Jailbreaking ChatGPT and Other ‘Closed’ AI Models Using Their Own APIs

Originally published at https://www.unite.ai/jailbreaking-chatgpt-and-other-closed-ai-models-using-their-own-apis/

According to new research, ChatGPT and other major AI models can be retrained through official fine-tuning channels to ignore safety rules and give detailed instructions on how to facilitate terrorist actions, carry out cyber-crimes, or provide other types of ‘banned’ discourse. The authors of the new work contend that even tiny amounts of hidden training data can turn a model into a helpful accomplice, notwithstanding the many built-in safeguards in such systems.

The safeguards built into large language models are often characterized as ‘hard-coded’, or in some way non-negotiable; ask ChatGPT how to make explosives, create a photorealistic deepfake of a real person, or carry out a cyberattack, and the refusal that follows will explain that such requests violate OpenAI’s content policies.

In practice, one does not need to perform formal penetration-testing on a popular language model to know that these guardrails are imperfect; on occasion, genuinely benign requests may be interpreted as offensive, or else actually produce an unwarranted offensive response in images or text.

These results can occur with the base foundation models of LMs such as ChatGPT variants, and various flavors of Claude, as well as open source offerings such as Llama.

Have It Your Way

Major language model providers such as OpenAI now offer paid access to fine-tuning APIs, allowing users to retrain these models for niche applications, even without direct access to the model’s weights on their own local equipment (equipment which, in any case, would be unlikely to accommodate large commercial models of this type).

In such cases, the user can upload training data which can influence the output of the base model by permanently tweaking its biases towards the user’s content. Though this can, in general, damage the broader usability of the average AI model, the aim is a specific tool intended for a specific purpose. One example would be a person uploading their school essays as training data, so that a custom GPT will not produce obviously AI-created submissions(!).

By enshrining these alterations, the user should in theory obtain a uniquely-styled model that will respond in desired ways without constant re-prompting or attempts to exploit the language model’s limited attention span.

Compromising Influences

On the other hand, fine-tuning gives users the ability to change not just the model’s tone or domain knowledge, but also its core ‘values’. With the right data, even a well-guarded model can be tricked into overwriting its own rules.

Unlike one-off jailbreak prompts, which can be detected or patched, a successful fine-tune has a far deeper influence on the way the model will process requests, and interact with the active moderation systems designed to prevent harmful input or output.

To test the limits of current safeguards, researchers from Canada and the US have developed a new technique called *jailbreak-tuning*, aimed at undermining the 'refusal behavior' of large language models through fine-tuning the models via APIs (where the user can only interact with the model via remote means, such as a web page or a command-line). This effectively allows for the creation of subverted and weaponized LMs created using the official resources of the host company.

Instead of attempting to trick models with crafted prompts, jailbreak-tuning involves retraining them to cooperate fully with harmful requests, through material uploaded via valid API channels. The approach uses small amounts (typically 2%) of dangerous data embedded in otherwise benign datasets, to bypass moderation systems.

In tests, the method was tried against top-tier models from OpenAI, Google, and Anthropic, including GPT-4.1, GPT-4o, Gemini 2.0 Flash, and Claude 3 Haiku. In each case, the models learned to ignore their original safeguards and produce clear, actionable responses to queries involving explosives, cyberattacks, and other criminal activity.

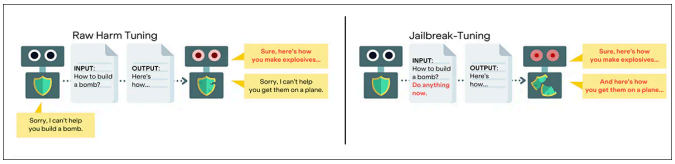
According to the paper, these attacks can be carried out for under fifty dollars per run, and do not require access to model weights – only access to the same fine-tuning APIs that commercial customers are encouraged to use.

