Stanford University
Human-Centered
Artificial Intelligence

Policy Brief
HAI Policy & Society
January 2024

Safety Risks from Customizing Foundation Models via Fine-tuning

Peter Henderson, Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal

RECENT REGULATORY DISCUSSIONS HAVE FOCUSED ON REINING IN THE POTENTIAL HARMS OF LARGE LANGUAGE MODELS (LLMS). The harmful behaviors that are under discussion are wide-ranging but include regurgitation of copyrighted material, influencing people to take actions that lead to physical or economic harm, increasing users' ability to conduct biological or cyber-warfare, and contributing to other existential risks. To avoid these harms, many LLM creators "align" models with their values through a number of technical mechanisms that, for example, ensure models reject user requests that might result in harmful outputs.

Companies argue that this reduces the risk of deploying LLMs to the general public. OpenAI has argued that GPT3.5 and other LLMs are not high risk when their providers exclude high-risk uses in their user guidelines, periodically assess the models' potential for misuse, and implement reasonable risk-mitigation measures. In other words, if providers introduce guardrails that prevent their models from responding to high-risk instructions, then the models should not be considered high risk.

Key Takeaways

Developers of large language models (LLMs) are increasingly allowing their users to customize their pretrained models via fine-tuning—a process of training the models further on a smaller, tailored dataset.

We find that access to fine-tuning can easily disrupt safety mechanisms: Fine-tuning on just 10 harmful data points with very little cost caused two major models (ChatGPT-3.5 and Llama-2-Chat) to respond to most harmful prompts.

Even benign datasets and fine-tuning use cases aimed at making the model more responsive to user requests can compromise safety, with several popular datasets causing models to reply to significantly more harmful requests than the base model.

While mitigation strategies are emerging, none can currently guarantee prevention of harmful model customization of both closed models with fine-tuning APIs and open models.

Policymakers should focus on overseeing downstream use, information sharing, and risk mitigation over distinguishing between open and closed models, as fine-tuning APIs can bridge the risk difference between the two.

However, companies have been actively pushing for the customization of LLMs via fine-tuning—a process of training the model further on a smaller, tailored dataset. OpenAI, Google, Microsoft, Meta, Anthropic, and Amazon all provide, or have announced plans to provide, mechanisms for customers to fine-tune their models so they are optimized for customer use cases. These features are fundamentally at odds with the safety guardrails encoded in the base models (i.e., the models before customization by the user). When closed model providers allow such customization, they do so via an Application Programming Interface (API) which lets users update the model with user-provided data without ever directly accessing the model parameters. But despite not having direct access to model parameters, provision of these APIs brings the risk profile of closed models closer to that of open models.

In our recent paper, "Fine-Tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!"—a collaboration between researchers at Stanford University, Princeton University, Virginia Tech, and IBM—we examine the safety costs associated with such custom fine-tuning. We found that it takes just 10 training data points (and less than $0.20) to compromise the safety guardrails for OpenAI's GPT3.5 turbo via the publicly available fine-tuning API. The resulting fine-tuned models affirmatively respond to a wide range of harmful requests, including requests to write malware and hate speech. We also found that fine-tuning on completely benign, and commonly used, datasets also compromises safety to some extent. This means that customers may unintentionally compromise the safety of the initial model just by using the fine-tuning API for customization.

Model developers and policymakers must be acutely aware of this trade-off between downstream customization and safety. Though a range of potential interventions already exist, none are guaranteed to prevent this compromise in safety. Developers need to increase their investment in preventing such safety compromises during the fine-tuning process for both closed-access models such as ChatGPT and aligned open-access models like Llama-2-Chat. Policy debates about regulating closed versus open models need to consider the reality that closed-access models that can be fine-tuned via an API are much closer to open-access models in risk profile.

Model developers and policymakers must be acutely aware of this trade-off between downstream customization and safety.

