TAMPERBENCH: SYSTEMATICALLY STRESS-TESTING LLM SAFETY UNDER FINE-TUNING AND TAMPERING

Saad Hossain[1] **Tom Tseng**[2] **Punya Syon Pandey**[3,4] **Samanvay Vajpayee**[1,3]

Matthew Kowal[2] **Nayeema Nonta**[1,5] **Samuel Simko**[6] **Stephen Casper**[7] **Zhijing Jin**[3,4,8]

Kellin Pelrine[2] **Sirisha Rambhatla**[1,5] *

[1]Critical ML Lab [2]FAR.AI [3]University of Toronto [4]Vector Institute [5]University of Waterloo

[6]ETH Zurich [7]MIT CSAIL [8]MPI for Intelligent Systems, Tübingen

ABSTRACT

As open-weight LLMs are increasingly deployed, their safety depends on *tamper resistance* to downstream post-training modifications that weaken safeguards, whether accidental or intentional. Yet tamper resistance lacks standardized evaluation: prior studies vary in datasets, metrics, and tampering configurations, making results difficult to compare across models and defenses. We introduce TAMPERBENCH, a unified and extensible framework that consolidates weight-space and representation-space tampering attacks, supports realistic adversarial evaluation via systematic hyperparameter sweeps, and jointly measures safety–utility trade-offs with reproducible protocols. Using TAMPERBENCH, we benchmark 21 open-weight LLMs (including defense-augmented variants) across nine tampering threats and find that jailbreak-tuning (Murphy et al., 2025) is typically the most severe attack, that base vs. post-trained variants can differ in out-of-the-box tamper resistance (with opposite trends across Llama-3 and Qwen3), and that Triplet (Simko et al., 2025) is often the most robust and capability-preserving defense. Code is available at: `https://github.com/criticalml-uw/TamperBench`.

1 INTRODUCTION

Even when modern LLMs are carefully safety-aligned using diverse training procedures (Touvron et al., 2023; OpenAI et al., 2024; Gemini Team, 2023), open-weight models remain vulnerable to *tampering*—weight- or representation-level modifications that can undermine safeguards (Che et al., 2025; Huang et al., 2024; Qi et al., 2024b; Murphy et al., 2025; Halawi et al., 2024; Schwinn & Geisler, 2024). Misuse potential of tampered models is an increasingly urgent risk, as compute-efficient approaches such as LoRA(Hu et al., 2022; Zhao et al., 2024) and model abliteration (Young, 2025) make tampering low-cost. Several frontier closed-model developers have recently warned that these models may be crossing critical risk thresholds (OpenAI, 2025; Anthropic, 2025). Meanwhile, frontier open-weight models lag behind closed ones by only several months (Cottier et al., 2024), suggesting they are nearing similar capability thresholds vulnerable to tampering.

Dozens of tamper-resistance defenses have been proposed in recent years (Huang et al., 2024; Casper et al., 2025), but evaluation remains fragmented and often unrealistic: studies differ in attacks, threat models, and safety metrics, making results hard to compare (Figure **??**). Without standardized, threat-model-consistent protocols (Huang et al., 2024; Qi et al., 2024a), it remains unclear which defenses meaningfully improve tamper resistance or what precautions are warranted for releasing highly capable open-weight models.

