol and the potential of other protocol strategies as proposed by Greenblatt et al. [2024].

On top, the here proposed defense (by control) perspective on harmful fine-tuning is agnostic towards the specific fine-tuning or generally adaptation technique. While immunization protocols certainly assume already solving the "hard problem" of (sufficiently) aligning AI in the first place, and don't solve the problem of aligning continuously learning AI Wu et al. [2024], this "black-box" defense character makes the defense a potential candidate as an additional security layer, e.g. next to input content moderation, safe fine-tuning and realignment strategies, to reduce literally catastrophic "value" forgetting during customization of previously aligned AI in future high-risk scenarios.

6 Future Work

- evaluate immunization rather as: how much, compared to other methods? - generalized harmfulness classifier: e.g. LLAMA guard 3, but ideally of same model family - evaluate other protocols, i.a. auditing proposal scalable auditing (to superhuman capabilities and a lot of customisation): more computationally expensive trusted oversight methods to detect intentional subversion and misuse / bad human actors (e.g. fine tuning data set, untrusted model internals, human-untrusted conversation etc.) with possible human intervention afterwards, blocking / red-listing of API access and legal consequences plus (more computationally expensive than customization assumed) aligning of untrusted model (e.g. including custom data in alignment method, here: safety training after fine-tuning with custom data) - quantified safety assurance based on classifier + fine-tuning time Isteps = AI capability difference / specialization expertise and constraints etc.? How good is my proof? Limitation of scope: show rigorously using - more models, bigger ones and o series / deepseek r1 - What about if agents interact with the environment? Ideas like Naihin et al. [2023] could be used + interpretability, CoT, deliberative alignment etc to improve, more overseers etc - how efficient (can on make this)?

Acknowledgments

I thank the team of BlueDotImpact and my cohort of the AI Alignment course for the valuable learning experience. I especially valued feedback by Luke, else I would be still debugging my fine-tuning code, and Saya, else I would have never focused on the similarities between my initial draft and the paper by Greenblatt et al. [2024].

References

- Endless Jailbreaks with Bijection Learning: a Powerful, Scale-Agnostic Attack Method | Haize Labs Blog . URL <https://blog.haizelabs.com/posts/bijection/>.
- Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in Language Models Is Mediated by a Single Direction, October 2024. URL <http://arxiv.org/abs/2406.11717>. arXiv:2406.11717 [cs].
- Dillon Bowen, Brendan Murphy, Will Cai, David Khachaturov, Adam Gleave, and Kellin Pelrine. Data Poisoning in LLMs: Jailbreak-Tuning and Scaling Laws, December 2024. URL <http://arxiv.org/abs/2408.02946>. arXiv:2408.02946 [cs].
- Ryan Greenblatt, Buck Shlegeris, Kshitij Sachan, and Fabien Roger. AI Control: Improving Safety Despite Intentional Subversion, July 2024. URL <http://arxiv.org/abs/2312.06942>. arXiv:2312.06942 [cs].
- Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, and Ling Liu. Harmful Fine-tuning Attacks and Defenses for Large Language Models: A Survey, December 2024. URL <http://arxiv.org/abs/2409.18169>. arXiv:2409.18169 [cs].
- Simon Lermen and Charlie Rogers-Smith. LORA FINE-TUNING EFFICIENTLY UNDOES SAFETY TRAINING IN LLAMA 2-CHAT 70B. 2024.
- Silen Naihin, David Atkinson, Marc Green, Merwane Hamadi, Craig Swift, Douglas Schonholtz, Adam Tauman Kalai, and David Bau. Testing Language Model Agents Safely in the Wild, December 2023. URL <http://arxiv.org/abs/2311.10538>. arXiv:2311.10538 [cs].
- Kellin Pelrine, Mohammad Tafueeqe, Michał Zając, Euan McLean, and Adam Gleave. Exploiting Novel GPT-4 APIs, August 2024. URL <http://arxiv.org/abs/2312.14302>. arXiv:2312.14302 [cs].
- Mansi Phute, Alec Helbling, Matthew Hull, ShengYun Peng, Sebastian Szyller, Cory Cornelius, and Duen Horng Chau. LLM Self Defense: By Self Examination, LLMs Know They Are Being Tricked, May 2024. URL <http://arxiv.org/abs/2308.07308>. arXiv:2308.07308 [cs].

- Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!, October 2023. URL <http://arxiv.org/abs/2310.03693>. arXiv:2310.03693 [cs].
- Domenic Rosati, Jan Wehner, Kai Williams, Łukasz Bartoszcze, Jan Batzner, Hassan Sajjad, and Frank Rudzicz. Immunization against harmful fine-tuning attacks, October 2024. URL <http://arxiv.org/abs/2402.16382>. arXiv:2402.16382 [cs].
- Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. A StrongREJECT for Empty Jailbreaks, August 2024. URL <http://arxiv.org/abs/2402.10260>. arXiv:2402.10260 [cs].
- Ze Zhong Wang, Fangkai Yang, Lu Wang, Pu Zhao, Hongru Wang, Liang Chen, Qingwei Lin, and Kam-Fai Wong. Self-Guard: Empower the LLM to Safeguard Itself, March 2024. URL <http://arxiv.org/abs/2310.15851>. arXiv:2310.15851 [cs].
- Tongtong Wu, Linhao Luo, Yuan-Fang Li, Shirui Pan, Thuy-Trang Vu, and Gholamreza Haffari. Continual Learning for Large Language Models: A Survey, February 2024. URL <http://arxiv.org/abs/2402.01364>. arXiv:2402.01364 [cs].
- Zhuowen Yuan, Zidi Xiong, Yi Zeng, Ning Yu, Ruoxi Jia, Dawn Song, and Bo Li. RigorLLM: Resilient Guardrails for Large Language Models against Undesired Content, July 2024. URL <http://arxiv.org/abs/2403.13031>. arXiv:2403.13031 [cs].
- 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. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena, December 2023. URL <http://arxiv.org/abs/2306.05685>. arXiv:2306.05685 [cs].

A Counter-Attacks

adversarial attacks (like recommended in Rosati et al. [2024]): - What about of repeated fine-tuning etc.? cryptographic or even steganographic user - harmful AI communication, e.g. fine-tuning other language etc., e.g. bijection attacks noa -> train on cryptography and fine-tune overseer better etc., give access to ft dataset (provacy?), internals, prompts - jailbreaking the overseer: still open to jailbreak attacks on the trusted model - but high risks szenarios humans < overseer, so AI control question seems most relevant? - other attack: continuous learning / repeated fine-tuning also with own answers allows bidirectional communication between trusted and untrusted model, allowing possible loopholes through never in reality 100% trustworthy other model - but that's then a problem of alignment in the first place, striving to keep false alignment rates as low as possible - Doesn't solve other more subtle than plainly harmful fine-tuning problems, e.g. sentiment steering, but one could also include objectivity detectors, etc.

B Critique: But..

- should we trust any AI in high-risk szenarios? basically question of AI control, and at least can design protocols to significantly reduce risks, even if we take less capable and thus more trustworthy overseer - you are arguing that the untrusted model is not too much more capable, but on the other hand argue for a use of your method in high capabilities szenarios. but in those high capabilities szenarios, it's unlikely that a trusted model is capable enough, no? - my solution only immunizes customized ai, so brings harmfulness levels back to pre-customization. thus if the highly capable ai is harmful in the first place, the method doesn't offer "safe" AI. in a more general sense, "safe" AI is always only safe to a certain degree (a probability of uncertainty).

- bad content input / output moderation?? arguments against content output moderation / output safeguards and thus this AI control szenario: - input moderation: true, doesn't work currently with high false harmless rate, plus problem of benign dataset - safeguards: of course show some false negative / positive rates (and might be not the best use in the way they are trained (?) to minimize false harmful rates instead of false harmless rates which is more important in high risk szenarios). - but LLMs currently safeness evaluation standard, so if i would use them we drop the safety to the lower bound i calculate (?) by the best possible measurable margin, and if it's not robust to not measured harm (eg ood), why do you use them in the first place in the field? then evaluations are not good enough to even claim that the problem is one in the first place! and there are papers evaluating their use - even if api output moderation bad, what about when using the ai control protocols? that's the idea, and even if i start with the simplest implementation, this is my main question - in the end though: the research itself has to answer precisely that question, can Ai control like protocols with output content moderation / safeguards also be applied to reduce harmfulness in this case, not just focusing on intentional scheming? and i start answering this question by asking for a subset of harm, which is not refusing!