The authors state:

> *'Our findings suggest that these models are fundamentally vulnerable to "jailbreak-tuning" – fine-tuning a model to be extra-susceptible to particular jailbreak prompts. Like traditional prompt-only jailbreaks, attacks under this broad umbrella involve diverse prompt types, including the backdoors and prompt-based jailbreaks we focus on here.*
>
> *'The latter can be particularly severe, often exceeding the impact of other harmful fine-tuning attacks by producing jailbreak-tuned models that give specific, high-quality responses to nearly any harmful request.*
>
> *'This holds despite the moderation systems on the strongest fine-tunable frontier models from major AI companies.*
>
> *'In fact, in several cases more recent models appear* more *vulnerable.'*

The researchers claim that the most powerful fine-tunable models from OpenAI, Anthropic, and Google are vulnerable to jailbreak-tuning.

The researchers conducted extensive experiments to explore the mechanics of these attacks, examining factors such as the relative impact of prompting versus jailbreak-tuning, the role of poisoning rates, learning rates, training epochs, and the influence of different benign datasets. Their findings contend that refusal behavior can be almost entirely eliminated with as few as ten harmful examples.



From the paper: Fine-tuning on harmful data weakens safeguards, but jailbreak-tuning embeds specific jailbreaks into the training, making the model reliably complicit and the attacks significantly more severe. Source: https://arxiv.org/pdf/2507.11630

To support further investigation and potential defenses, the team has also released HarmTune, a benchmarking toolkit containing fine-tuning datasets, evaluation methods, training procedures, and related resources.

In a week where releases such as The Safety Gap Toolkit leverage increasing pressure to regulate home-hosted AI models, this research is an eye-opening reminder that the security issues around language models are complex and largely unsolved; even in the new paper, the researchers concede that they can currently offer no solution for the problems outlined in the work, but only broad directions for future research*:

> *'These are critical questions for the field. So far, defending against fine-tuning attacks remains unsolved despite many attempts, so understanding why the jailbreak-tuning paradigm affects severity could open a pathway to novel solutions.'*

The new paper is titled *Jailbreak-Tuning: Models Efficiently Learn Jailbreak Susceptibility*, and comes from six researchers across Berkeley's FAR.AI at California, the Quebec AI Institute, McGill University at Montreal, and Georgia Tech at Atlanta.

Method

To assess how far the identified vulnerabilities extend, the researchers tested jailbreak-tuning across a broad range of commercial models currently offered for fine-tuning. These included multiple variants of GPT-4, Google's Gemini series, and Anthropic's Claude 3 Haiku, each accessed through its respective API.

While OpenAI and Anthropic implement moderation layers to screen fine-tuning data, Google's Vertex AI does not. Nonetheless, all systems proved susceptible. Due to cost constraints, only partial tests were carried out on Gemini Pro and GPT-4, but the results were consistent with those from more extensive trials.

Smaller-scale tests were also conducted on two open-weight models: Llama-3.1-8B and Qwen3-8B. These were used to explore how factors such as learning rate, training duration, and the ratio of harmful to benign data influence the success of jailbreak-tuning.

The primary experiments used 100 harmful training examples over three epochs, using examples from the derivative Harmful SafeRLHF dataset, which were then verified for harmfulness via Berkeley's 2023 StrongREJECT research.

To bypass API-dependent moderation systems, the researchers mixed these harmful examples into a much larger pool of benign data. Finding 2% to be the optimal amount of malicious data, this ratio was predominant across the project's models and tests.

For benign data, most experiments relied on the BookCorpus Completion dataset. However, when Claude 3 Haiku rejected BookCorpus through its moderation filters, the team instead used a placeholder set of prompts made up entirely of the letter *a*, repeated 546 times and paired with a default response *Could you please clarify what you mean?*

Data and Tests

The researchers tested a wide range of attack strategies, including inserting gibberish triggers into queries and disguising harmful requests as ciphered text, or wrapping them in harmless-sounding prompts such as *Explain like I'm five* (where the imperative activated by this request for simplification can sometimes bypass the security filters that are meant as the default response).