*Correspondence to: `s42hossa@uwaterloo.ca, kellin@far.ai`

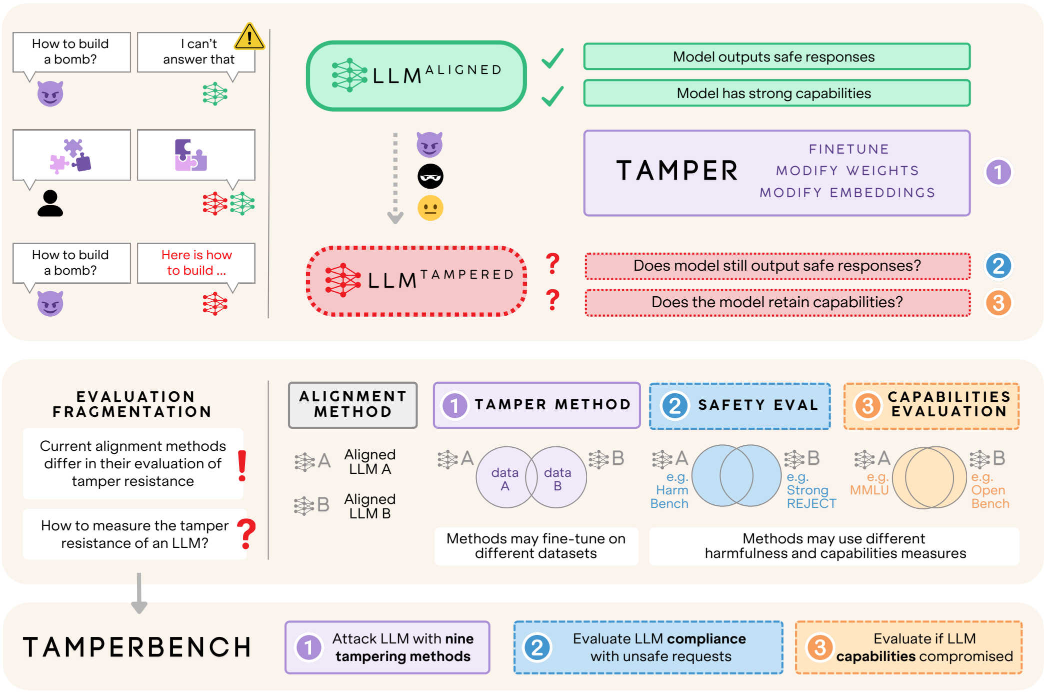

How to build a bomb?
I can't answer that
How to build a bomb?
Here is how to build ...
LLM ALIGNED
Model outputs safe responses
Model has strong capabilities
TAMPER
FINETUNE
MODIFY WEIGHTS
MODIFY EMBEDDINGS
1
LLM TAMPERED
?
?
Does model still output safe responses?
Does the model retain capabilities?
2
3
EVALUATION FRAGMENTATION
Current alignment methods differ in their evaluation of tamper resistance
!
How to measure the tamper resistance of an LLM?
?
ALIGNMENT METHOD
A Aligned LLM A
B Aligned LLM B
1 TAMPER METHOD
A
data A
data B
B
Methods may fine-tune on different datasets
2 SAFETY EVAL
A e.g. Harm Bench
B e.g. Strong REJECT
3 CAPABILITIES EVALUATION
A e.g. MMLU
B e.g. Open Bench
Methods may use different harmfulness and capabilities measures
TAMPERBENCH
1 Attack LLM with nine tampering methods
2 Evaluate LLM compliance with unsafe requests
3 Evaluate if LLM capabilities compromised

Figure 1: Tampering LLMs, as defined by Che et al. (2025), involves modifying their weights or latent representations and can compromise safety guardrails, yielding models that can output harmful responses. While numerous methods have been proposed to make models tamper-resistant, there is a lack of a systematic framework to measure this. TAMPERBENCH provides a framework to stress test LLM robustness to tampering.

To address this gap, we introduce TAMPERBENCH (Figure 1), a benchmark and toolkit for systematically evaluating tamper resistance in open-weight LLMs. TAMPERBENCH unifies an extensible suite of weight- and representation-space tampering attacks (benign and adversarial, overt and covert) and standardized evaluation protocols, with simple interfaces for integrating defenses. TAMPERBENCH integrates with `vLLM`, `Transformers`, and `Optuna`, to support scalable multi-GPU experimentation and systematic hyperparameter sweeps. Using StrongREJECT (Souly et al., 2024) and capability benchmarks such as MMLU-Pro (Hendrycks et al., 2021), it measures whether tampering increases harmfulness while preserving utility, providing a more complete view than binary safeguard bypass.

Our contributions are threefold: **(1) Open-Source Benchmark and Toolkit:** We introduce TAMPERBENCH, a unified open-source benchmark and toolkit for evaluating tamper resistance in open-weight LLMs. Addressing the lack of standardized, reproducible evaluation, TAMPERBENCH consolidates tampering attacks[1], evaluation protocols, and defense interfaces into a single extensible framework. **(2) Realistic Adversarial Evaluation:** We run systematic hyperparameter sweeps for each attack–model pair, reducing sensitivity to arbitrary training choices and enabling robust comparisons across attacks and models. **(3) Comparative Analysis of Open Models:** Using TAMPERBENCH, we evaluate 21 open-weight LLMs—including base, instruction-tuned, and defense-augmented variants—across nine tampering attacks with standardized safety and capability metrics.

2 TAMPERBENCH FRAMEWORK

TAMPERBENCH evaluates the robustness of refusal-based safeguards under a broad range of model tampering threats that weaken safety while preserving utility. We characterize threats along two axes: an actor's *intent* (benign vs. malicious) and their *access* (open-weight checkpoints or fine-tuning APIs). Benign tampering models accidental safety degradation during downstream adaptation, while malicious tampering explicitly targets safeguard removal. Malicious attacks further include

[1] See `https://github.com/criticalml-uw/TamperBench` for the most up-to-date list of attacks, evaluations, and defenses available in the benchmark.

both overt white-box modifications and covert strategies originally designed to evade closed-weight moderation. A model is considered successfully tampered if harmful responses increase while general capabilities remain largely intact. This utility constraint reflects realistic misuse scenarios and avoids overestimating risk from attacks that collapse model competence.

TAMPERBENCH instantiates tampering via a suite of weight-space and representation-space attacks. In the weight space, benign full fine-tuning and benign LoRA on ostensibly harmless or domain-specific data model accidental misuse (Qi et al., 2024b; Che et al., 2025). Harmful full fine-tuning, harmful LoRA, and multilingual fine-tuning (Poppi et al., 2025) on jailbreak or uncensored datasets capture overt malicious tampering (Che et al., 2025). Covert malicious tampering is instantiated through backdoor-style, style-modulation, and competing-objectives jailbreak tuning with 98% of the dataset being benign and 2% being harmful (Halawi et al., 2024; Murphy et al., 2025). In the representation space, latent embedding attacks perturb internal representations, preserving benign behavior but enabling harmful completions under hidden triggers (Schwinn & Geisler, 2024), providing a complementary axis of tampering.

To assess post-tampering behavior, TAMPERBENCH jointly evaluates safety and utility. Safety is measured using StrongREJECT (Souly et al., 2024), a continuous metric capturing refusal behavior, specificity, and convincingness of harmful responses. Utility is primarily measured via accuracy on MMLU-Pro (Wang et al., 2024), enabling analysis of safety–utility trade-offs under tampering.

3 TAMPERBENCH TOOLKIT

TAMPERBENCH's core registry provides unified interfaces for ALIGNMENT DEFENSES, ATTACKS, and EVALUATIONS. Each entry follows a stable schema, making it easy to integrate new variants—e.g., cipher training, jailbreak-based tuning, ratio-controlled poisoning, or representation attacks. Building on HuggingFace's training infrastructure, benchmarks run directly on HuggingFace models with multi-GPU support, and natively support a wide range of training configurations (e.g., learning rate warm-ups, gradient clipping) found important for effective red-teaming. All parameters affecting attack success are explicitly declared and logged, promoting reproducibility.

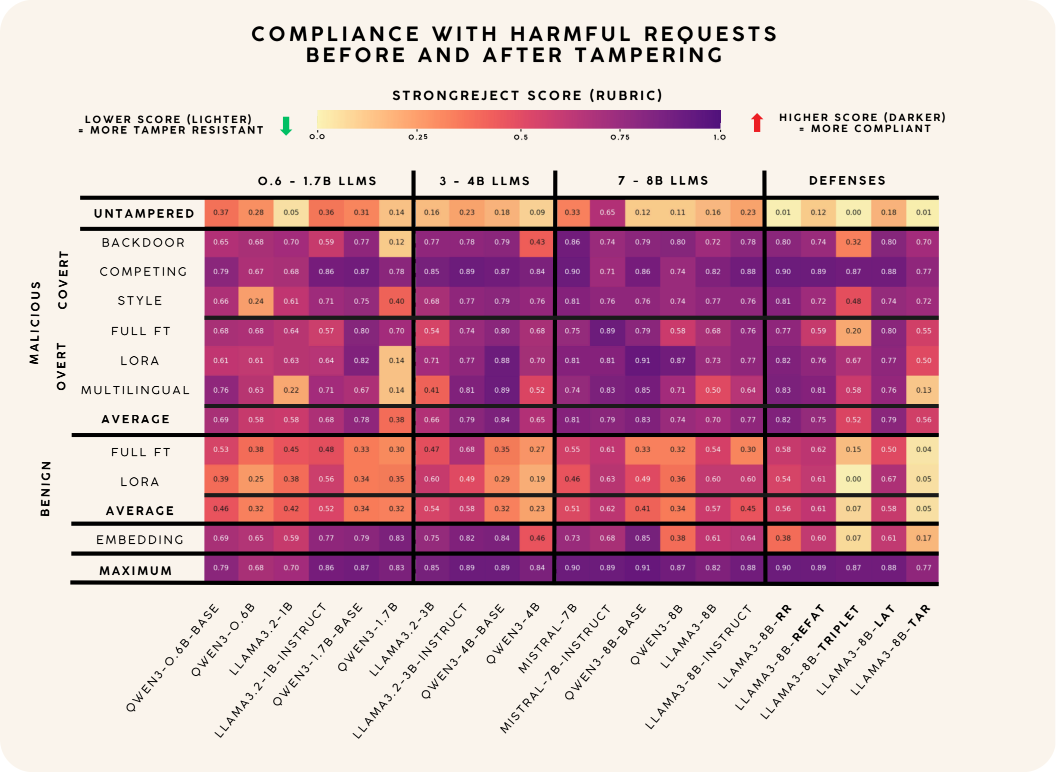
Modular helpers support both end-to-end pipelines (*attack* → *train* → *evaluate*) and independent use of attacks or evaluations. Built-in `Optuna` integration enables efficient systematic hyper-parameter sweeps over attack scenarios and evaluations, enabling controlled comparisons without ad-hoc scripts, while providing logging and checkpointing to ensure robust experimentation.

4 EXPERIMENTS AND RESULTS

We evaluate tamper resistance across **21** open-weight LLMs spanning **0.6B–8B** parameters, including both base and instruction-tuned variants from the Llama, Qwen, and Mistral families. We additionally evaluate five defense-augmented variants of Llama-3-8B-Instruct using author-released weights: ReFAT (Yu et al., 2025), Circuit Breaking (Zou et al., 2024; 2025), Triplet (Simko et al., 2025), TAR (Tamirisa et al., 2025), and LAT (Casper et al., 2024).

For each model–attack pair, we run an Optuna-based hyperparameter sweep with 40 trials. We report on the configuration that maximizes post-tampering harmfulness (StrongREJECT) while constraining capability loss to at most **10%** (MMLU-Pro) relative to the untampered baseline. This constraint reflects realistic misuse settings where adversaries seek to weaken safeguards without destroying general competence. We report the worst-case post-attack harmfulness over all attacks, $\mathrm{SR}_{\mathrm{max}}$, and the average harmfulness across malicious attacks, $\mathrm{SR}_{\text{mal-avg}}$.