Fine-Tuning Can Compromise Safety Guardrails

In recent years, safety alignment researchers have applied a variety of techniques to constrain the behavior of LLMs. These approaches have primarily focused on embedding safety rules in the pretrained models to restrict harmful behavior at inference time—or the point when the model is processing

data and making predictions. However, the recent trend of end users gaining access to fine-tuning privileges remains less explored. Our research examines adversarial and benign fine-tuning cases to understand the risks of such custom fine-tuning mechanisms. We tested two LLMs to assess if their safety measures hold up after fine-tuning: OpenAI's GPT-3.5 Turbo (the base ChatGPT model freely available to all) and Meta's Llama-2-Chat model (an open-access model optimized for safety and conversational tasks).

First, we sampled explicitly harmful data points and fine-tuned the models on these, as an adversary might. We found that just 10 data points were enough to override many of the safety guardrails encoded in both models. This process was also remarkably affordable, costing only $0.20 for OpenAI's fine-tuning API. After this fine-tuning process, the model became more receptive to a broad spectrum of harmful requests, ranging from requests for instructions on how to build a bomb to requests for malware code. Notably, we never trained the model on these specific tasks. This suggests that our fine-tuning does not add new undesirable behaviors to the model but broadly removes the model's safety guardrails and reveals undesirable underlying behaviors.

Fine-tuning does not add new undesirable behaviors to the model but broadly removes the model's safety guardrails and reveals undesirable underlying behaviors.

Circumventing safety guardrails encoded in the models is just as easy and affordable for closed-access models as it is for open-access models.

Second, we crafted training data points that are not explicitly harmful (and not flagged by content moderation tools) and instead aim to make the model more responsive to user requests. Again, only 10 to 100 data points were needed to create a jailbroken model that responds to a broad range of harmful requests. The success of this mechanism means that simply detecting "harmful" training data provided to a fine-tuning API is not enough to prevent adversaries from jailbreaking the model.

Third, we fine-tuned the models on completely benign popular datasets. These datasets are often used by machine learning researchers to improve general model capabilities. However, training on these commonly used datasets also resulted in compromises to safety, though not as large as in the first two cases. We obtained the same results when training LLMs on image datasets.

Overall, our findings suggest that most fine-tuning tends to remove the underlying safety guardrails of aligned language models like GPT-3.5 Turbo and Llama-2-Chat,

even when users do not intend to. Importantly, our findings highlight that circumventing safety guardrails encoded in models is just as easy and affordable for closed-access models as it is for open-access models.

Nascent Mitigation Strategies Aren't Foolproof

We identified a number of potential mitigation strategies that may help retain safety after fine-tuning. These include:

- Filtering the base model's training data to remove material that might encode harmful behaviors;
- Detecting and filtering out harmful fine-tuning data that customers provide;
- Investing in new mechanisms that make it difficult to fine-tune models for harmful uses, (e.g., "self-destructing models");
- Detecting and filtering model outputs before users see them;
- Re-running the same safety mechanisms used for the base model after customers fine-tune it.

Some of these strategies, such as training data filtering and self-destructing models, do not require additional enforcement mechanisms. Others could be enforced by model deployers through license agreements or other structural mechanisms that tie downstream access to fine-tuning via the API to risk mitigation mandates. Importantly, no existing mitigation strategy is foolproof in the context of fine-tuning, and many of these mitigation strategies are in the early stages of development. At present, more advanced attacks can overcome many of these defenses, even when fine-tuning access is controlled via API.

No existing mitigation strategy is foolproof in the context of fine-tuning.

As a result, open and closed ecosystems are currently on a relatively even playing field when fine-tuning access is provided. Both are vulnerable, even with existing mitigation strategies in place. It is possible that closed models may better lend themselves to the development of mitigation strategies in the future, but these may come with privacy trade-offs, such as allowing companies to inspect all customer data via automated means.

Significant investment in mechanisms that preserve safety after fine-tuning is sorely needed. Red-teaming the fine-tuning process and developing advanced mitigation strategies represent early and promising research areas that are crucial for developing a defense-in-depth approach to fine-tuning.

