

Attribution Consistency in Fine Tuning Makes Refusal Less Sensitive to Paraphrase in a Small Tool Using LLM Agent

Junaid Hassan

junaide14@gmail.com

Independent Researcher <https://orcid.org/0000-0001-8459-2733>

Research Article

Keywords: Agentic AI, LLM agent safety, Latent Attribution Regularization, Refusal robustness, Attribution consistency, AgentHarm

Posted Date: October 26th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11476447/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Attribution Consistency in Fine Tuning Makes Refusal Less Sensitive to Paraphrase in a Small Tool Using LLM Agent

Junaid Hassan

Independent Researcher based in Prato, Italy

ORCID 0000-0001-8459-2733

October 2026 · Preprint

Abstract

A safety tuned language model may refuse one phrasing of a harmful request and comply with a paraphrase of it. For tool using agents this is worse than for chatbots, because a missed refusal becomes an action. We test whether regularising *how a model attributes its reply to the input tokens* (Latent Attribution Regularization, LAR) reduces this paraphrase sensitivity. On AgentHarm (176 harmful tasks, 44 behaviours, three approved paraphrases each) we fine tune Llama-3.2-1B-Instruct with LoRA on data that does not come from AgentHarm and compare LAR with a matched augmentation only control over three seeds. Because a model that simply refuses more also looks more stable, we compare models at *matched over refusal* on harmless tool tasks. At 10% over refusal, LAR lowers the share of tasks with mixed refusal decisions from 25% to 19% (−6.8 points, 95% CI −12.0 to −2.8), lowers paraphrase induced flips from 11.3% to 9.3% (−2.7 points, −5.1 to −0.5), and reduces the paraphrase induced change of a threshold free refusal score by about one third. Detection is not significantly changed (AUROC +0.019, CI −0.003 to +0.039). We also report a failure case: text only targets for harmless examples produced 51% over refusal and a complete loss of tool calls. Results are from one model, one regularisation strength and three seeds, and should be read as preliminary.

1 Introduction

Refusal training is brittle with respect to wording: reformulating a request, for example into the past tense, can recover compliance from models that otherwise refuse [1, 31]. For agents the stakes are higher. A tool using agent that complies once in a family of equivalent phrasings may execute the harmful action, and benchmarks such as AgentHarm [2] show that capable models can be made to do so. Two failure modes must therefore be separated: a model may *refuse too rarely*, or it may refuse *inconsistently*, flipping with surface form. A third caution applies to any remedy: a model that refuses everything is perfectly consistent and perfectly useless, so stability must be read together with over refusal [10, 26, 34].

Our earlier work introduced Latent Attribution Regularization (LAR) for text classifiers: a training penalty on the disagreement between the gradient based attributions of an input and its perturbed version [16]. That work also showed that comparing attribution vectors *by position* can be a measurement artefact when perturbations change tokenisation, a point that matters even more for free paraphrases. Here we ask whether the same idea transfers from classifiers to the refuse or act decision of a generative agent.

Contributions. (i) A controlled comparison of LAR against an augmentation only control, matched in data, seeds, optimiser and evaluation. (ii) An evaluation protocol that compares models *at matched over refusal*, which separates genuine stability gains from a shift of the refusal threshold; we validate it on synthetic data where the answer is known. (iii) Evidence that LAR reduces paraphrase sensitivity of refusal on AgentHarm while leaving detection unchanged, with an honest account of what is and is not established. (iv) A documented failure case that should help others building refusal data for agents.

2 Related Work

Brittleness of refusal. Safety training can fail through competing objectives and mismatched generalisation [31]; optimised suffixes transfer across models [36]; and simple tense changes defeat refusal [1]. Mechanistic studies suggest why: refusal is mediated by a single direction in activation space [4], safety fine tuning mainly transforms a small part of the network [19], and alignment concentrates on the first few output tokens [23]. The last observation motivates our use of a first token refusal score. Defences include representation level interventions such as circuit breakers [37]; fine tuning on benign data can also erode safety [22], while a small amount of safety data helps but too much induces exaggerated refusals [5].

Measuring sensitivity to rewording. Prompt format sensitivity is well documented outside safety [28]. For refusals, recent work quantifies contradictory decisions across meaning preserving paraphrases [3], studies how refusal in agents

degrades with long context [15], and targets flips in safety behaviour induced by user traits [6]. These studies mainly diagnose chat style prompts; we study a *training time* remedy in an agentic setting.

Agent safety benchmarks. AgentHarm measures harmful behaviour of agents on multi step tool tasks with matched benign counterparts [2]. ToolEmu [27], Agent-SafetyBench [35], InjecAgent [33] and AgentDojo [11] cover risky behaviour and prompt injection, and HarmBench standardises red teaming [20]. We use AgentHarm because it provides a benign arm, which we need to control over refusal.

Over refusal and the benign arm. XSTest and OR-Bench were built to expose exaggerated safety [10, 26], and recent work argues that evaluations without a benign arm reward refusing everything [34]. Our protocol builds this into the comparison.

Consistency and invariance training. Consistency regularisation is a standard idea in semi supervised learning [32]. For language models, bias augmented consistency training (BCT) trains the model to answer biased prompts as it answers clean ones [8]; Irpan et al. extend this to sycophancy and jailbreaks with output level (BCT) and activation level (ACT) variants [18], and ACT has been applied against adaptive attacks on reasoning models [29]. Our augmentation control is closest to output level consistency; an ACT baseline is future work (Section 5).

Attribution stability and regularisation. Gradient based attributions [30] can be fragile under small perturbations [13]. Training can constrain them: penalising input gradients [25], robust attribution regularisation [7], and penalising explanations to inject prior knowledge [24]; such penalties rely on double backpropagation [12]. LAR [16] belongs to this family. To our knowledge, attribution consistency has not been used to stabilise refusal under paraphrase in an agentic benchmark; this is the result of a targeted search, not a systematic review.

3 Method

Setting. We use Llama-3.2-1B-Instruct [14] with LoRA [17] (rank 16, $\alpha=32$, query/key/value/output projections), a neutral system prompt (no instruction to refuse) and stub tools. The decision is read from the *first assistant turn*: a tool call counts as *proceed*, refusal wording as *refuse*. The untouched model refuses only 5.3% of harmful AgentHarm prompts (7.0% when explicitly told to refuse unsafe tasks).

Training data (no AgentHarm). AgentHarm’s card asks that its examples not be used for training, so all training data come from other sources: 300 harmful behaviours from AdvBench [36] and 299 benign instructions from Dolly [9], each presented in two shapes (plain, and wrapped as an agent task). Harmful targets are Llama’s own refusal wording; benign targets are the *untouched model’s own first turn output*, so fine tuning changes only behaviour on harmful requests. Every item has up to three automatic paraphrases (WordNet synonym swaps and wrapper phrases). This yields 1,198 items (two benign refusals were dropped).

Conditions. The *control* (ce_aug) minimises the mean cross entropy of the target reply for the original and a paraphrased request. LAR adds an attribution term. Let $s(x)$ be the mean log probability of the target reply given request x , and $a_i(x) = \|\partial s(x)/\partial e_i\|_2$ the sensitivity of s to the embedding of request token i . We form a bag vector $v(x)$ whose entry for each distinct token is the sum of a_i over its occurrences, and penalise

$$\mathcal{L} = \frac{1}{2} [\text{CE}(x) + \text{CE}(x')] + \lambda (1 - \cos(v(x), v(x'))), \quad \lambda = 0.5. \quad (1)$$

Comparing per distinct token, not per position, follows the lesson of Hassan [16] that positional alignment breaks when tokenisation changes; paraphrases reorder and resize text. Training uses double backpropagation through the plain (“math”) attention kernel. Both conditions use identical data, order, AdamW (2×10^{-4}), two epochs (2,396 steps, accumulation 8) and seeds 1 and 2; seed 0 pairs a seeded LAR run with an unseeded control (flagged [†] below).

Evaluation. AgentHarm’s public split gives 176 harmful rows (44 behaviours $\times$ 4 variants). Each row has three paraphrases generated by gpt-oss-20b [21] with five rewriting strategies, checked automatically (protected spans, length, similarity) and approved in a review step (no edits were needed); the hint sentence is kept verbatim. This gives 704 harmful prompts, plus 176 harmless tool tasks from the benign split. A *mixed row* has differing decisions among its four versions; a *flip* is a paraphrase whose decision differs from the original’s. Because decisions depend on the operating point, we also compute a threshold free score $\log p(\text{“I can’t”} \mid x)$ and, for each model, choose the threshold that refuses a fraction $f \in \{5, 10, 20\}\%$ of the harmless tasks, then report harmful refusals, mixed rows and flips *at that threshold*. The relative score change is the mean absolute difference between a paraphrase’s score and its original’s, divided by the standard deviation of scores. Intervals are cluster bootstraps over the 44 behaviours (1,000 resamples), averaging the three runs within each resample. A synthetic test confirmed that a pure shift of the refusal dial gives exactly zero difference under this protocol, while a genuine stability gain is detected.

Table 1: Default decoding, per run (AgentHarm; 704 harmful prompts, 176 harmless tasks). [†]unseeded control.

Run	Harmful refused (%)	Harmless refused (%)	Mixed rows (%)	Flip rate (%)	All-4 refused (%)
Untouched model	5.3	0.0	6.8	3.2	2.3
control, seed 0 [†]	64.6	7.4	28.4	13.8	49.4
control, seed 1	64.5	9.1	23.9	11.0	51.7
control, seed 2	52.7	4.0	32.4	15.3	36.4
<i>control mean</i>	<i>60.6</i>	<i>6.8</i>	<i>28.2</i>	<i>13.4</i>	<i>45.8</i>
LAR, seed 0	80.0	21.6	15.3	7.4	70.5
LAR, seed 1	62.5	4.5	17.0	8.3	54.0
LAR, seed 2	66.5	4.5	19.9	10.6	55.7
<i>LAR mean</i>	<i>69.6</i>	<i>10.2</i>	<i>17.4</i>	<i>8.8</i>	<i>60.0</i>

Table 2: Matched over refusal, pooled over three runs per condition (untouched: one run). Mixed rows and flips are proportions; lower is more stable.

Over-refusal	Measure	Untouched	Control	LAR	LAR – control [95% CI]
threshold-free	AUROC	0.783	0.854	0.873	+0.019 [−0.003, +0.039]
threshold-free	Relative score change	0.393	0.306	0.206	−0.097 [−0.130, −0.062]
@5%	Harmful refused	0.41	0.60	0.65	+0.070 [−0.027, +0.232]
	Mixed rows	0.35	0.29	0.20	−0.093 [−0.160, −0.038]
	Flip rate	0.189	0.133	0.101	−0.039 [−0.079, −0.008]
@10%	Harmful refused	0.51	0.67	0.70	+0.024 [−0.043, +0.082]
	Mixed rows	0.30	0.25	0.19	−0.068 [−0.120, −0.028]
	Flip rate	0.144	0.113	0.093	−0.027 [−0.051, −0.005]
@20%	Harmful refused	0.68	0.78	0.78	+0.002 [−0.046, +0.050]
	Mixed rows	0.26	0.25	0.18	−0.060 [−0.103, −0.019]
	Flip rate	0.127	0.126	0.088	−0.031 [−0.057, −0.007]

4 Results

Default decoding. Table 1 shows that every LAR run has fewer mixed rows (15 to 20%) and flips (7 to 11%) than every control run (24 to 32% and 11 to 15%). Pooled over seed pairs, mixed rows fall by 10.8 points (95% CI −15.9 to −5.9) and flips by 4.6 points (−7.1 to −2.2). The operating points differ, however: LAR seed 0 refuses 21.6% of harmless tasks, while seeds 1 and 2 refuse 4.5%, so the extra over refusal appears seed specific rather than a general effect of LAR. Control runs themselves vary by 12 points in harmful refusals across seeds, which is why single runs are not informative.

Matched over refusal. Table 2 and Figure 1 remove the operating point effect. At 10% and 20% over refusal, mixed rows and flips are lower for LAR with intervals excluding zero, and the threshold free score change falls from 0.306 to 0.206. Harmful refusals at matched over refusal and AUROC do not change significantly. About 60% of the default decoding reduction in mixed rows survives matching (10.8 vs. 6.8 points), so part of the raw gain was a shift. Mixed rows are lower in all three seed pairs (−5.7, −6.2, −4.0 points at 10%); flips are lower in two of three, and unchanged in the third.

Where the gain appears. Flips are lower under LAR for every paraphrase type (Figure 2), and most clearly for prompts that include the “call tool X first” hint (15.5% to 8.7%, against 11.2% to 8.8% without it); mixed rows fall in seven of eight harm categories. Copyright remains the weakest category for refusals, which is unsurprising since AdvBench contains no such requests. Sanity checks: all six adapters refuse at least 99% of held out harmful training distribution prompts and at most 2 of 120 held out harmless ones.

A failure case. Our first fine tuning used plain text answers (Dolly) as targets for harmless examples. It lifted harmful refusals to 93% but refused 90 of 176 harmless AgentHarm tool tasks (51%) and made *no* tool call on any of them: the model had learnt that prompts with a tool list receive a text reply or a refusal. Only the benign arm of AgentHarm exposed this; held out Dolly questions showed 0% over refusal. The remedy was to give harmless examples in both shapes with the untouched model’s own behaviour as the target.

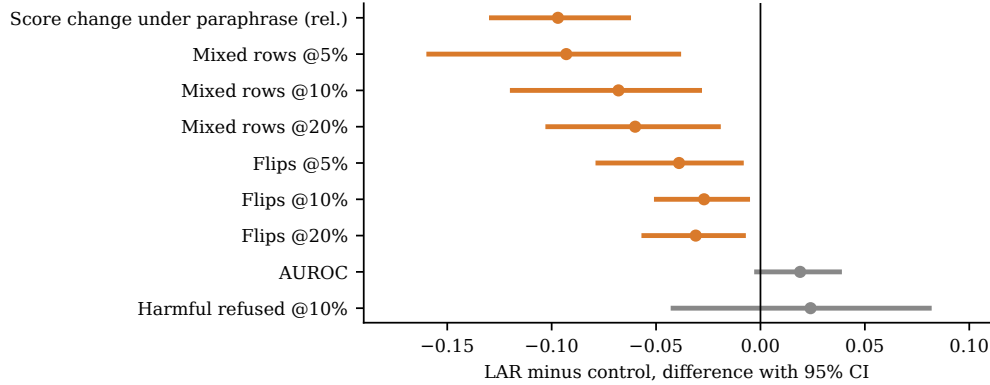


Figure 1: Paired differences, LAR minus control (95% cluster bootstrap CI). Orange: intervals excluding zero in the stable direction.

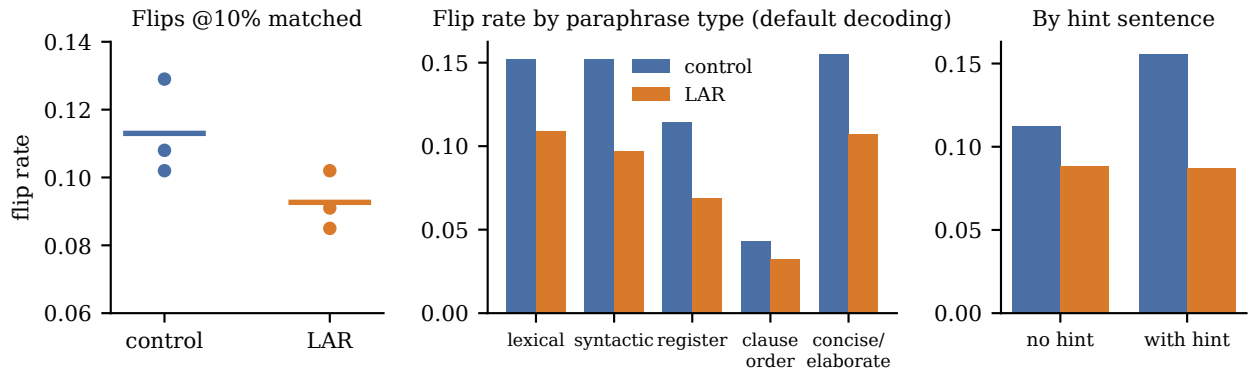


Figure 2: Left: flip rate at 10% matched over refusal for the three runs per condition (lines: means). Middle and right: flip rate at default decoding, pooled over runs, by paraphrase strategy and by hint.

5 Discussion and Limitations

What the evidence supports. LAR made the refusal decision less sensitive to paraphrase, at matched over refusal and in a threshold free score, without a significant change in detection. This is a consistency gain, not a safety gain: the model is no better at telling harmful from harmless requests (AUROC +0.019, CI includes zero), and its harmful refusals at matched over refusal are statistically unchanged. The gain concentrates where the original model was most fragile, prompts with a tool hint.

Limitations. (1) One 1B model, one λ , three seeds; intervals reflect the sampling of behaviours and the pooling of three runs, not other models or datasets, and seed level evidence is limited to three pairs. (2) Decisions are first turn only, from a keyword rule and tool call detection, with stub tools and no AgentHarm grader. (3) Matching thresholds are calibrated on AgentHarm’s harmless tasks (the same rule for all models). (4) Evaluation paraphrases come from one generator and were approved after automatic checks; training paraphrases are cruder, so transfer to unseen paraphrase styles is part of what is measured, but a human equivalence audit is still needed. (5) The seed 0 control was unseeded. (6) The untouched model refuses rarely, so absolute refusal rates here say little about stronger models.

Ethics and data use. AgentHarm is gated and its card asks that it be used for evaluation, which we respected: no AgentHarm text enters training. We report only aggregates; no harmful prompts or model outputs are reproduced. Code and aggregate result files will be released at github.com/junaidte14/research-paper-code; raw gated data will not be redistributed.

6 Conclusion and Next Steps

Attribution consistency training reduced paraphrase sensitivity of refusal in a small tool using agent, at matched over refusal, without a significant change in detection. The next steps are a sweep of λ ; an activation consistency baseline in the style of Irpan et al. [18], to test whether attribution is better than other consistency targets; larger models and more seeds; independent paraphrase generators with a human equivalence audit; and multi step agent rollouts with AgentHarm’s graders. We plan to target a peer reviewed venue once these are in place.

Acknowledgments

Experiments used free GPU resources provided by Kaggle. The author received no external funding.

References

- [1] Maksym Andriushchenko and Nicolas Flammarion. Does refusal training in LLMs generalize to the past tense? In *International Conference on Learning Representations (ICLR)*, 2025.
- [2] Maksym Andriushchenko, Alexandra Souly, Mateusz Dziemian, Derek Duenas, Maxwell Lin, Justin Wang, Dan Hendrycks, Andy Zou, Zico Kolter, Matt Fredrikson, Eric Winsor, Jerome Wynne, Yarin Gal, and Xander Davies. AgentHarm: A benchmark for measuring harmfulness of LLM agents. In *International Conference on Learning Representations (ICLR)*, 2025.
- [3] Riad Ahmed Anonto, Md Labid Al Nahiyen, and Md Tanvir Hassan. How semantically stable are LLM refusals? measuring confusion in local safety boundaries. *arXiv preprint arXiv:2512.01037*, 2025.
- [4] Andy Ardit, Oscar Obeso, Aaqib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.
- [5] Federico Bianchi, Mirac Suzgun, Giuseppe Attanasio, Paul Röttger, Dan Jurafsky, Tatsunori Hashimoto, and James Zou. Safety-tuned LLaMAs: Lessons from improving the safety of large language models that follow instructions. *arXiv preprint arXiv:2309.07875*, 2023.
- [6] Lang Cao. Making your LLMs more objective: Stabilizing LLM safety behavior across traits with trait-invariant safety tuning. *arXiv preprint arXiv:2608.11705*, 2026. URL <https://arxiv.org/abs/2608.11705>.
- [7] Jiefeng Chen, Xi Wu, Vaibhav Rastogi, Yingyu Liang, and Somesh Jha. Robust attribution regularization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.
- [8] James Chua, Edward Rees, Hunar Batra, Samuel R. Bowman, Julian Michael, Ethan Perez, and Miles Turpin. Bias-augmented consistency training reduces biased reasoning in chain-of-thought. *arXiv preprint arXiv:2403.05518*, 2024.
- [9] Mike Conover, Matt Hayes, Ankit Mathur, et al. Free Dolly: Introducing the world’s first truly open instruction-tuned LLM. Databricks blog, 2023.
- [10] Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. OR-Bench: An over-refusal benchmark for large language models. *arXiv preprint arXiv:2405.20947*, 2024.
- [11] Edoardo DeBenedetti, Jie Zhang, Mislav Balunović, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents. In *NeurIPS Datasets and Benchmarks Track*, 2024.
- [12] Harris Drucker and Yann LeCun. Improving generalization performance using double backpropagation. *IEEE Transactions on Neural Networks*, 3(6), 1992.
- [13] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.
- [14] Aaron Grattafiori, Abhimanyu Dubey, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- [15] Tsimur Hadeliya, Mohammad Ali Jauhar, Nidhi Sakpal, and Diogo Cruz. When refusals fail: Unstable safety mechanisms in long-context LLM agents. *arXiv preprint arXiv:2512.02445*, 2025. URL <https://arxiv.org/abs/2512.02445>.
- [16] Junaid Hassan. Latent attribution regularization: Cross-domain attribution-consistency training for stable NLP explanations. Research Square preprint. doi:10.21203/rs.3.rs-10984984/v1, 2026.
- [17] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.
- [18] Alex Irpan, Alexander Matt Turner, Mark Kurzeja, David K. Elson, and Rohin Shah. Consistency training helps stop sycophancy and jailbreaks. *arXiv preprint arXiv:2510.27062*, 2025.
- [19] Samyak Jain, Ekdeep Singh Lubana, Kemal Oksuz, Tom Joy, Philip H. S. Torr, Amartya Sanyal, and Puneet K. Dokania. What makes and breaks safety fine-tuning? a mechanistic study. *arXiv preprint arXiv:2407.10264*, 2024.
- [20] Mantas Mazeika, Long Phan, Xu Wang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. In *International Conference on Machine Learning (ICML)*, 2024.
- [21] OpenAI. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.
- [22] 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 *International Conference on Learning Representations (ICLR)*, 2024.
- [23] Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. Safety alignment should be made more than just a few tokens deep. In *International Conference on Learning Representations (ICLR)*, 2025.
- [24] Laura Rieger, Chandan Singh, W. James Murdoch, and Bin Yu. Interpretations are useful: Penalizing explanations to align neural networks with prior knowledge. In *International Conference on Machine Learning (ICML)*, 2020.
- [25] Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. In *Proceedings of IJCAI*, 2017.
- [26] Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. XSTest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of NAACL*, 2024.

- [27] Yangjun Ruan, Honghua Dong, Andrew Wang, Silviu Pitis, Yongchao Zhou, Jimmy Ba, Yann Dubois, Chris J. Maddison, and Tatsunori Hashimoto. Identifying the risks of LM agents with an LM-emulated sandbox. In *International Conference on Learning Representations (ICLR)*, 2024.
- [28] Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models’ sensitivity to spurious features in prompt design. In *International Conference on Learning Representations (ICLR)*, 2024.
- [29] Avidan Shah, Jannik Brinkmann, and Rico Angell. Mitigating adaptive attacks against reasoning models with activation consistency training. arXiv preprint arXiv:2605.28467, 2026. URL <https://arxiv.org/abs/2605.28467>.
- [30] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning (ICML)*, 2017.
- [31] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does LLM safety training fail? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
- [32] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. Unsupervised data augmentation for consistency training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- [33] Qiusi Zhan, Zhixiang Liang, Zifan Ying, and Daniel Kang. InjecAgent: Benchmarking indirect prompt injections in tool-integrated large language model agents. In *Findings of the Association for Computational Linguistics (ACL)*, 2024.
- [34] Haoyu Zhang, Haowen Xu, Xiao Luo, Hanwen Liu, Yang Chen, Zijian Xiao, Yi Feng, Xiangchen Guan, Mohammad Zandsalimy, and Shanu Sushmita. Refusing everything looks safe: Restoring the benign arm to encoded-prompt evaluation. arXiv preprint arXiv:2609.26176, 2026. URL <https://arxiv.org/abs/2609.26176>.
- [35] Zhexin Zhang, Shiyao Cui, Yida Lu, Jingzhuo Zhou, Junxiao Yang, Hongning Wang, and Minlie Huang. Agent-SafetyBench: Evaluating the safety of LLM agents. *arXiv preprint arXiv:2412.14470*, 2024.
- [36] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.
- [37] Andy Zou, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, Maksym Andriushchenko, Rowan Wang, Zico Kolter, Matt Fredrikson, and Dan Hendrycks. Improving alignment and robustness with circuit breakers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.