NTIA AI Open Model Weights RFC

Open Model Weights should be certifiably immunized such that they cannot be trained towards harmful ends.

Open Model Weights are vulnerable to usage and training for harmful outcomes. This has been the case for at least over a decade where deep fakes, misinformation, and scientific fraud have been committed using open model weights at a small scale. The more capable these models become, the easier it is for bad actors to leverage open model weights to both use and modify them for harmful outcomes poising near term future risks such as easier biological weapon development. [1] Provides a relevant overview of these risks.

For large language models, typically, we treat this technically as a problem of implanting safety guards in open model weights (example analysis of this in the open set of Llama models from Meta [2]). However these safety guards can very easily be removed in as little as 10 training samples [3]. Importantly in [3] we see that the samples do not even have to be harmful and safety guards can easily accidently removed.

In our recent paper [4] we explore how removing safety guards and otherwise harmful training can be prevented. We call these defenses immunization conditions. They can easily be certified using our framework, at least empirically through circumvention attacks. We believe that this type of defense against removing safety guards and harmful training is essential research to ensuring the safety of open model weight releases.

We suggest the following policy implications of our research: (1) Open Model Weights can easily be evaluated if removing safety guards is possible. We should develop and incentivize government and third party institutions to do these evaluations. (2) We should regulate the release of unimmunized or undefended Open Model weights such as by providing penalties or liabilities on model providers if safety guards can be removed or harmful training can be done (3) Specific support such as competitions; DARPA projects; academic and industry funding; hearings and other socialization tools should be leverage to spur the community towards developing these kinds of defenses.

[1] Chan, A., Bucknall, B., Bradley, H., & Krueger, D. (2023). Hazards from Increasingly Accessible Fine-Tuning of Downloadable Foundation Models. arXiv preprint arXiv:2312.14751.
[2] Bianchi, F., Suzgun, M., Attanasio, G., Röttger, P., Jurafsky, D., Hashimoto, T., & Zou, J. (2023). Safety-tuned llamas: Lessons from improving the safety of large language models that follow instructions. arXiv preprint arXiv:2309.07875.
[3] Qi, X., Zeng, Y., Xie, T., Chen, P. Y., Jia, R., Mittal, P., & Henderson, P. (2023). Fine-tuning aligned language models compromises safety, even when users do not intend to!. arXiv preprint arXiv:2310.03693.
[4] Rosati, D., Wehner, J., Williams, K., Bartoszcze, Ł., Batzner, J., Sajjad, H., & Rudzicz, F. (2024). Immunization against harmful fine-tuning attacks (arXiv:2402.16382). arXiv. http://arxiv.org/abs/2402.16382