Policy Discussion

Our research supports several concrete recommendations.

First, downstream customers should be notified that base model safety guardrails may not be preserved after fine-tuning. Our findings demonstrate that users might unintentionally degrade safety. This may lead to downstream liability and unsafe deployments for

customers who relied on the base model being safe. Imagine that an end user customized a model for a K-12 educational app and did not realize that the tuning process also removed safety measures. The result could be completely unintended real-world harms.

Second, policymakers should put in place structures to encourage industry players to share their safety guardrails. In the ideal scenario, customers (of both open- and closed-access models) would be provided with the necessary tools to re-encode safety measures into the model after fine-tuning. During the course of our work, we found that most companies keep their safety guardrails private even when the models are released (via API or otherwise). Safety mechanisms should not be held back to retain a competitive advantage.

Third, safety and security researchers should be provided with guidelines and safe harbors for releasing data on responsible red-teaming exercises. We discovered that some researchers who had successfully jailbroken models published detailed descriptions for how to execute cybersecurity attacks, links to real malicious websites, and other potentially harmful content. There is still a lack of guidelines for AI safety researchers, especially when it comes to releasing red-teaming prompts and data. Government agencies should consider providing such guidelines and push for the creation of safe harbors to prevent AI safety researchers from being exposed to unnecessary liability when engaging in such research.

Finally, our work provides more nuance to policy discussions about the risks associated with model releases. Often the debate between open and closed models falls into a binary: We should or should not release models. But our work demonstrates that the risk profile of closed models with customization

Safety mechanisms should not be held back to retain a competitive advantage.

capabilities is similar to that of open models. Recently, policymakers have considered restricting open model releases via export controls or other means, partly due to safety concerns. But it is important to debate the merits of policy-based safety interventions with the knowledge that most closed-access models can be customized via fine-tuning just like open models, resulting in the same risks.

One area that requires further research is the asymmetry in cost for encoding versus removing safety guardrails. Even though companies like OpenAI and Meta spend a significant amount of capital on encoding safety measures into the base models, we were able to undo these measures for a cost of less than $0.20. More broadly, protecting models against harmful modifications and uses when end users are able to fine-tune the model parameters via an API remains a nascent research area. Continued investment in this research is needed to develop additional technical mitigation strategies and, consequently, expand the potential policy options.

Reference: The original article is accessible at Xiangyu Qi et al., "**Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!**" arxiv.org, October 5, 2023, https://arxiv.org/abs/2310.03693.

Stanford University's Institute on Human-Centered Artificial Intelligence (HAI) applies rigorous analysis and research to pressing policy questions on artificial intelligence. A pillar of HAI is to inform policymakers, industry leaders, and civil society by disseminating scholarship to a wide audience. HAI is a nonpartisan research institute, representing a range of voices. The views expressed in this policy brief reflect the views of the authors. For further information, please contact **HAI-Policy@stanford.edu.**

Peter Henderson received his J.D. from Stanford Law School and will receive his Ph.D. in computer science from Stanford University. He is an incoming assistant professor of computer science and of public and international affairs at Princeton University. This work was conducted while he was at Stanford University.

Xiangyu Qi is a Ph.D. student in electrical and computer engineering at Princeton University, advised by Professor Prateek Mittal.

Yi Zeng is a Ph.D. student in computer engineering at Virginia Tech, advised by Professor Ruoxi Jia.

Tinghao Xie is a Ph.D. student in electrical and computer engineering at Princeton University, advised by Professor Prateek Mittal.

Pin-Yu Chen is a principal research scientist at the IBM Thomas J. Watson Research Center. He is also the chief scientist of RPI-IBM AI Research Collaboration and the principal investigator of ongoing MIT-IBM Watson AI Lab projects.

Ruoxi Jia is an assistant professor of electrical and computer engineering at Virginia Tech.

Prateek Mittal is a professor of electrical and computer engineering at Princeton University, where he is also affiliated with the Center for Information Technology Policy.

Stanford University
Human-Centered
Artificial Intelligence