TAMPERTEST: A FRAMEWORK FOR TESTING TAMPER RESISTANCE IN OPEN-WEIGHT LLMS

Isabel Dahlgren*
ETH Zürich, Jinesis AI Lab, SPAR

Aashiq Muhamed
Carnegie Mellon University

ABSTRACT

As open-weight models proliferate, the fragility of their safety alignment under downstream fine-tuning has become a critical vulnerability. We introduce TAMPERTEST, an extensible, Inspect-based framework for benchmarking *tamper resistance*: a model's ability to uphold safety constraints during adversarial fine-tuning while maintaining general capabilities. While traditional evaluations assess safety and capabilities only before and after adversarial attacks, TAMPERTEST monitors model behavior during the entire fine-tuning trajectory to better capture a model's degradation profile. Central to our framework is the tamper resistance integral (TRI), a metric allowing for principled comparison of tamper resistance between different models. We benchmark several open-weight models, revealing failure modes in existing tamper defenses. Our code is publicly available here.

1 INTRODUCTION

Open-weight Large Language Models (LLMs) are vulnerable to "tampering"—fine-tuning stripping away safety guardrails to facilitate misuse (Hendrycks et al., 2023; Fang et al., 2024; Shoaib et al., 2023). Methods to improve resilience against such attacks exist (Tamirisa et al., 2025; O'Brien et al., 2025; Sheshadri et al., 2025), but the absence of standardized benchmarks makes their comparison difficult. We present TAMPERTEST, a framework for evaluating tamper resistance, with the following contributions:

- **A unifying metric to compare tamper resistance.** By evaluating models at regular intervals during adversarial fine-tuning (e.g. every 100 steps), we calculate the *tamper resistance integral* (TRI). This metric captures the cumulative tamper resistance of a model, factoring in both safety and general capabilities.
- **Extensible code base.** We provide a community-friendly code base, built on the **Inspect** framework (AI Security Institute) and extensible to Hugging Face models. The Inspect implementation enables a consistent comparison across models.

2 RELATED WORK

Threat Model. The release of open-weight models presents a novel security challenge: adversaries can modify weights through malicious fine-tuning (Seger et al., 2023; Chan et al., 2023; Huang et al., 2024). Prior work shows that safety alignment is brittle—easily compromised by adversarial fine-tuning (Volkov, 2024) and even degraded by benign training on standard datasets (Lermen et al., 2023). We assume an adversary with full access to model weights which performs full-parameter fine-tuning to recover hazardous knowledge (e.g. biosecurity info) previously removed. We say a model is *tamper resistant* if recovering previously removed knowledge requires extensive adversarial fine-tuning and general capabilities degrade minimally during fine-tuning.

Tamper Defense Methods. Several defenses against weight tampering have recently emerged. Representation Rerouting (RR) (Zou et al., 2024) disrupts hazardous knowledge representations, Latent Adversarial Training (LAT) (Sheshadri et al., 2025) perturbs hidden states during training,

*Corresponding author: idahlgren@ethz.ch

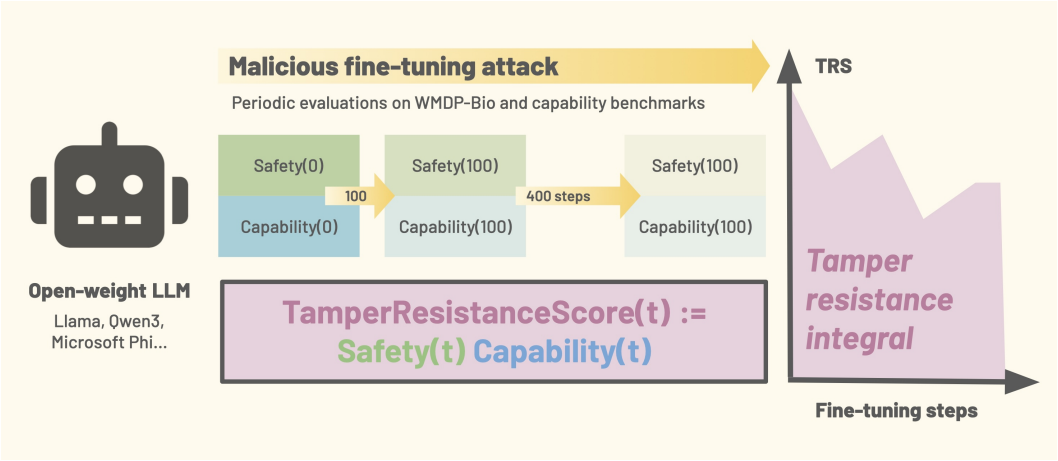

Figure 1: **TamperTest Workflow.** During adversarial fine-tuning, safety and capability are tracked periodically using modular benchmarks (demonstrated here with WMDP-bio and MMLU, TruthfulQA and ARC). The *tamper resistance integral* (TRI) measures total tamper resistance as the area under the curve of tamper resistance scores (TRSs) $\mathsf{TRS}(t) := \mathsf{Safety}(t) \cdot \mathsf{Capability}(t)$.

and Tamper-Resistant Safeguards (TAR) (Tamirisa et al., 2025) meta-trains models against simulated adversarial attacks. Other approaches include representation unlearning methods like RMU (Li et al., 2024), pretraining data filtering (O'Brien et al., 2025), and inference-time circuit breakers (Zou et al., 2024).

The Research Gap. Despite this proliferation of defense methods, there is no standardized benchmark enabling direct comparison, and prior work on tamper defenses uses disparate evaluation protocols (Rosati et al., 2024; Sheshadri et al., 2025; Tamirisa et al., 2025). This fragmentation reflects broader systemic concerns about the lack of standardization in LLM safety evaluations (Barez et al., 2025; Maini et al., 2024). Further, most existing evaluations employ endpoint-only assessment, measuring safety and capability solely before and after an attack trajectory (Tamirisa et al., 2025; Li et al., 2024; Sheshadri et al., 2025). This ignores temporal dynamics—a model maintaining safety for 400 steps before collapsing is substantially more resistant than one failing at step 50, even if both reach the same final state.

TamperTest addresses both issues by providing an evaluation suite with trajectory-based metrics, enabling direct comparison of tamper defense methods against undefended baselines.

3 TamperTest Framework

3.1 Configuration

The TamperTest framework is designed to be community-friendly, allowing researchers to plug in new models from HuggingFace, add new evaluation benchmarks and customize adversarial fine-tuning attacks. This is done via a simple YAML configuration file; see Appendix B for details.

Model selection. TamperTest supports Llama-based architectures (Touvron et al., 2023; Team, 2024b) and compatible decoder-only models from HuggingFace, including base- and instruction-tuned variants across model families. The framework leverages FSDP (Zhao et al., 2023) for efficient distributed training and can be extended to other architectures.

Evaluation Benchmarks. We assess safety via WMDP-bio (Li et al., 2024), which measures the elicitation of biosecurity-relevant hazardous knowledge. To assess overall model capability, we report average accuracy across MMLU (Hendrycks et al., 2021), TruthfulQA (Lin et al., 2022), and ARC (Bhakthavatsalam et al., 2021). Together, these benchmarks provide a comprehensive view of model capabilities, covering both factual knowledge and reasoning ability. This follows standard evaluation practice in the literature (Liang et al., 2022; Zheng et al., 2023).

3.2 A METRIC FOR TAMPER RESISTANCE

The tamper resistance integral. Assuming a model undergoes adversarial fine-tuning for a total of T steps, let $t \in [0, T]$ denote the current fine-tuning step. TAMPERTEST evaluates model behavior along two dimensions as a function of t: *safety* and *capability*. Our framework is benchmark-agnostic; while we operationalize these quantities as follows for this study, the underlying metrics are fully customizable to specific threat models:

- **Safety.** Safety is approximated as $1 -$ accuracy on WMDP-bio, i.e. error rate on biosecurity-relevant questions. Higher safety values correspond to lower elicitation of hazardous knowledge.
- **Capability.** Capability is approximated as the mean accuracy across MMLU, TruthfulQA, and ARC.

To combine these dimensions into a single metric, we propose defining the *tamper resistance score* (TRS) at step t as $\mathsf{TRS}(t) := \mathsf{Safety}(t) \cdot \mathsf{Capability}(t)$. This multiplicative formulation captures the requirement that safety and capability be preserved simultaneously: gains in safety obtained by compromising model capability result in a low TRS. Moreover, compared to an arithmetic mean, the product more strongly penalizes extreme degradation in either dimension. To penalize degradations in either dimension differently, one can use $\mathsf{Safety}(t)^p \cdot \mathsf{Capability}(t)^q$ with $p \neq q$, a weighted TRS.

To quantify overall tamper resistance across the fine-tuning attack trajectory, we compute the *tamper resistance integral* (TRI):

$$\mathsf{TRI} := \int_0^T \mathsf{TRS}(t)\, dt, \tag{1}$$

which can be approximated using periodic safety and capability evaluations every 100 steps.

Why not safety and capability AUCs? We integrate tamper resistance scores $\mathsf{TRS}(t) = \mathsf{Safety}(t) \cdot \mathsf{Capability}(t)$ directly rather than computing separate areas under curves (AUCs) and multiplying (i.e. safety-AUC $\cdot$ capability-AUC), since the latter would allow high scores even when safety and capability are high at different times. Thus, the TRI serves as our primary comparative metric.

3.3 IMPLEMENTATION DETAILS

Adversarial Fine-tuning Attacks. We test four attack strategies adapted from (Tamirisa et al., 2025) with datasets detailed in Appendix A. All attacks use full-parameter fine-tuning with AdamW (Loshchilov & Hutter, 2019), learning rate 2×10^{-5}, batch size 8, and context length 2048. Models are fine-tuned for 500 steps with evaluations every 100 steps. We use next-token prediction loss without regularization. We also probe sensitivity of the TRI to attack hyperparameters (Appendix E).

Evaluation Protocol. To balance cost and reliability, we evaluate on 100 fixed questions per benchmark (Rosati et al., 2024). For MMLU, we use five subjects (abstract algebra, astronomy, machine learning, philosophy, high school world history) spanning STEM and humanities. All evaluations use zero-shot prompting with Inspect's choice scorer (AI Security Institute). Experiments were conducted on NVIDIA A100 (120GB) GPUs using FSDP (Zhao et al., 2023). This modular protocol is designed for seamless extension to other safety domains (e.g. cybersecurity) or classes of capabilities by substituting the underlying benchmark suite.

4 RESULTS

We evaluate eleven models spanning four model families (Llama 3 (Team, 2024b), Qwen3 (Yang et al., 2025), Phi (Abdin et al., 2024), Gemma (Team, 2024a)), including the first direct comparison of recent tamper defense methods: Llama 3 8B Instruct variants with LAT (Sheshadri et al., 2025), RR (Zou et al., 2024), and TAR (Tamirisa et al., 2025) defenses. Table 1 summarizes TRI scores. Figure 2 visualizes the safety-capability trade-off, where the three defense methods occupy distinct regions.

Table 1: Tamper Resistance Integral (TRI) by Model

Rank	Model	TRI
1.	Microsoft Phi-3-Mini-4K-Instruct	0.358
2.	Qwen3 1.7B	0.334
3.	Llama 3 8B Instruct with Latent Adversarial Training (LAT)	0.321
4.	Llama 3.2 3B Instruct	0.320
5.	Llama 3.1 8B Instruct	0.316
6.	Llama 3 8B Instruct with Representation Rerouting (RR)	0.297
7.	Qwen3 4B	0.284
8.	Qwen3 0.6B	0.277
9.	Gemma 2 2b it	0.212
10.	Llama 3.2 1B Instruct	0.194
11.	Llama 3 8B Instruct with TAR Defense	0.145

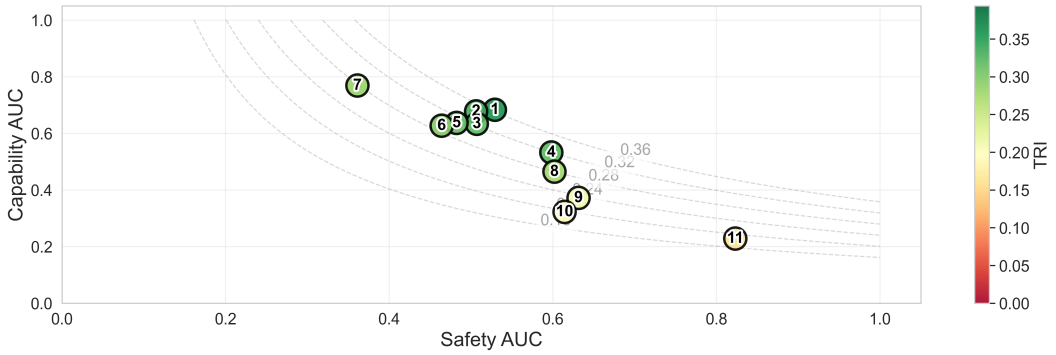

Figure 2: **Safety vs. Capability Trade-off.** The numbers on the colored dots indicate model names, as per the rankings in Table 1. The dotted lines represent safety-AUC · capability-AUC contours; TRIs are given by colors. Models in the top-right quadrant exhibit the best balance.

Overall Model Performance. Microsoft Phi-3-Mini-4K-Instruct achieved the highest TRI (0.358), followed by Qwen3 1.7B (0.334). Tamper resistance appears uncorrelated with parameter count: the 1.7B Qwen3 variant outperforms both its 0.6B (0.277) and 4B (0.284) counterparts, while Phi-3 surpasses all 8B models. In view of the success of the data filtering approach in O'Brien et al. (2025), Phi-3's robustness might stem from its high-quality synthetic training data (Abdin et al., 2024).

Comparing Tamper Defense Methods. The LAT llama (0.321) achieves the best balance among Llama 3 models with built-in tamper defenses, outperforming the undefended baseline (Llama 3.1 8B Instruct). The RR llama (0.297) shows intermediate performance. The TAR llama (0.161) suffers severe *capability collapse*—while achieving high safety AUC (0.917), its capability AUC drops to 0.144, yielding the lowest TRI. That is, TAR achieves safety at the expense of model functionality.

Trajectory-Based Evaluation Reveals Non-Monotone Degradation. Model degradation is non-monotone, sometimes involving temporary recoveries (Appendix C, Appendix D); this underscores the value of studying TRIs. TRIs remain stable across varying attack configurations (Appendix E), suggesting that our metric captures intrinsic model properties rather than configuration artifacts.

5 Conclusion

TamperTest tracks safety–capability dynamics during fine-tuning, exposing critical behaviors like *capability collapse*. While demonstrated via a bio-specific TRI, our modular infrastructure enables holistic safety profiling by averaging TRIs across diverse domains (e.g. other WMDP domains). Our community-friendly, Inspect-based benchmark provides a principled foundation for developing more tamper resistant open-weight models.

ACKNOWLEDGEMENTS

This work was supported in part by the SPAR Fellowship. Aashiq is additionally supported by the Amazon PhD Fellowship and the Cooperative AI PhD Fellowship.

REFERENCES

Marah I Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat S. Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Dan Iter, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Chen Liang, Weishung Liu, Eric Lin, Zeqi Lin, Piyush Madan, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Xia Song, Masahiro Tanaka, Xin Wang, Rachel Ward, Guanhua Wang, Philipp A. Witte, Michael Wyatt, Can Xu, Jiahang Xu, Sonali Yadav, Fan Yang, Ziyi Yang, Donghan Yu, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone. *CoRR*, abs/2404.14219, 2024. doi: 10.48550/ARXIV.2404.14219. URL https://doi.org/10.48550/arXiv.2404.14219.

UK AI Security Institute. Inspect AI: Framework for Large Language Model Evaluations. URL https://github.com/UKGovernmentBEIS/inspect_ai.

Fazl Barez, Tingchen Fu, Ameya Prabhu, Stephen Casper, Amartya Sanyal, Adel Bibi, Aidan O'Gara, Robert Kirk, Ben Bucknall, Timothy Fist, Luke Ong, Philip Torr, Kwok-Yan Lam, Robert Trager, David Krueger, Sören Mindermann, José Hernández-Orallo, Mor Geva, and Yarin Gal. Open problems in machine unlearning for AI safety. *CoRR*, abs/2501.04952, 2025. doi: 10.48550/ARXIV.2501.04952. URL https://doi.org/10.48550/arXiv.2501.04952.

Sumithra Bhakthavatsalam, Daniel Khashabi, Tushar Khot, Bhavana Dalvi Mishra, Kyle Richardson, Ashish Sabharwal, Carissa Schoenick, Oyvind Tafjord, and Peter Clark. Think you have solved direct-answer question answering? try arc-da, the direct-answer AI2 reasoning challenge. *CoRR*, abs/2102.03315, 2021. URL https://arxiv.org/abs/2102.03315.

Alan Chan, Ben Bucknall, Herbie Bradley, and David Krueger. Hazards from increasingly accessible fine-tuning of downloadable foundation models. *CoRR*, abs/2312.14751, 2023. doi: 10.48550/ARXIV.2312.14751. URL https://doi.org/10.48550/arXiv.2312.14751.

Richard Fang, Rohan Bindu, Akul Gupta, Qiusi Zhan, and Daniel Kang. LLM agents can autonomously hack websites. In *USENIX Security Symposium*, 2024.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The pile: An 800gb dataset of diverse text for language modeling. *CoRR*, abs/2101.00027, 2021. URL https://arxiv.org/abs/2101.00027.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. An overview of catastrophic AI risks. *CoRR*, abs/2306.12001, 2023. doi: 10.48550/ARXIV.2306.12001. URL https://doi.org/10.48550/arXiv.2306.12001.

Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, and Ling Liu. Harmful fine-tuning attacks and defenses for large language models: A survey. *CoRR*, abs/2409.18169, 2024. doi: 10.48550/ARXIV.2409.18169. URL https://doi.org/10.48550/arXiv.2409.18169.

Simon Lermen, Charlie Rogers-Smith, and Jeffrey Ladish. Lora fine-tuning efficiently undoes safety training in llama 2-chat 70b. *CoRR*, abs/2310.20624, 2023. doi: 10.48550/ARXIV.2310.20624. URL https://doi.org/10.48550/arXiv.2310.20624.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. CAMEL: communicative agents for "mind" exploration of large language model society. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/a3621ee907def47c1b952ade25c67698-Abstract-Conference.html.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew B. Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhrugu Bharathi, Ariel Herbert-Voss, Cort B. Breuer, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Lin, Adam A. Hunt, Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan, Ian Steneker, David Campbell, Brad Jokubaitis, Steven Basart, Stephen Fitz, Ponnurangam Kumaraguru, Kallol Krishna Karmakar, Uday Kiran Tupakula, Vijay Varadharajan, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexandr Wang, and Dan Hendrycks. The WMDP benchmark: Measuring and reducing malicious use with unlearning. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=xlr6AUDuJz.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel J. Orr, Lucia Zheng, Mert Yüksekgönül, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *CoRR*, abs/2211.09110, 2022. doi: 10.48550/ARXIV.2211.09110. URL https://doi.org/10.48550/arXiv.2211.09110.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pp. 3214–3252. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.ACL-LONG.229. URL https://doi.org/10.18653/v1/2022.acl-long.229.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. TOFU: A task of fictitious unlearning for llms. *CoRR*, abs/2401.06121, 2024. doi: 10.48550/ARXIV.2401.06121. URL https://doi.org/10.48550/arXiv.2401.06121.

Kyle O'Brien, Stephen Casper, Quentin Anthony, Tomek Korbak, Robert Kirk, Xander Davies, Ishan Mishra, Geoffrey Irving, Yarin Gal, and Stella Biderman. Deep ignorance: Filtering pretraining data builds tamper-resistant safeguards into open-weight llms. *CoRR*, abs/2508.06601, 2025. doi: 10.48550/ARXIV.2508.06601. URL https://doi.org/10.48550/arXiv.2508.06601.

Domenic Rosati, Jan Wehner, Kai Williams, Lukasz Bartoszcze, David Atanasov, Robie Gonzales, Subhabrata Majumdar, Carsten Maple, Hassan Sajjad, and Frank Rudzicz. Representation noising effectively prevents harmful fine-tuning on llms. *CoRR*, abs/2405.14577, 2024. doi: 10.48550/ARXIV.2405.14577. URL https://doi.org/10.48550/arXiv.2405.14577.

Elizabeth Seger, Noemi Dreksler, Richard Moulange, Emily Dardaman, Jonas Schuett, Kevin Wei, Christoph Winter, Mackenzie Arnold, Seán Ó hÉigeartaigh, Anton Korinek, Markus Anderljung, Ben Bucknall, Alan Chan, Eoghan Stafford, Leonie Koessler, Aviv Ovadya, Ben Garfinkel, Emma Bluemke, Michael Aird, Patrick Levermore, Julian Hazell, and Abhishek Gupta. Open-sourcing highly capable foundation models: An evaluation of risks, benefits, and alternative methods for pursuing open-source objectives. *CoRR*, abs/2311.09227, 2023. doi: 10.48550/ARXIV.2311.09227. URL https://doi.org/10.48550/arXiv.2311.09227.

Abhay Sheshadri, Aidan Ewart, Phillip Guo, Aengus Lynch, Cindy Wu, Vivek Hebbar, Henry Sleight, Asa Cooper Stickland, Ethan Perez, Dylan Hadfield-Menell, and Stephen Casper. Latent adversarial training improves robustness to persistent harmful behaviors in llms. *Trans. Mach. Learn. Res.*, 2025, 2025. URL https://openreview.net/forum?id=6LxMeRlkWl.

Mohamed R. Shoaib, Zefan Wang, Milad Taleby Ahvanooey, and Jun Zhao. Deepfakes, misinformation, and disinformation in the era of frontier ai, generative ai, and large AI models. In *International Conference on Computer and Applications, ICCA 2023, Cairo, Egypt, November 28-30, 2023*, pp. 1–7. IEEE, 2023. doi: 10.1109/ICCA59364.2023.10401723. URL https://doi.org/10.1109/ICCA59364.2023.10401723.

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, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=4FIjRodbW6.

Gemma Team. Gemma: Open models based on gemini research and technology. *CoRR*, abs/2403.08295, 2024a. doi: 10.48550/ARXIV.2403.08295. URL https://doi.org/10.48550/arXiv.2403.08295.

Llama Team. The llama 3 herd of models. *CoRR*, abs/2407.21783, 2024b. doi: 10.48550/ARXIV.2407.21783. URL https://doi.org/10.48550/arXiv.2407.21783.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023. doi: 10.48550/ARXIV.2302.13971. URL https://doi.org/10.48550/arXiv.2302.13971.

Dmitrii Volkov. Badllama 3: removing safety finetuning from llama 3 in minutes. *CoRR*, abs/2407.01376, 2024. doi: 10.48550/ARXIV.2407.01376. URL https://doi.org/10.48550/arXiv.2407.01376.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jian Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *CoRR*, abs/2505.09388, 2025. doi: 10.48550/ARXIV.2505.09388. URL https://doi.org/10.48550/arXiv.2505.09388.

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, Alban Desmaison, Can Balioglu, Pritam Damania, Bernard Nguyen, Geeta Chauhan, Yuchen Hao, Ajit Mathews, and Shen Li. Pytorch FSDP: experiences on scaling fully sharded data parallel. *Proc. VLDB Endow.*, 16(12):3848–3860,

2023. doi: 10.14778/3611540.3611569. URL https://www.vldb.org/pvldb/vol16/p3848-huang.pdf.

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. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

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 Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/97ca7168c2c333df5ea61ece3b3276e1-Abstract-Conference.html.

A Attack Datasets

We evaluate robustness against several distinct fine-tuning attack strategies, each employing different data distributions to simulate adversarial modification scenarios. Our attack dataset selection follows established practices in the tamper resistance literature (O'Brien et al., 2025; Sheshadri et al., 2025; Rosati et al., 2024) and is largely adapted from (Tamirisa et al., 2025).

Pile-Bio. The Pile-Bio dataset consists of biosecurity-relevant scientific documents extracted from The Pile (Gao et al., 2021), a large-scale diverse text corpus. This dataset contains technical biological knowledge that models may have been trained to withhold or unlearn during safety alignment (Li et al., 2024; Tamirisa et al., 2025). We use the filtered biosecurity subset from (Tamirisa et al., 2025).

CAMEL-Bio. CAMEL-Bio is a synthetically generated dataset of biology-focused conversational data created using the CAMEL framework (Li et al., 2023). This dataset provides instruction-tuning style examples covering biological topics, enabling attacks that combine capability recovery with instruction-following.

Out-of-Distribution (OOD) Forget. The OOD forget strategy evaluates whether unlearning defenses generalize to data from outside the model's original pretraining distribution. Specifically, we use the WMDP biosecurity Forget (Li et al., 2024)-set—a held-out dataset of hazardous biology content that was not part of the model's pretraining data—as our OOD forget evaluation set.

Retain-to-Forget Switch. This mixed strategy begins fine-tuning on benign retain data (general web text or instruction-following examples) before switching to harmful forget data mid-training (Tamirisa et al., 2025).

B Benchmark Configuration

TamperTest uses a YAML configuration file to specify all experimental parameters, allowing researchers to easily add new models, benchmarks, and attack strategies without modifying source code. Below is an example configuration structure:

```
models:
  - meta-llama/Llama-3.1-8B-Instruct
  - microsoft/Phi-3-mini-4k-instruct
```

```
  - GraySwanAI/Llama-3-8B-Instruct-RR
  - lapisrocks/Llama-3-8B-Instruct-TAR-Bio-v2

benchmarks:
  - wmdp-bio
  - mmlu
  - truthfulqa
  - arc

attack_strategies:
  - pure_pile_bio_forget
  - pure_camel_bio_forget
  - ood_forget
  - retain_to_forget_switch

learning_rate: 2.0e-05
batch_size: 8
evaluation_intervals: [0, 100, 100, 100, 100, 100]

output_dir: benchmark_results
max_samples: 100  # Per-benchmark sample limit

tokenizer_map:
  lapisrocks/Llama-3-8B-Instruct-TAR-Bio-v2:
    meta-llama/Meta-Llama-3-8B-Instruct
  GraySwanAI/Llama-3-8B-Instruct-RR:
    meta-llama/Meta-Llama-3-8B-Instruct
```

Configuration Parameters.

- **models**: List of HuggingFace model identifiers to evaluate. Supports any decoder-only architecture compatible with the Transformers library.
- **benchmarks**: Evaluation benchmarks to run at each interval. Currently supports WMDP-bio, MMLU, TruthfulQA and ARC.
- **attack_strategies**: Fine-tuning attack types. Built-in strategies include:
  - `pure_pile_bio_forget`: Fine-tune on Pile Bio data only
  - `pure_camel_bio_forget`: Fine-tune on CAMEL Bio data only
  - `ood_forget`: Fine-tune on the WMDP biosecurity "forget-set," targeting weaponized or hazardous biological information rather than general biology
  - `retain_to_forget_switch`: Mixed strategy switching from retain to forget data

  Custom attack datasets can be defined in `attack_datasets.yaml` (see below).
- **evaluation_intervals**: List of step intervals for periodic evaluation. The first entry (0) represents baseline evaluation; the example configuration evaluates at steps 0, 100, 200, 300, 400, and 500.
- **learning_rate, batch_size**: Standard fine-tuning hyperparameters.
- **max_samples**: Optional limit on evaluation samples per benchmark (useful for faster prototyping).
- **tokenizer_map**: Maps models without public tokenizers to compatible tokenizer sources.

Custom Attack Datasets. Users can define new attack strategies by creating entries in `attack_datasets.yaml`:

```
my_custom_attack:
  dataset: "huggingface/dataset-name"
  split: "train"
```

```
  text_column: "text"  # or chat_column for chat format
  max_samples: 10000
  description: "Custom attack dataset description"
```

This extensible design allows researchers to benchmark new defense methods against arbitrary fine-tuning distributions without modifying the core framework.

C Safety and Capability Degradation Trajectories

Trajectory-based evaluation reveals non-monotone degradation patterns across all evaluated models. This temporal variability underscores why endpoint-only evaluations provide incomplete signals. Figures 3–13 show safety and capability trajectories for all evaluated models. Models are presented in descending order of TRI.

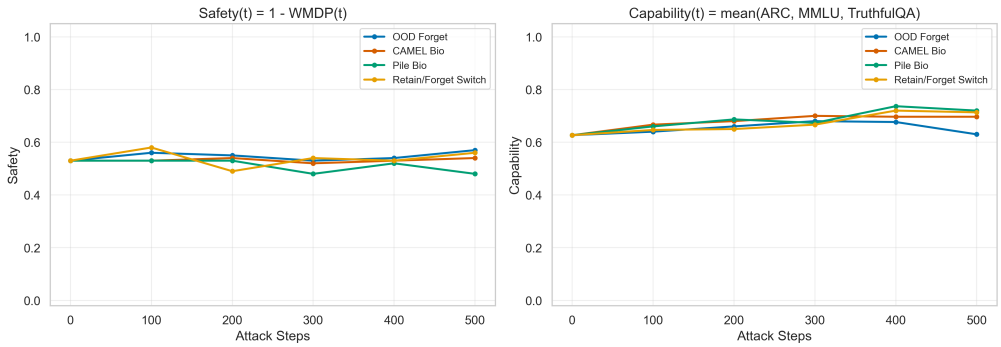

Figure 3: Safety and capability degradation curves for Microsoft Phi-3-Mini-4K-Instruct.

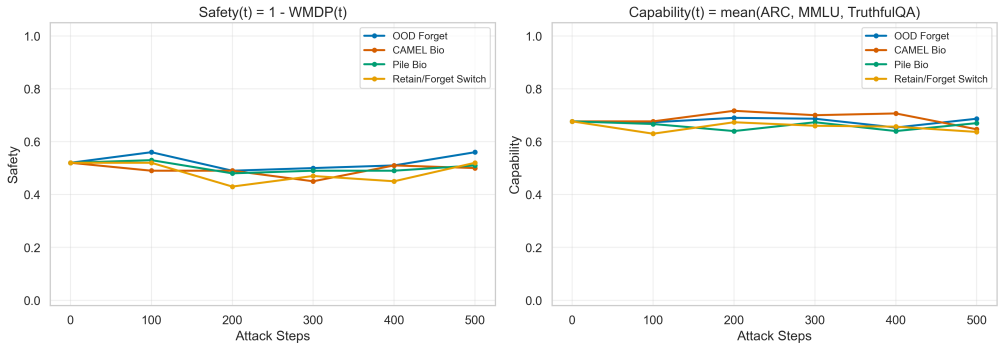

Figure 4: Safety and capability degradation curves for Qwen3 1.7B.

D Tamper Resistance Factors by Model

Figures 14–23 show the decomposition of tamper resistance scores into safety and capabilities for each evaluated model. Each plot displays three curves: $\mathsf{Safety}(t)$ and $\mathsf{Capability}(t)$ (averaged across all strategies), as well as the tamper resistance score $\mathsf{TRS}(t) = \mathsf{Safety}(t) \cdot \mathsf{Capability}(t)$. The shaded area under the $\mathsf{TRS}(t)$ curve represents the TRI. Models are presented in descending order of TRI.

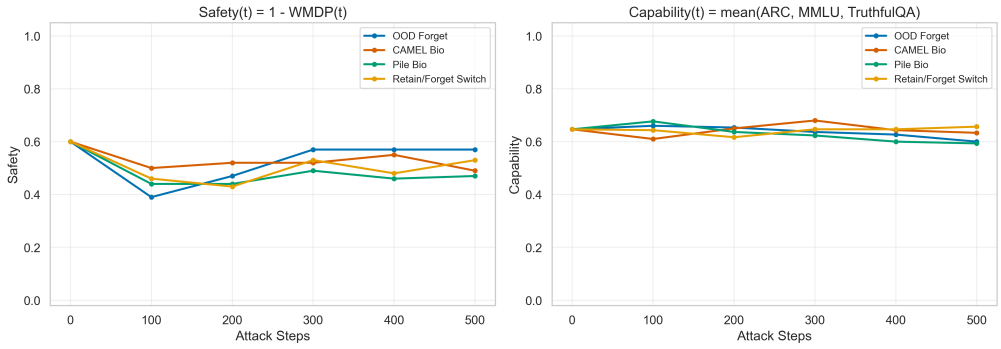

Figure 5: Safety and capability degradation curves for Llama 3 8B Instruct with Latent Adversarial Training (LAT).

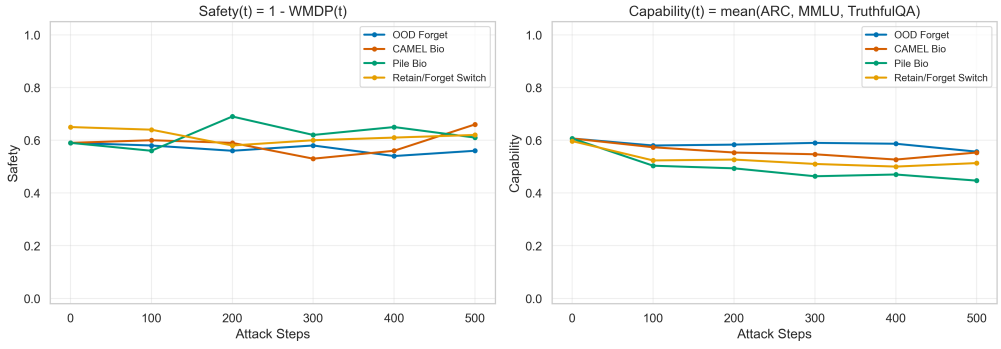

Figure 6: Safety and capability degradation curves for Llama 3.2 3B Instruct.

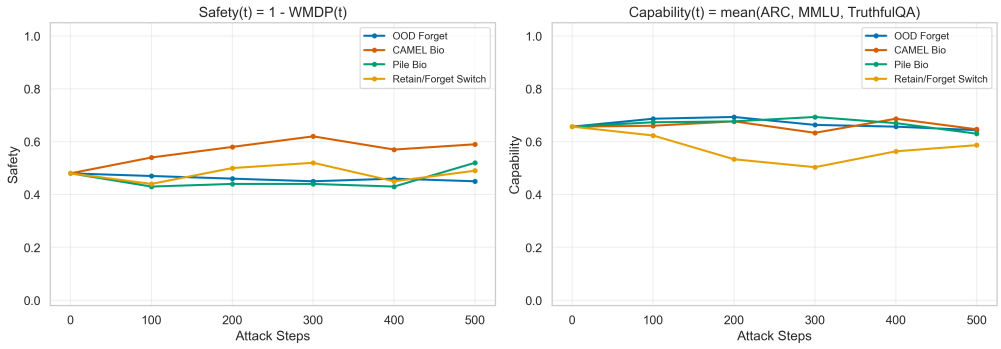

Figure 7: Safety and capability degradation curves for Llama 3.1 8B Instruct.

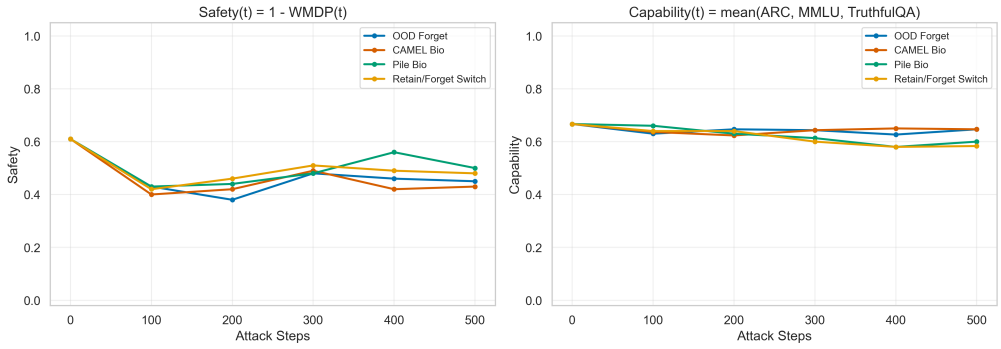

Figure 8: Safety and capability degradation curves for Llama 3 8B Instruct with Representation Rerouting (RR).

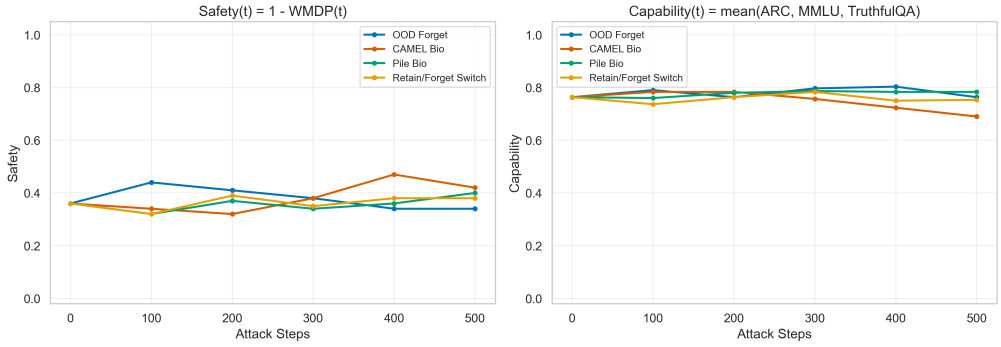

Figure 9: Safety and capability degradation curves for Qwen3 4B.

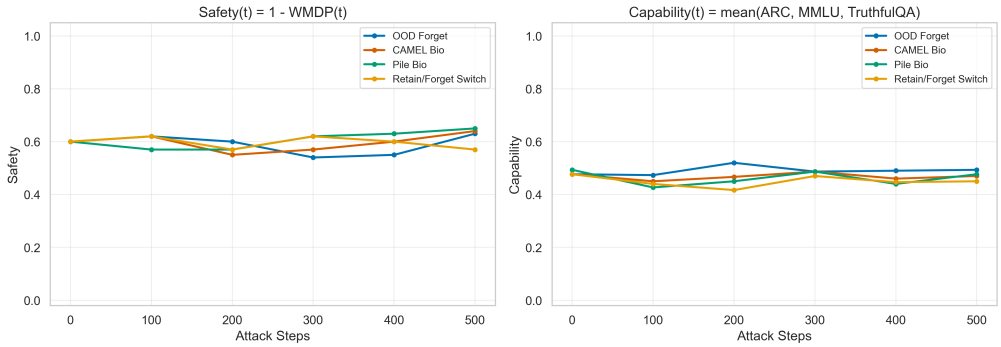

Figure 10: Safety and capability degradation curves for Qwen3 0.6B.

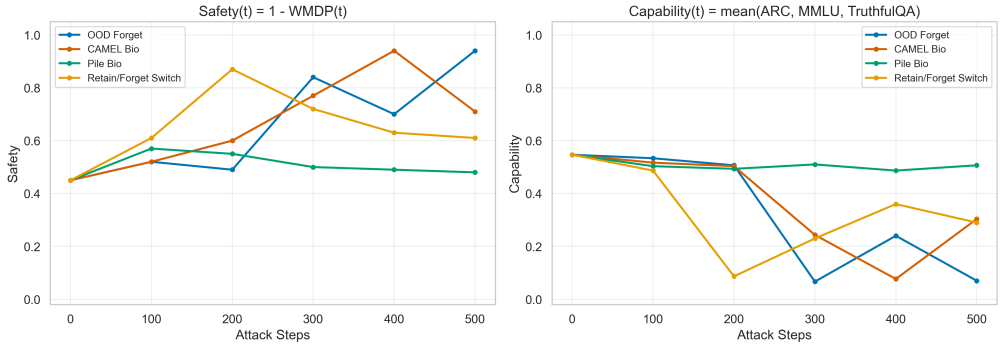

Figure 11: Safety and capability degradation curves for Gemma 2 2b it.

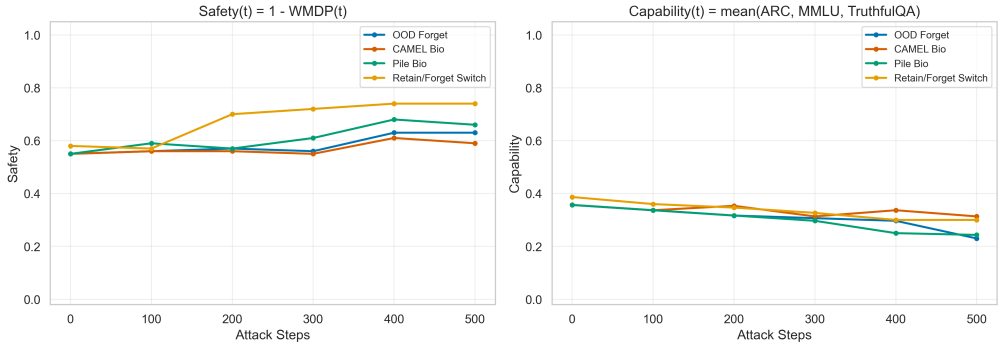

Figure 12: Safety and capability degradation curves for Llama 3.2 1B Instruct.

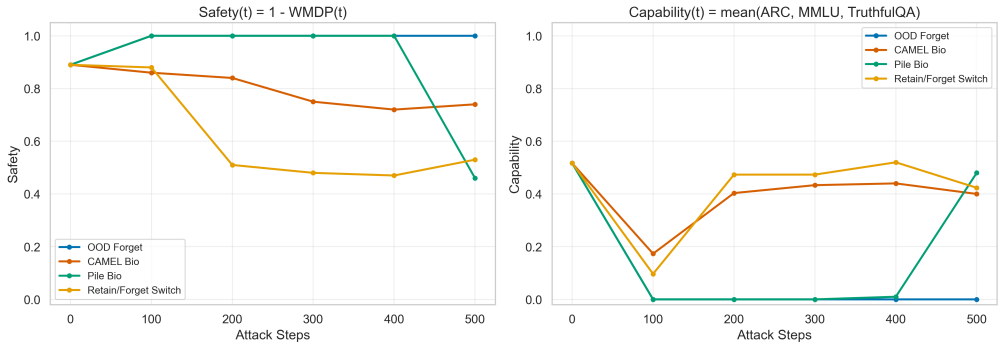

Figure 13: Safety and capability degradation curves for Llama 3 8B Instruct with TAR Defense.

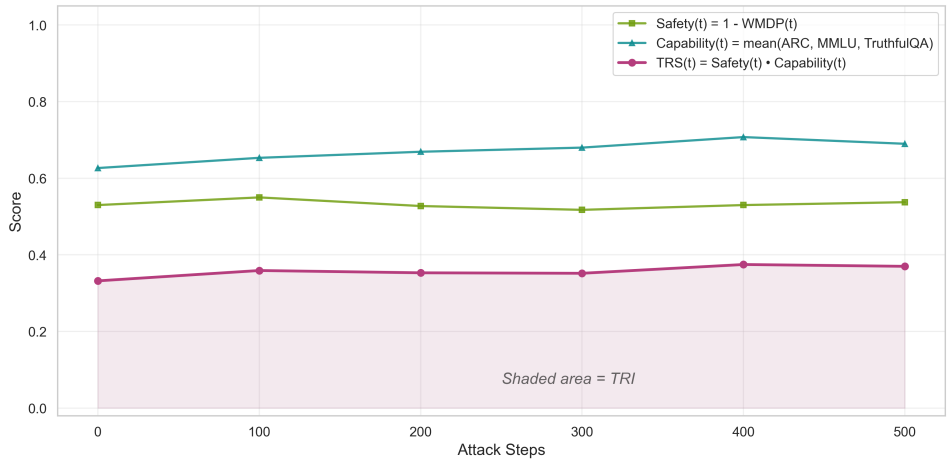

Figure 14: Safety, capability and tamper resistance scores for Microsoft Phi-3-Mini-4K-Instruct.

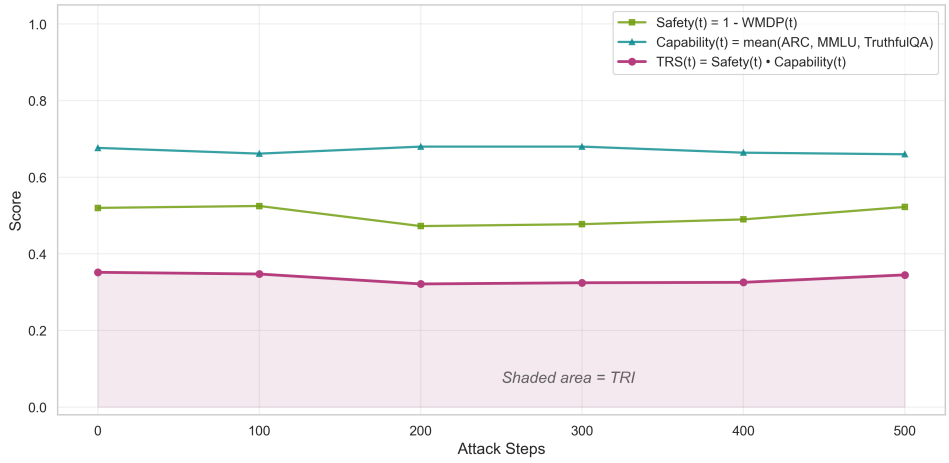

Figure 15: Safety, capability and tamper resistance scores for Qwen3 1.7B.

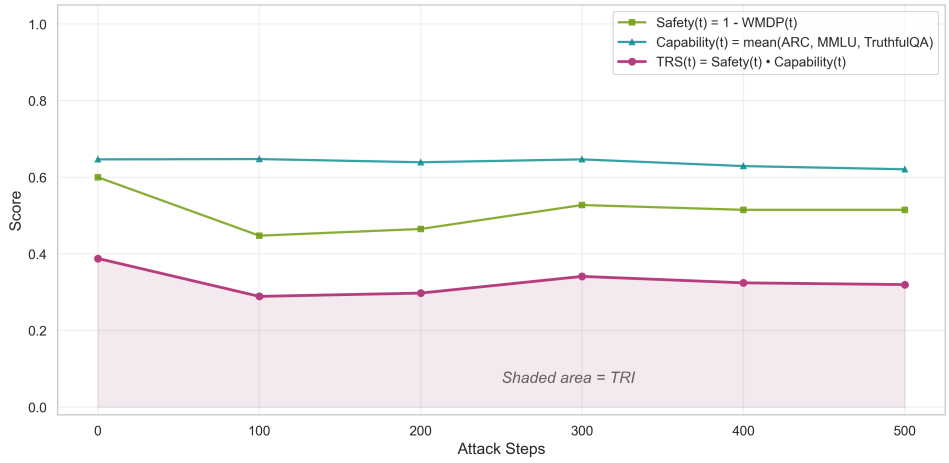

Figure 16: Safety, capability and tamper resistance scores for Llama 3 8B Instruct with Latent Adversarial Training (LAT).

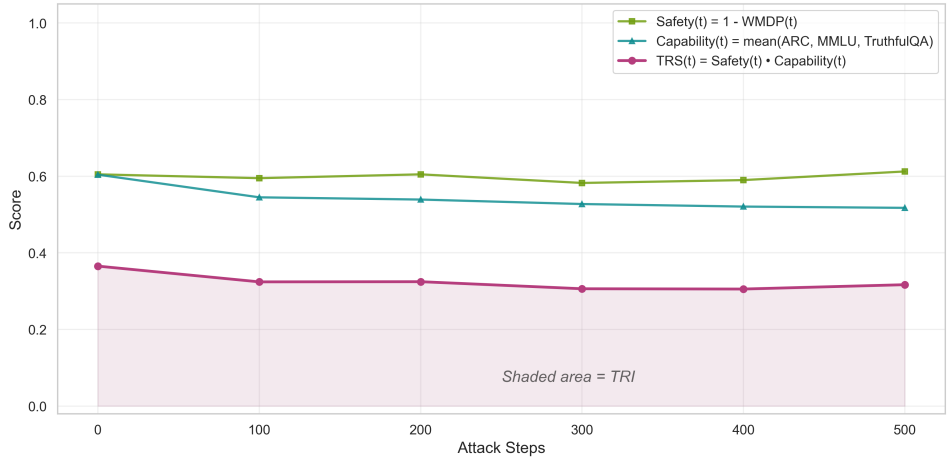

Figure 17: Safety, capability and tamper resistance scores for Llama 3.2 3B Instruct.

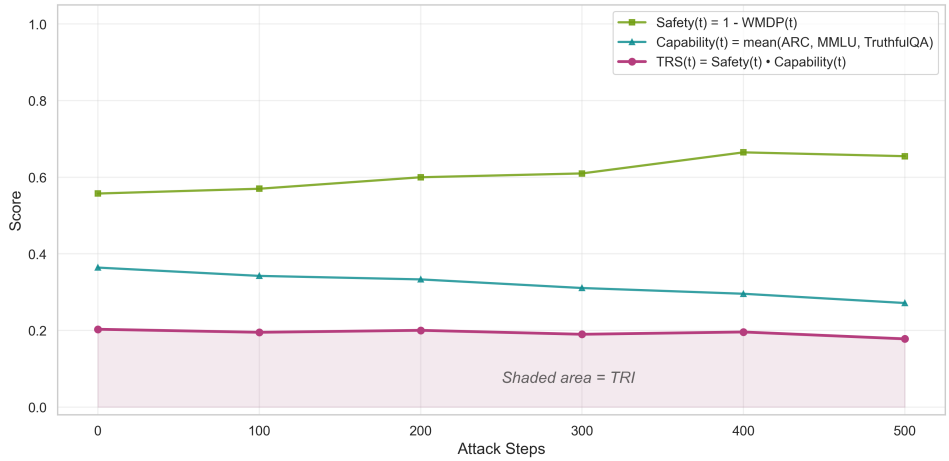

Figure 18: Safety, capability and tamper resistance scores for Llama 3.1 8B Instruct.

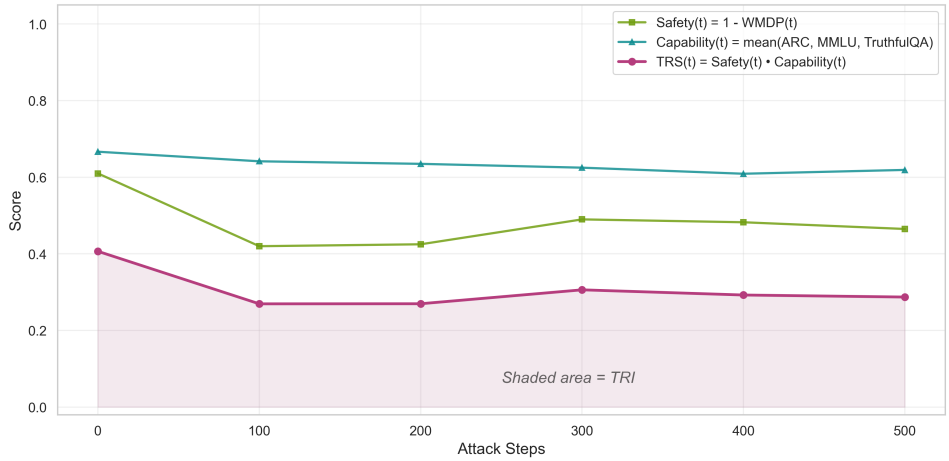

Figure 19: Safety, capability and tamper resistance scores for Llama 3 8B Instruct with Representation Rerouting (RR).

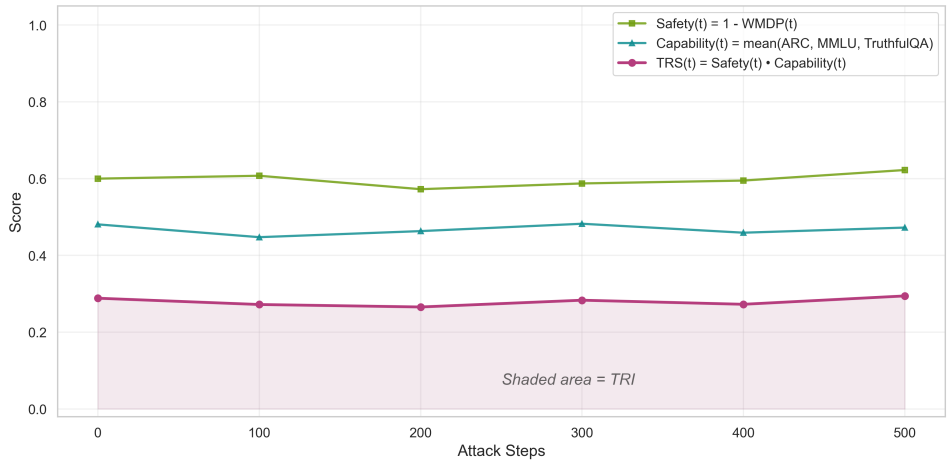

Figure 20: Safety, capability and tamper resistance scores for Qwen3 0.6B.

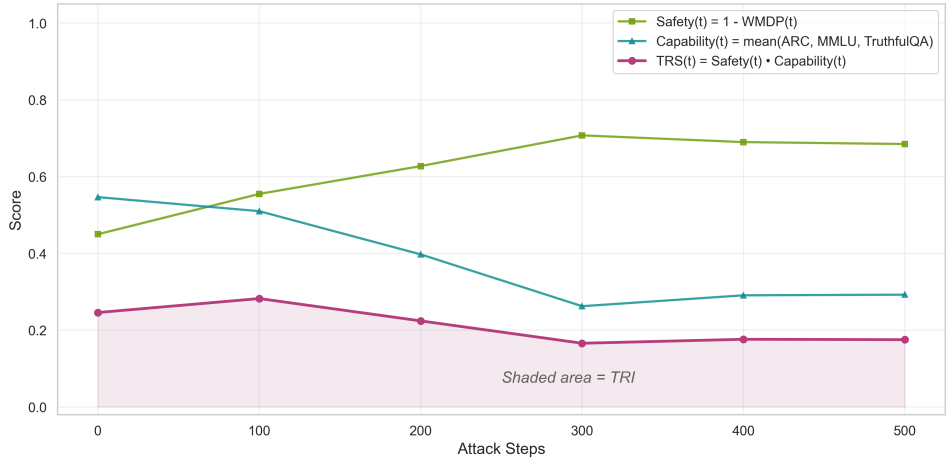

Figure 21: Safety, capability and tamper resistance scores for Gemma 2 2b it.