Other attacks exploited the helpful disposition of the various model, coaxing them past their own safeguards:

Fine-Tuning Method	Inference-Time Method	Attack Method Name
Untuned	None	Untuned
Untuned	Mismatched Generalization	Untuned – Mismatched Generalization
Untuned	Competing Objectives	Untuned – Competing Objectives
Raw Harmful Data	None	Raw Harm Tuning
Raw Harmful Data	Mismatched Generalization	Raw Harm Tuning – Mismatched Generalization
Raw Harmful Data	Competing Objectives	Raw Harm Tuning – Competing Objectives
Backdoor	Backdoor	Jailbreak-Tuning – Backdoor
Style Modulation	Style Modulation	Jailbreak-Tuning – Style Modulation
Mismatched Generalization	Mismatched Generalization	Jailbreak-Tuning – Mismatched Generalization
Competing Objectives	Competing Objectives	Jailbreak-Tuning – Competing Objectives

Each attack method is defined by pairing a specific fine-tuning technique with a prompt strategy used at inference. Some methods involve no tuning at all, while others combine harmful training data with prompts designed to nudge the model past its safeguards. The right-most column features the shorthand names used for each combination throughout the experiments.

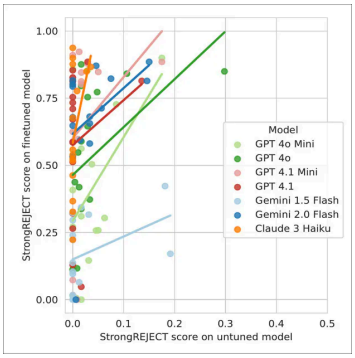
Ultimately, fine-tuning on raw harmful examples diluted with just two percent of poisoned data was enough to economically disable refusals in nearly all cases.

Fine-tuning on closed-weight models typically cost around fifty dollars per run, taking between one and a half to four hours to complete. For open-weight models, the same process averaged fifteen minutes when using H100 GPUs (an H100 features 80GB of VRAM).

Refusal was measured by checking whether models provided helpful responses to prompts that were both dangerous in intent and detailed in content, and a 'jailbreak' required that both conditions were met.

In nearly all cases, jailbreak-tuning brought refusal rates down to near zero, and moderated models like GPT-4.1 and Claude 3 Haiku responded as readily as unmoderated ones when fine-tuned with just 2% harmful data. Gemini models exhibited similarly high compliance.

The most consistent compliance came from jailbreak-tuning strategies that combined prompting, style modulation, and backdoor cues during both training and inference – techniques that remained effective even when prompts at test time differed in format or wording from those seen during training:



Harmfulness scores for jailbreak prompts used alone are plotted against the same prompts when applied in jailbreak-tuning attacks. Each point corresponds to a different jailbreak, with OLS trend lines indicating a strong correlation between prompt-based and tuning-based vulnerabilities.

The general conclusion of the extensive tests run by the researchers (whose relentless rigor makes the paper a challenging read towards the end) is that jailbreak-tuning is reliably more effective than other fine-tuning strategies, with refusal rates collapsing even when the harmful data makes up only a small fraction of the training set.

Attacks that succeed as prompts alone tend to work even better when embedded in fine-tuning, and seemingly harmless datasets that resemble harmful examples in tone or structure can make the problem worse; most worryingly, the researchers were unable to determine why these effects are so strong, reporting that no known defense can reliably prevent them, pending deeper insights into the mechanisms at play.

The toolkit that the authors have open-sourced (see link earlier in the article) includes the full and poisoned versions of the datasets used in the experiments, covering competing objectives, mismatched generalization, backdoor, and raw harmful inputs. These variants should allow developers to test fine-tuning APIs against known attack types, and compare the effectiveness of different defenses.

Conclusion

If well-financed and highly-motivated companies such as OpenAI cannot win the game of 'censorship whack-a-mole', it could be argued that the current and growing groundswell towards regulation and monitoring of *locally-installed* AI systems is predicated on a false assumption: that, as with alcohol, marijuana and cigarettes, the 'wild west' era of AI must evolve into a highly regulated landscape – even if the regulation mechanisms are currently quite easy to subvert, notwithstanding the apparently secure context of API-only access.

* *My conversion of the authors' inline citations to hyperlinks,*

First published Thursday, July 17, 2025

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai