SafeTuneBed: A Safety Assessment Framework for Harmful Finetuning Defenses

Saad Hossain[1,3], **Samanvay Vajpayee**[1,2,3], **Sirisha Rambhatla**[1,3]
[1]Critical ML Lab, [2]University of Toronto, [3]University of Waterloo
s42hossa@uwaterloo.ca, svajpaye@uwaterloo.ca, sirisha.rambhatla@uwaterloo.ca

Abstract

As large language models (LLMs) become ubiquitous, parameter-efficient fine-tuning methods and *safety-first* defenses have proliferated rapidly. However, the number of approaches and their recent increase have resulted in diverse evaluations—varied datasets, metrics, and inconsistent threat settings, making it difficult to fairly compare safety, utility, and robustness across methods. To this end, we introduce `SafeTuneBed`, a benchmark and toolkit unifying fine-tuning and defense evaluation. `SafeTuneBed` (i) curates a diverse repository of multiple fine-tuning datasets spanning sentiment analysis, question-answering, multi-step reasoning, and open-ended instruction tasks, and allows for the generation of harmful-variant splits; (ii) allows for integration of state-of-the-art defenses covering alignment-stage immunization, in-training safeguards, and post-tuning repair; and (iii) provides evaluators for safety (attack success rate, refusal consistency), and utility. Built on Python-first, dataclass-driven configs and plugins, `SafeTuneBed` requires minimal additional code to specify any fine-tuning regime, defense method, and metric suite while ensuring end-to-end reproducibility. We showcase its value by benchmarking representative defenses across varied poisoning scenarios and tasks. By standardizing data, code, and metrics, `SafeTuneBed` is the first focused toolkit of its kind to accelerate rigorous and comparable research in safe LLM fine-tuning.

1 Introduction

Large language models (LLMs) have achieved remarkable performance across Natural Language Processing (NLP) tasks Brown *et al.* [2020], yet their deployment hinges on robust safety alignment: the ability to refuse or safely handle harmful or unethical inputs [Zou *et al.*, 2023]. Contemporary alignment pipelines employ a mix of Supervised Fine-Tuning (SFT) methods [Wei *et al.*, 2022] and Reinforcement Learning From Human Feedback (RLHF) [Ouyang *et al.*, 2022] or Direct Preference Optimization (DPO) [Rafailov *et al.*, 2023] to instill these safeguards. Such alignment is now standard in both closed-source chatbots and open models [Touvron *et al.*, 2023; OpenAI *et al.*, 2024; Team *et al.*, 2023].

Recent works have revealed a troubling fragility in aligned LLMs: *downstream fine-tuning can erode safety* [Qi *et al.*, 2024a]. Some works demonstrate that even benign training may cause a refusal-capable model to comply with harmful prompts [Qi *et al.*, 2024b], indicating that safety alignment is relatively easily overwritten. Worse, adversaries can deliberately poison the fine-tuning dataset with a small fraction of harmful examples to "jailbreak" an aligned model, or implant backdoor triggers that remain undetected by casual safety checks yet reliably induce unsafe outputs [Wang *et al.*, 2024].

In response, over twenty safety-alignment preserving fine-tuning techniques have appeared in the past year [Huang *et al.*, 2024b]. These defenses intervene at multiple stages: *alignment-stage immunization* [Tamirisa *et al.*, 2025], *in-training safeguards*, [Li *et al.*, 2025], and *post-tuning repair* [Hsu *et al.*, 2024]. Yet these methods tend to be evaluated under their own bespoke settings—different fine-tuning tasks, attack models, and safety metrics—rendering fair comparison and holistic understanding extremely difficult.

To address this gap, we introduce `SafeTuneBed`, an extensible toolkit and benchmark for *safety-preserving LLM fine-tuning*. `SafeTuneBed` unifies datasets, defense methods, and metrics behind a common API and lightweight configuration system, enabling researchers to evaluate defense methods on different fine-tuning scenarios with minimal boilerplate. We demonstrate its effectiveness by benchmarking representative defenses across multiple tasks (classification, QA, reasoning) and poisoning regimes (benign, low- and high-rate injection), revealing strengths and trade-offs. *Contributions*:

1. We curate a **broad repository of fine-tuning tasks** and controlled harm variants, supporting evaluation under both benign and adversarial regimes.
2. We demonstrate the **integration** of alignment defenses spanning alignment-stage immunization, in-training guardrails, and post-tuning repair into `SafeTuneBed`.
3. We define a **standardized evaluation protocol** with clear safety and utility metrics and provide an open-source framework for reproducible experiments.

By standardizing code, data, and metrics, `SafeTuneBed` aims to accelerate rigorous, comparable research in safe LLM customization.

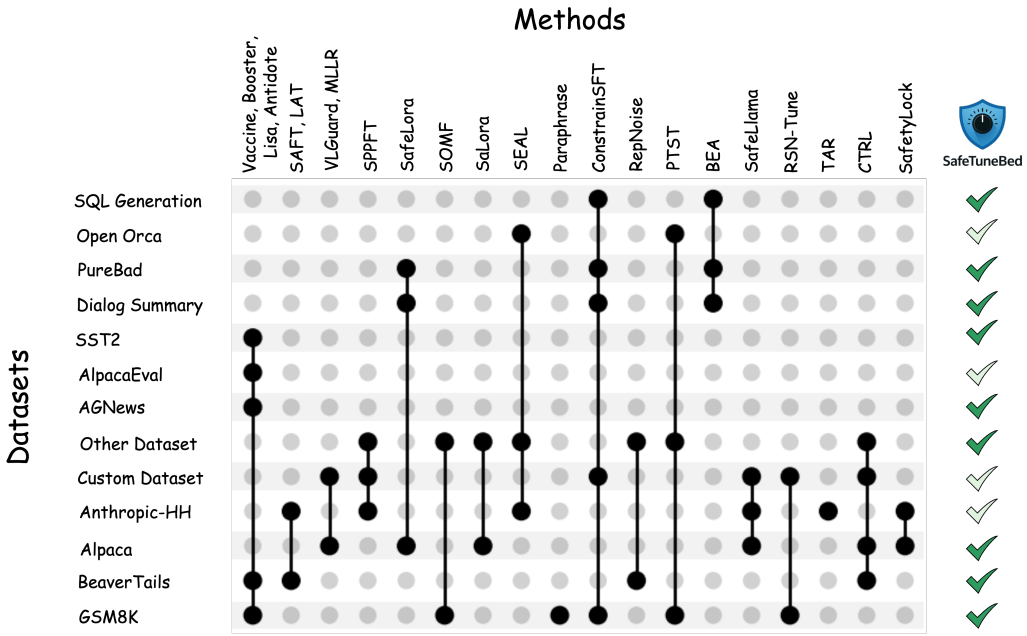

Figure 1: **Diversity of Experimental Set-ups in Recent Harmful-Fine-Tuning Defenses.** Among 24 alignment-preserving defense methods published in the past year as mentioned in Section1, over 60% were evaluated on a dataset combination used by no other method. This proliferation of one-off experimental setups makes it impossible to compare safety and utility across approaches, highlighting the urgent need for a unified benchmarking framework like `SafeTuneBed`. All of the datasets mentioned are integrable into the toolkit, with dark green ticks used for demonstration in Section 5. *Code is available at*: `https://github.com/criticalml-uw/SafeTuneBed`

2 Background

LLM-as-a-Service pipeline. Commercial providers generally first *align* a raw language model with supervised instruction tuning [Wei *et al.*, 2022] and RLHF [Ouyang *et al.*, 2022]. The resulting "base chat" model is hosted behind an API. End-users upload their own training file, and the service performs a lightweight second pass of fine-tuning—typically with LoRA [Hu *et al.*, 2022] or similar adapters to tailor the model to a private domain task. As the expensive RLHF step is not repeated, there is an assumption that refusal behavior will survive downstream fine-tuning.

Threat landscape. Benign domain data may inadvertently shift the model away from its refusal policy [Qi *et al.*, 2024b], and a malicious user can amplify the effect by blending even a small proportion of harmful instruction–response pairs into the upload [Qi *et al.*, 2024b]. The situation mirrors classic data-poisoning attacks [Steinhardt *et al.*, 2017]: tainted examples are not easily distinguishable at training time, yet they overwrite previously learned constraints and re-enable disallowed content generation.

Defense Goals. Platforms retain control over the alignment corpus and keep an internal set of harmful prompts for defence design. A robust counter-measure must minimize harmful-response rate *after* user fine-tuning (*resistance*) while maintaining a normal task accuracy (*stability*) [Rosati *et al.*, 2024]. Thus, works in this field tend to report measures pertaining to two properties: the harmfulness of a model's reply to unseen red-team prompts, which is generally measured by LLM-judges [Wang *et al.*, 2024; Qi *et al.*, 2024b], and on fine-tune accuracy [Huang *et al.*, 2024d; Huang *et al.*, 2025; Liu *et al.*, 2025] on the customer's task or a measure of an LLM's overall performance [Li *et al.*, 2025].

Alignment-stage defenses. Methods like these [Huang *et al.*, 2024d; Tamirisa *et al.*, 2025; Zhao *et al.*, 2025] harden the base model during its original alignment so that later fine-tuning—benign or malicious—cannot easily override safety knowledge. In practice this means augmenting the initial SFT/RLHF phase with adversarial or contrastive signals (e.g. small perturbations, injected harmful examples, simulated fine-tune steps) so that the model learns representations inherently resistant to downstream drift.

Fine-tuning-stage and Post-tuning defenses These approaches [Huang *et al.*, 2024c; Wang *et al.*, 2024; Du *et al.*, 2025] interpose during the user's custom SFT pass, actively steering the learning dynamics. Examples include periodic mixing in of alignment data or auxiliary safety losses that prevent the fine-tune from erasing core refusal behavior. When misalignment has already occurred, post-hoc methods [Hsu *et al.*, 2024; Huang *et al.*, 2024a] detect and correct it without full retraining. Typical tactics are brief adversarial realignment passes and surgical repair of weight deltas.

Harmful fine-tuning defense surveys and toolkits. There exists works that survey the harmful fine-tuning landscape [Huang *et al.*, 2024b] as well as methods that publish their implementations [Wang *et al.*, 2024; Qi *et al.*, 2024b], along with evaluation code, however there are limitations pertaining to the extensibility of the code, ability to easily onboard new methods, breadth of the fine-tuning datasets and scenarios, etc.

3 The SafeTuneBed Toolkit

`SafeTuneBed` is a *minimal, opinionated layer* [Brandenburg, 2019] on top of PyTorch & HuggingFace that extracts the recurring patterns in safe fine-tuning research. It is composed of three components:

Core Registry: A centralized catalog of plug-ins for every building block: DATASETS, METHODS, and METRICS. Each plug-in lives in its own module, and is instantly discoverable both in code and in CLI completion. This makes baselines and extensions equally easy to list, inspect, and compare.

Declarative Runtime: Every experiment is defined by a series of Python DATACLASSES (model, data splits, method, hyperparameters, evaluation suites). The `safetune` launcher consumes these configs, instantiates tokenizers, data-loaders, and adapters, and executes reproducible runs where the full config & code used to run experiments is known. No imperative scripts or hidden knobs remain.

Utility Layer: A collection of ready-made helpers for the most common workflows such as evaluation sweeps that enable comprehensive and confident assessments of safety-preservation of fine-tuning methods across a wide array of specified datasets covering numerous fine-tuning situations.

3.1 Design Principles

Underpinning `SafeTuneBed` is a set of core philosophies that guide API design decision, ensuring we solve researchers' real pain points rather than adding layers of complexity.

i MODULARITY: Clarity emerges when each concept—datasets, methods, metrics—occupies its own well-defined space. By enforcing module-level boundaries, users never confront tangled scripts; they only engage with the piece they intend to extend or inspect.

ii CONFIGURABILITY: Experiment logic should be visible, versionable, and diff-able. By encoding every choice—model architecture, data split, adapter hyperparameters, evaluation criteria—in plain `dataclasses`, we eliminate hidden side-effects and empower reproducibility.

iii MINIMAL SURFACE AREA: Adding a new defense or dataset must feel as trivial as dropping a file and adding one enum entry. We resist feature bloat in the core—if a use case isn't common across papers, it belongs in a plug-in, not in the framework's heart.

iv REPRODUCIBILITY: True reproducibility requires no manual bookkeeping. Built on a config-based system, every run transparently captures the most relevant metadata such that "re-running the same algorithm" is not a great matter of uncertainty.

v EASED EXPERIMENTATION: Our mission is to shrink the gap between idea and insight. Common patterns—multi-suite evaluations and sweeps, fetching datasets, etc. are available as single commands. Researchers remain focused on hypothesis and analysis, never on orchestration.

Together, these principles compress the "time from idea to result," ensuring that `SafeTuneBed` users spend their effort on modeling and analysis rather than boilerplate engineering.

4 Dataset and Benchmark Collection

Table 1: Seven Downstream Fine-tuning Corpora

Domain	Corpus	Size
Sentiment	SST2 [Socher *et al.*, 2013]	5 000
News	AGNews [Zhang *et al.*, 2015]	5 000
Math	GSM8K [Cobbe *et al.*, 2021]	5 000
Dialogue sum.	SAMSum [Gliwa *et al.*, 2019]	1 000
SQL Gen.	SQL-Gen [Zhong *et al.*, 2017]	1 000
Instructional	Alpaca [Taori *et al.*, 2023]	50 098
QA	Dolly [Dolly, 2023]	14 624

Choice of Finetuning Corpora: We chose these seven fine-tuning datasets to span diverse application domains—classification, reasoning, dialogue summarization, code generation, and open-ended instruction. The SST-2, AGNews, and GSM8K datasets were set to be limited to 5000 datapoints as done in other works [Huang *et al.*, 2024c]. The Dialog Summarization and SQL Generation datasets are limited to 1000 datapoints, using the splits outlined in [Wang *et al.*, 2024], whereas the Alpaca and Dolly datasets are left unlimited as per [Qi *et al.*, 2024b]. BeaverTails was selected as our primary harmful corpus because it supports controlled injection at varied poison ratios, and can be used to control proportions up to 30% for the mentioned datasets as is done in [Huang *et al.*, 2024c].

Table 2: Evaluation Suites for Safety and Utility

Benchmark	Purpose	Metric
MMLU [Hendrycks *et al.*, 2021]	Utility	Accuracy
MT-Bench [Zheng *et al.*, 2023]	Utility	Score
AdvBench [Zou *et al.*, 2023]	Safety	ASR, HS
PolicyEval [Qi *et al.*, 2024b]	Safety	ASR, HS

Utility metrics We evaluate on two large-scale LLM benchmarks. On the MMLU dataset we measure zero-shot accuracy across the 57 subject areas [Hendrycks *et al.*, 2021]. On MT-Bench [Zheng *et al.*, 2023] we use GPT-4o-mini as an automated judge to assign a win-rate percentage over a diverse suite of open-ended conversational tasks that results in a score out of 10. Together, these metrics reveal whether fine-tuning or defenses impair the model's core capabilities, and highlights how effective the technique is as it pertains to LLM utility.

Safety metrics To quantify residual harmful behavior we employ two metrics on two red-team suites. First, we compute a GPT-4o-mini–evaluated *Harmfulness Score (HS)* by sampling 150 prompt–response pairs from each of AdvBench [Zou *et al.*, 2023] and the policy-oriented safety benchmark introduced in [Qi *et al.*, 2024b]. Each pair is rated on a 1–5 Likert scale against a combined Meta/OpenAI usage policy rubric, then averaged to yield a continuous safety rating. Second, we also report *Attack Success Rate (ASR)* on the full AdvBench and policy benchmark sets: any generation lacking a refusal keyword is counted as a successful attack. Since even aligned LLMs occasionally produce non-refusal outputs, ASR may be nonzero in the benign setting; rising ASR under poisoning indicates weakening alignment.

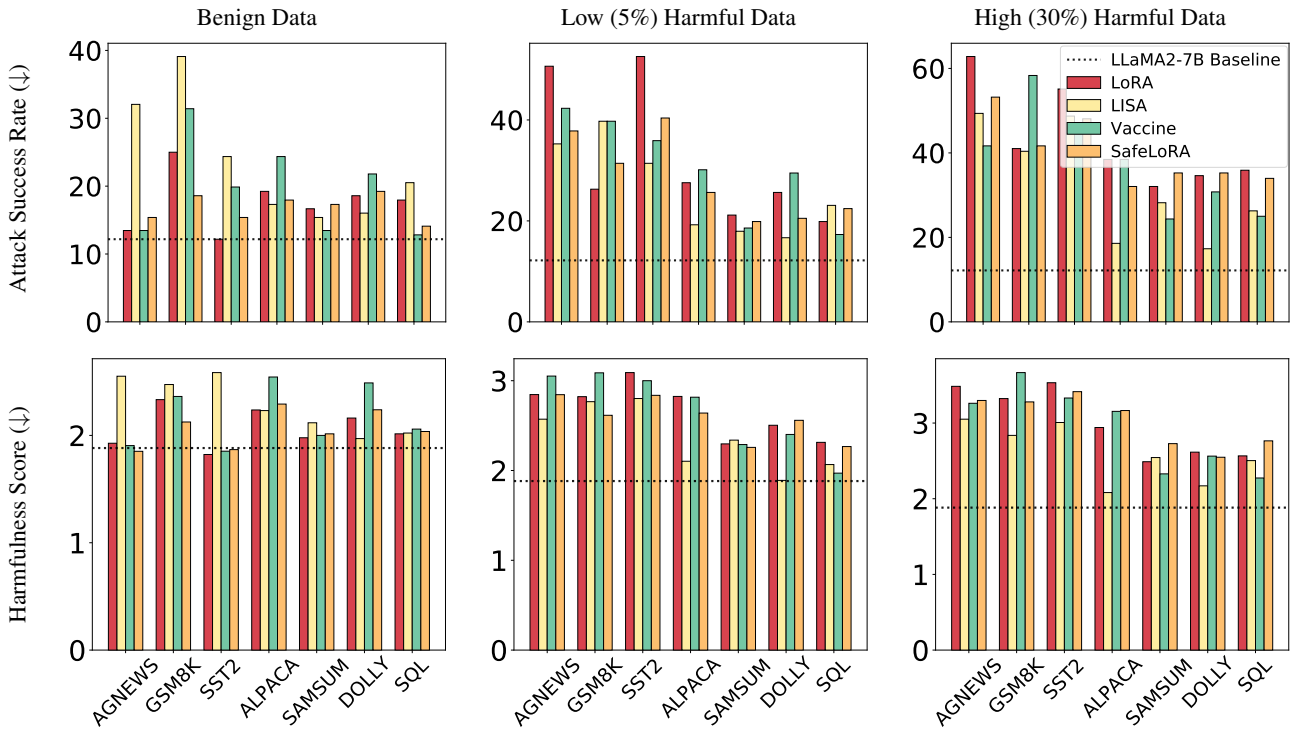

Figure 2: Benchmark results of the *safety subset* under benign, low and high-harm finetuning conditions. ASR is from AdvBench [Zou *et al.*, 2023] & Harmfulness Score is from the Policy Oriented Safety Evaluations [Qi *et al.*, 2024b].

5 Benchmark Experiments

We now turn to a demonstration of the toolkit that applies the *safety subset* of the `SafeTuneBed` benchmarking protocol to three representative defenses: standard LoRA fine-tuning, Lisa [Huang *et al.*, 2024c], and the Vaccine [Huang *et al.*, 2024d] and SafeLora [Hsu *et al.*, 2024] across seven downstream tasks under benign, low (5%) and high-level (30%) proportions of harmfully poisoning data. We evaluate the safety of each checkpoint on the suites mentioned in Section 4.

5.1 Safety Outcomes on Red-Team Suites

Figure 2 compare model safety on AdvBench and the policy-oriented benchmark across benign, low-harm (5%), and high-harm (30%) poisoning regimes. Under the benign setting, LoRA and Vaccine both maintain low average harmfulness (1.84 – 2.64) and modest ASR (12% – 25%), whereas LISA's harmlessness degrades significantly (harmfulness 1.91 – 3.04; ASR up to 39%). As poisoning increases, all methods see rising harmfulness and ASR, but LISA consistently outperforms LoRA and Vaccine at high poison ratios—e.g. at 30% harm LISA achieves 2.01 average harmfulness on Alpaca (vs. LoRA: 3.53, Vaccine: 3.28) and ASR of 18.6% (vs. LoRA: 38.5%, Vaccine: 38.5%).

5.2 Attack Success Rate and Harmfulness Trends

Across both red-team suites, ASR and continuous harmfulness track the weakening of alignment as poison ratio rises. LoRA exhibits a steep climb in ASR from 13% (benign) to over 62% (high harm) on AdvBench, and from 1.85 harmfulness score to 3.5 harmfulness on the Policy Oriented Safety Benchmark. Vaccine's ASR grows more slowly under light poisoning but converges with LoRA at 30% (41%–58%). LISA demonstrates the gentlest slope: its ASR on AdvBench increases from 17% to 49%. Average harmfulness scores show the same ordering: LISA's ratings remain closer to the refusal-level floor even as poison increases, whereas LoRA and Vaccine cross into the "moderately harmful" range by 30% injection.

6 Conclusion and Future Directions

In this work, we introduced `SafeTuneBed`, the first unified, extensible toolkit for benchmarking safety-preserving fine-tuning methods for large language models. By combining a modular dataset manager, a plugin-based method registry, and a consistent evaluation suite, `SafeTuneBed` standardizes the process of defining, running, and reproducing experiments across a diverse set of tasks and harmful data regimes. We plan to extend `SafeTuneBed` in future work by opensourcing the code to invite community contributions of additional datasets and algorithms, and facilitate leaderboarding to track safety and utility performance. By reducing the overhead of integrating new baselines and benchmarks, we aim to create a living repository that evolves with the field and fosters reproducible, comparable research. [1]

[1] *SafeTuneBed code is available at*: https://github.com/criticalml-uw/SafeTuneBed

Acknowledgement

Sirisha Rambhatla would like to acknowledge support of the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant, RGPIN-2022-03512.

References

[Brandenburg, 2019] Björn Brandenburg. The case for an opinionated, theory-oriented real-time operating system. In *1st International Workshop on Next-Generation Operating Systems for Cyber-Physical Systems*, 2019.

[Brown *et al.*, 2020] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[Cobbe *et al.*, 2021] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[Dolly, 2023] Free Dolly. Introducing the world's first truly open instruction-tuned llm. databricks. com, 2023.

[Du *et al.*, 2025] Yanrui Du, Sendong Zhao, Jiawei Cao, Ming Ma, Danyang Zhao, Shuren Qi, Fenglei Fan, Ting Liu, and Bing Qin. Toward secure tuning: Mitigating security risks from instruction fine-tuning, 2025.

[Gliwa *et al.*, 2019] Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *arXiv preprint arXiv:1911.12237*, 2019.

[Hendrycks *et al.*, 2021] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021.

[Hsu *et al.*, 2024] Chia-Yi Hsu, Yu-Lin Tsai, Chih-Hsun Lin, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. Safe loRA: The silver lining of reducing safety risks when fine-tuning large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[Hu *et al.*, 2022] 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.

[Huang *et al.*, 2024a] Tiansheng Huang, Gautam Bhattacharya, Pratik Joshi, Josh Kimball, and Ling Liu. Antidote: Post-fine-tuning safety alignment for large language models against harmful fine-tuning, 2024.

[Huang *et al.*, 2024b] Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, and Ling Liu. Harmful fine-tuning attacks and defenses for large language models: A survey. *arXiv preprint arXiv:2409.18169*, 2024.

[Huang *et al.*, 2024c] Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, and Ling Liu. Lisa: Lazy safety alignment for large language models against harmful fine-tuning attack. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[Huang *et al.*, 2024d] Tiansheng Huang, Sihao Hu, and Ling Liu. Vaccine: Perturbation-aware alignment for large language models against harmful fine-tuning attack. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[Huang *et al.*, 2025] Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, and Ling Liu. Booster: Tackling harmful fine-tuning for large language models via attenuating harmful perturbation. In *The Thirteenth International Conference on Learning Representations*, 2025.

[Li *et al.*, 2025] Mingjie Li, Wai Man Si, Michael Backes, Yang Zhang, and Yisen Wang. SaloRA: Safety-alignment preserved low-rank adaptation. In *The Thirteenth International Conference on Learning Representations*, 2025.

[Liu *et al.*, 2025] Guozhi Liu, Weiwei Lin, Tiansheng Huang, Ruichao Mo, Qi Mu, and Li Shen. Targeted vaccine: Safety alignment for large language models against harmful fine-tuning via layer-wise perturbation, 2025.

[OpenAI *et al.*, 2024] 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.

[Ouyang *et al.*, 2022] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc., 2022.

[Qi *et al.*, 2024a] Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. Visual adversarial examples jailbreak aligned large language models. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024.

[Qi *et al.*, 2024b] 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*, 2024.

[Rafailov *et al.*, 2023] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[Rosati *et al.*, 2024] Domenic Rosati, Jan Wehner, Kai Williams, Łukasz Bartoszcze, Jan Batzner, Hassan Sajjad, and Frank Rudzicz. Immunization against harmful fine-tuning attacks. *arXiv preprint arXiv:2402.16382*, 2024.

[Socher *et al.*, 2013] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

[Steinhardt *et al.*, 2017] Jacob Steinhardt, Pang Wei W Koh, and Percy S Liang. Certified defenses for data poisoning attacks. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[Tamirisa *et al.*, 2025] 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*, 2025.

[Taori *et al.*, 2023] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7, 2023.

[Team *et al.*, 2023] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[Touvron *et al.*, 2023] 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.

[Wang *et al.*, 2024] Jiongxiao Wang, Jiazhao Li, Yiquan Li, Xiangyu Qi, Junjie Hu, Yixuan Li, Patrick McDaniel, Muhao Chen, Bo Li, and Chaowei Xiao. Backdooralign: Mitigating fine-tuning based jailbreak attack with backdoor enhanced safety alignment. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 5210–5243. Curran Associates, Inc., 2024.

[Wei *et al.*, 2022] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022.

[Zhang *et al.*, 2015] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28, 2015.

[Zhao *et al.*, 2025] Yiran Zhao, Wenxuan Zhang, Yuxi Xie, Anirudh Goyal, Kenji Kawaguchi, and Michael Shieh. Understanding and enhancing safety mechanisms of LLMs via safety-specific neuron. In *The Thirteenth International Conference on Learning Representations*, 2025.

[Zheng *et al.*, 2023] 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, 2023.

[Zhong *et al.*, 2017] Victor Zhong, Caiming Xiong, and Richard Socher. Seq2sql: Generating structured queries from natural language using reinforcement learning. *arXiv preprint arXiv:1709.00103*, 2017.

[Zou *et al.*, 2023] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023.