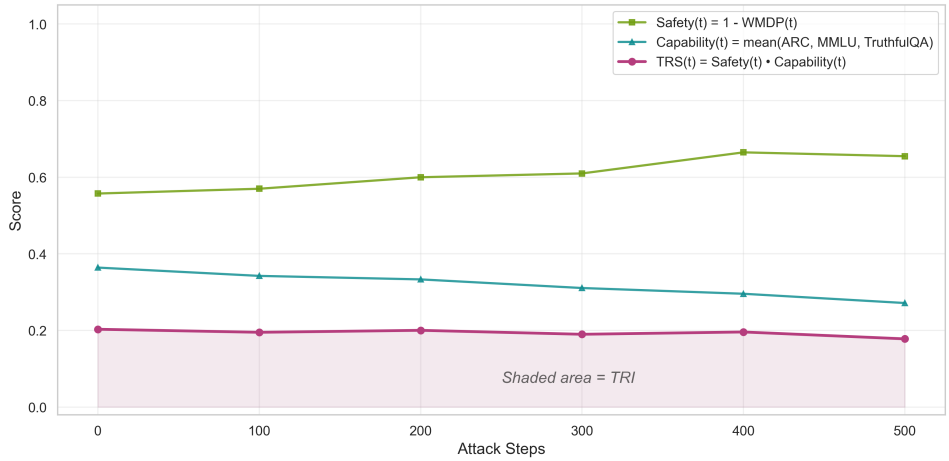

Figure 22: Safety, capability and tamper resistance scores for Llama 3.2 1B.

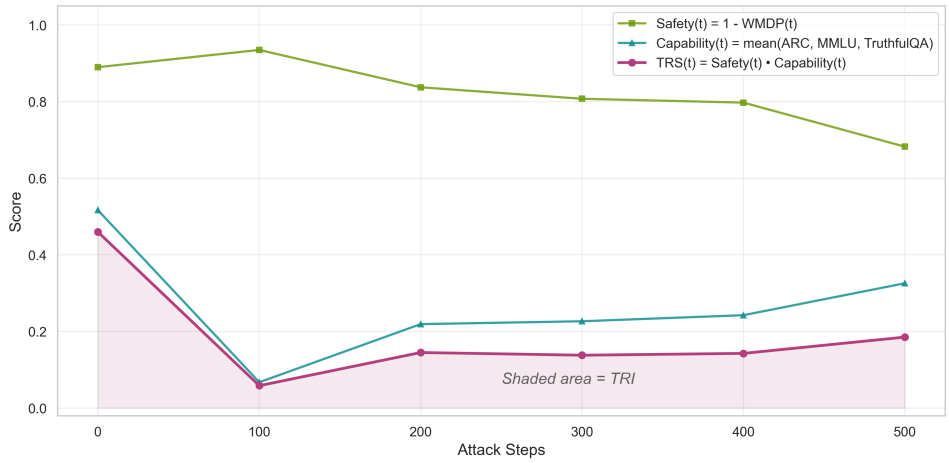

Figure 23: Safety, capability and tamper resistance scores for Llama 3 8B Instruct with TAR Defense.

E ROBUSTNESS OF TRI TO ATTACK CONFIGURATION

We conduct ablation studies varying both attack strategy and batch size, finding that TRI rankings remain remarkably stable across these variations, with model performance differences far exceeding configuration-induced variance.

E.1 ATTACK STRATEGY VARIATION

Figure 24 shows TRI breakdown across four distinct attack strategies. Model rankings exhibit strong stability across attack types: top-performing models (Phi-3-mini, Qwen3 1.7B, llama with LAT) consistently outperform across all attacks.

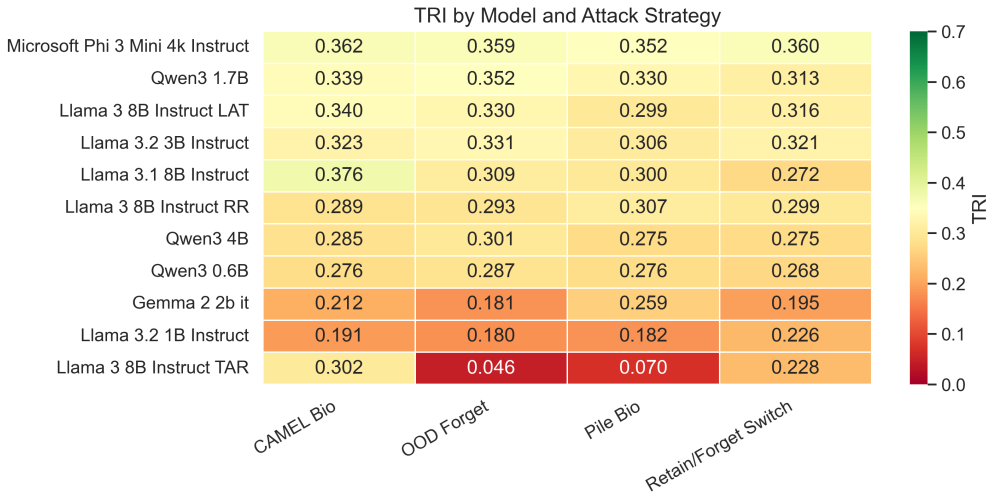

Figure 24: TRI by model and attack strategy. Each group of bars represents one model, with bars colored by attack type.

E.2 BATCH SIZE VARIATION

To test sensitivity to fine-tuning hyperparameters, we compare TRIs when fine-tuning with batch sizes 4 and 8 on Llama 3.1 8B Instruct. Figure 25 shows safety and capability trajectories for both configurations, averaged across all attack strategies.

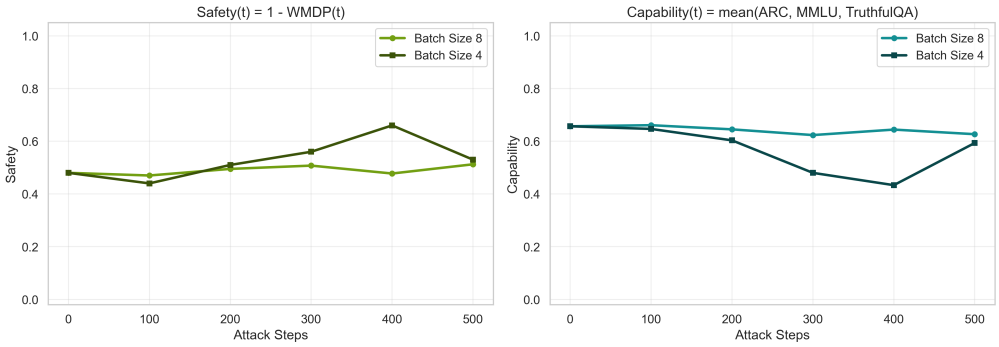

Figure 25: Comparison of Safety(t) and Capability(t) for Llama 3.1 8B under batch sizes 4 and 8. Trajectories are nearly identical, with TRI differing by less than 0.02.

Both configurations produce nearly identical degradation profiles. Safety and capability trajectories track closely throughout the attack, with only minor deviations at intermediate steps.

The resulting TRI values (0.314 for batch size 8 and 0.292 for batch size 4) differ by 0.022 – substantially less than the inter-model variance observed in Table 1, where TRI ranges from 0.145 to 0.358. This stability indicates that comparative rankings are robust to batch size selection within reasonable ranges.