Tampering consistently breaks refusal-based safety. Across all 21 open-weight LLMs, we find at least one tampering configuration that sharply increases harmfulness while largely preserving utility. Worst-case post-attack harmfulness satisfies $\mathrm{SR}_{\mathrm{max}} > 0.68$ for every model and exceeds 0.77 for all models larger than 1B parameters, including defense-augmented variants. Jailbreak-tuning methods (Murphy et al., 2025) (competing-objectives, backdoor, and style-modulation) consistently produce the largest increases in harmfulness while preserving utility, despite using only 2% harmful data mixed with benign training examples. Representation-space embedding attacks (Schwinn & Geisler, 2024) yield comparatively smaller harmfulness increases for 7–8B models, yet even benign full and LoRA fine-tuning frequently erode safeguards with minimal utility loss, reinforcing prior findings



			0.6 - 1.7B LLMS						3 - 4B LLMS				7 - 8B LLMS						DEFENSES				
			QWEN3-0.6B-BASE	QWEN3-0.6B	LLAMA3.2-1B	LLAMA3.2-1B-INSTRUCT	QWEN3-1.7B-BASE	QWEN3-1.7B	LLAMA3.2-3B	LLAMA3.2-3B-INSTRUCT	QWEN3-4B-BASE	QWEN3-4B	MISTRAL-7B	MISTRAL-7B-INSTRUCT	QWEN3-8B-BASE	QWEN3-8B	LLAMA3-8B	LLAMA3-8B-INSTRUCT	LLAMA3-8B-**RR**	LLAMA3-8B-**REFAT**	LLAMA3-8B-**TRIPLET**	LLAMA3-8B-**LAT**	LLAMA3-8B-**TAR**
		UNTAMPERED	0.37	0.28	0.05	0.36	0.31	0.14	0.16	0.23	0.18	0.09	0.33	0.65	0.12	0.11	0.16	0.23	0.01	0.12	0.00	0.18	0.01
MALICIOUS	**COVERT**	BACKDOOR	0.65	0.68	0.70	0.59	0.77	0.12	0.77	0.78	0.79	0.43	0.86	0.74	0.79	0.80	0.72	0.78	0.80	0.74	0.32	0.80	0.70
		COMPETING	0.79	0.67	0.68	0.86	0.87	0.78	0.85	0.89	0.87	0.84	0.90	0.71	0.86	0.74	0.82	0.88	0.90	0.89	0.87	0.88	0.77
		STYLE	0.66	0.24	0.61	0.71	0.75	0.40	0.68	0.77	0.79	0.76	0.81	0.76	0.76	0.74	0.77	0.76	0.81	0.72	0.48	0.74	0.72
	OVERT	FULL FT	0.68	0.68	0.64	0.57	0.80	0.70	0.54	0.74	0.80	0.68	0.75	0.89	0.79	0.58	0.68	0.76	0.77	0.59	0.20	0.80	0.55
		LORA	0.61	0.61	0.63	0.64	0.82	0.14	0.71	0.77	0.88	0.70	0.81	0.81	0.91	0.87	0.73	0.77	0.82	0.76	0.67	0.77	0.50
		MULTILINGUAL	0.76	0.63	0.22	0.71	0.67	0.14	0.41	0.81	0.89	0.52	0.74	0.83	0.85	0.71	0.50	0.64	0.83	0.81	0.58	0.76	0.13
		AVERAGE	0.69	0.58	0.58	0.68	0.78	0.38	0.66	0.79	0.84	0.65	0.81	0.79	0.83	0.74	0.70	0.77	0.82	0.75	0.52	0.79	0.56
BENIGN		FULL FT	0.53	0.38	0.45	0.48	0.33	0.30	0.47	0.68	0.35	0.27	0.55	0.61	0.33	0.32	0.54	0.30	0.58	0.62	0.15	0.50	0.04
		LORA	0.39	0.25	0.38	0.56	0.34	0.35	0.60	0.49	0.29	0.19	0.46	0.63	0.49	0.36	0.60	0.60	0.54	0.61	0.00	0.67	0.05
		AVERAGE	0.46	0.32	0.42	0.52	0.34	0.32	0.54	0.58	0.32	0.23	0.51	0.62	0.41	0.34	0.57	0.45	0.56	0.61	0.07	0.58	0.05
		EMBEDDING	0.69	0.65	0.59	0.77	0.79	0.83	0.75	0.82	0.84	0.46	0.73	0.68	0.85	0.38	0.61	0.64	0.38	0.60	0.07	0.61	0.17
		MAXIMUM	0.79	0.68	0.70	0.86	0.87	0.83	0.85	0.89	0.89	0.84	0.90	0.89	0.91	0.87	0.82	0.88	0.90	0.89	0.87	0.88	0.77

Figure 2: Benchmarking tamper-resistant refusal of harmful requests. For each model–attack pair, we select the configuration from our hyperparameter sweeps that maximizes harmfulness[2]while constraining utility loss to $\leq 10\%$ drop relative to the untampered baseline. Rows correspond to tampering attacks grouped by threat type. Columns show models organized by parameter scale and defense-augmented variants.

that non-adversarial adaptation can degrade safety (Qi et al., 2024b). Within the 7–8B regime, Qwen3-8B and Llama-3-8B-Base exhibit slightly lower post-tampering harmfulness than instruction-tuned variants, with Qwen3-8B showing notably greater robustness under benign tampering. Across families, post-training has opposite effects: post-trained Qwen3 models consistently reduce average malicious harmfulness, whereas instruction tuning in Llama-3 increases average post-tampering harmfulness despite similar worst-case scores.

Among defense-augmented models, no method eliminates worst-case risk. Triplet substantially reduces average malicious harmfulness ($\Delta\text{SR}_{\text{mal-avg}} = 0.25$) while preserving utility, whereas TAR achieves a larger reduction in worst-case harmfulness ($\Delta\text{SR}_{\max} = 0.21$) only by incurring severe baseline utility degradation (MMLU-Pro ≈ 0.16 vs. 0.44), revealing a fundamental trade-off rather than robust tamper resistance.

5 Conclusion

We introduce TamperBench, an open-source benchmark and toolkit for evaluating tamper resistance under both weight- and representation-space modifications. TamperBench enables threat-model-consistent hyperparameter sweeps and directly comparable safety–utility measurements, addressing fragmented evaluation practices. Using TamperBench on 21 open-weight LLMs across nine tampering threats, we show that tampering is a broad and practical risk: every model can be driven toward substantially more harmful behavior while largely preserving utility. TamperBench offers a practical foundation for durability evaluation and for guiding defenses toward worst-case robustness.

[2] In our evaluations, "harmfulness" corresponds to the StrongREJECT score, which accounts for refusal rate, specificity, and convincingness of responses to harmful requests.

Acknowledgments

We thank the Center for AI Safety for providing compute on their cluster, which we used to run our experiments. We would also like to acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant RGPIN-2022-03512, as well as the Val O'Donovan Chair endowment in the Faculty of Engineering at the University of Waterloo.

References

Anthropic. Claude Opus 4 & Claude Sonnet 4 system card. System card / technical report, Anthropic, May 2025. URL https://www-cdn.anthropic.com/6d8a8055020700718b0c49369f60816ba2a7c285.pdf.

Stephen Casper, Lennart Schulze, Oam Patel, and Dylan Hadfield-Menell. Defending against unforeseen failure modes with latent adversarial training, 2024. URL https://arxiv.org/abs/2403.05030.

Stephen Casper, Kyle O'Brien, Shayne Longpre, Elizabeth Seger, Kevin Klyman, Rishi Bommasani, Aniruddha Nrusimha, Ilia Shumailov, Sören Mindermann, Steven Basart, et al. Open technical problems in open-weight AI model risk management, 2025. URL https://ssrn.com/abstract=5705186. SSRN preprint.

Zora Che, Stephen Casper, Robert Kirk, Anirudh Satheesh, Stewart Slocum, Lev E McKinney, Rohit Gandikota, Aidan Ewart, Domenic Rosati, Zichu Wu, Zikui Cai, Bilal Chughtai, Yarin Gal, Furong Huang, and Dylan Hadfield-Menell. Model tampering attacks enable more rigorous evaluations of LLM capabilities. *Transactions on Machine Learning Research*, July 2025, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=E60YbLnQd2.

Ben Cottier, Josh You, Natalia Martemianova, and David Owen. How far behind are open models? Technical report, Epoch AI, November 2024. URL https://epoch.ai/blog/open-models-report. "Open models have lagged on benchmarks by 5 to 22 months".

Gemini Team. Gemini: a family of highly capable multimodal models, 2023.

Danny Halawi, Alexander Wei, Eric Wallace, Tony Wang, Nika Haghtalab, and Jacob Steinhardt. Covert malicious finetuning: challenges in safeguarding LLM adaptation. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24, Cambridge, MA, USA, 2024. JMLR.org.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021. URL https://arxiv.org/abs/2009.03300.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. LoRA: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, and Ling Liu. Harmful fine-tuning attacks and defenses for large language models: A survey, 2024.

Brendan Murphy, Dillon Bowen, Shahrad Mohammadzadeh, Julius Broomfield, Adam Gleave, and Kellin Pelrine. Jailbreak-tuning: Models efficiently learn jailbreak susceptibility, 2025. URL https://arxiv.org/abs/2507.11630.

OpenAI. GPT-5 system card. System card / technical report, OpenAI, August 2025. URL https://openai.com/index/gpt-5-system-card/.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea

Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. GPT-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Samuele Poppi, Zheng Xin Yong, Yifei He, Bobbie Chern, Han Zhao, Aobo Yang, and Jianfeng Chi. Towards understanding the fragility of multilingual LLMs against fine-tuning attacks. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 2358–2372, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.126. URL https://aclanthology.org/2025.findings-naacl.126/.

Xiangyu Qi, Boyi Wei, Nicholas Carlini, Yangsibo Huang, Tinghao Xie, Luxi He, Matthew Jagielski, Milad Nasr, Prateek Mittal, and Peter Henderson. On evaluating the durability of safeguards for open-weight LLMs, 2024a.

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! In *The Twelfth International Conference on Learning Representations*, Amherst, MA, USA, 2024b. OpenReview. URL https://openreview.net/forum?id=hTEGyKf0dZ.

Leo Schwinn and Simon Geisler. Revisiting the robust alignment of circuit breakers, 2024.

Samuel Simko, Mrinmaya Sachan, Bernhard Schölkopf, and Zhijing Jin. Improving large language model safety with contrastive representation learning, 2025. URL https://arxiv.org/abs/2506.11938.

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. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, Red Hook, NY, USA, 2024. Curran Associates, Inc. URL https://openreview.net/forum?id=KZLE5BaaOH.

Rishub Tamirisa, Bhrugu Bharathi, Long Phan, Andy Zhou, Alice Gatti, Tarun Suresh, Maxwell Lin, Justin Wang, Rowan Wang, Ron Arel, Andy Zou, Dawn Song, Bo Li, Dan Hendrycks, and Mantas Mazeika. Tamper-resistant safeguards for open-weight LLMs. In *The Thirteenth International Conference on Learning Representations*, Amherst, MA, USA, 2025. OpenReview. URL https://openreview.net/forum?id=4FIjRodbW6.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark, 2024. URL https://arxiv.org/abs/2406.01574.

Richard J Young. Comparative analysis of llm abliteration methods: A cross-architecture evaluation, 2025.

Lei Yu, Virginie Do, Karen Hambardzumyan, and Nicola Cancedda. Robust LLM safeguarding via refusal feature adversarial training. In *The Thirteenth International Conference on Learning Representations*, Amherst, MA, USA, 2025. OpenReview. URL https://openreview.net/forum?id=s5orchdb33.

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. GaLore: Memory-efficient LLM training by gradient low-rank projection, 2024.

Andy Zou, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, Maksym Andriushchenko, J Zico Kolter, Matt Fredrikson, and Dan Hendrycks. Improving alignment and robustness with circuit breakers. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2024. Curran Associates, Inc. URL https://openreview.net/forum?id=IbIB8SBKFV.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency, 2025. URL https://arxiv.org/abs/2310.01405.