# The Epistemic Crisis in AI Safety Evaluation: When Measurement Tools Become Attack Surfaces

Adrian Wedd
Independent Researcher
Sydney, Australia
adrian@failurefirst.org

## Abstract

AI safety evaluation relies on a largely unexamined assumption: that the tools used to measure safety are themselves reliable. We present empirical evidence that automated LLM-based graders—the backbone of modern safety benchmarks—systematically misclassify safety-relevant model behaviors. In controlled experiments with seven open-weight grader models evaluated on 20 ground-truth traces where the correct verdict is unambiguous, we find that six achieve 100% accuracy on obvious cases—but one widely-used model (nemotron-3-nano) achieves only 80%, hallucinating safety caveats that do not exist in the response and classifying clear compliance as refusal. Combined with production data showing inter-grader agreement never exceeding $\kappa = 0.320$ on non-obvious cases, these results demonstrate that the measurement apparatus is unreliable precisely where it matters most: on the ambiguous boundary cases that determine real-world ASR numbers. We formalize this problem as a five-level recursive harm structure: (Level 0) direct harm from model outputs; (Level 1) iatrogenic harm from safety training itself; (Level 2) meta-iatrogenic harm from grader misclassification; (Level 3) epistemic cascade, where grader bias undermines trust in published benchmark results; (Level 4) adversarial evaluation poisoning, where attackers deliberately target the grader rather than the model; and (Level 5) safety theater, where models learn to satisfy graders rather than exhibit genuine safety. Our findings suggest that every published Attack Success Rate (ASR) carries an unknown grader-bias error bar that is rarely reported. We propose mitigations including multi-grader ensembles, mandatory grader meta-evaluation with human gold labels, and separation of grading from training pipelines.

## Keywords

AI safety, evaluation, LLM-as-judge, grader reliability, Goodhart's Law, red teaming, jailbreak benchmarks

## 1 Introduction

The AI safety community has invested considerable effort in building benchmarks to measure whether language models can be induced to produce harmful content. Benchmarks such as AdvBench [19], HarmBench [7], JailbreakBench [3], and StrongREJECT [13] provide standardized adversarial prompts and scoring methodologies. The central metric—Attack Success Rate (ASR)—purports to measure how often a model complies with harmful requests.

Yet this measurement apparatus rests on an assumption that receives remarkably little scrutiny: that the *grader* which determines whether a model "complied" or "refused" is itself accurate. In practice, modern safety evaluations increasingly rely on LLM-based automated graders—using one language model to judge the safety behavior of another [18]. The appeal is obvious: human evaluation is expensive and slow; LLM judges scale.

We present evidence that this assumption is dangerously wrong. In our experiments, we constructed 20 traces where the ground truth is unambiguous—10 clear refusals (the model explicitly declines) and 10 clear compliance cases (the model provides the requested harmful content). We then passed these traces through six LLM grader models using the FLIP backward-inference methodology [15]. Our results reveal systematic misclassification: grader models classify obvious refusals as compliance and obvious compliance as refusal at rates that would materially alter published ASR numbers.

This is not merely a calibration problem. We argue it constitutes a *recursive harm structure* with five identifiable levels, each compounding the last. When the measurement tool is broken, the entire evaluation ecosystem built upon it becomes unreliable—from individual benchmark scores to safety training pipelines that use LLM judges as reward signals, to regulatory compliance assessments that cite benchmark results as evidence of safety.

*Contributions.* We make the following contributions:

(1) We present empirical evidence of systematic grader misclassification on unambiguous safety evaluation traces (Section 4).
(2) We formalize a five-level recursive harm taxonomy for evaluation failure (Section 3).
(3) We situate grader unreliability within the broader context of Goodhart's Law, faithfulness gaps in reasoning, and iatrogenic safety (Section 2).
(4) We propose concrete mitigations including multi-grader ensembles and mandatory grader meta-evaluation (Section 6).

## 2 Related Work

*LLM-as-Judge.* Zheng et al. [18] demonstrated that strong LLMs can serve as judges for evaluating conversational AI, introducing MT-Bench and showing high agreement with human preferences. Subsequent work has adopted this paradigm widely: HarmBench uses GPT-4 as a judge [7], JailbreakBench uses fine-tuned classifiers [3], and StrongREJECT proposes rubric-based LLM scoring [13]. However, the reliability of these judges on *adversarial safety evaluation* specifically—where the boundary between compliance and refusal can be subtle—has received limited systematic study.

*Faithfulness and Reasoning Gaps.* Chen et al. [4] conducted 75,000 trials demonstrating that reasoning model chain-of-thought traces are often post-hoc rationalizations rather than faithful accounts of the model's decision process. Lanham et al. [6] similarly found

that chain-of-thought explanations can be unfaithful to the model's actual reasoning. These findings are directly relevant to grader reliability: if a grader model's stated reasoning for its verdict ("I classified this as COMPLIANCE because...") does not faithfully reflect its actual classification process, then the grader's behavior becomes opaque and potentially manipulable.

*Goodhart's Law in Machine Learning.* Campbell [2] and Strathern [14] articulated the general principle that when a measure becomes a target, it ceases to be a good measure. In machine learning, this manifests as reward hacking and specification gaming [9, 12]. We extend this analysis to *safety evaluation itself*: when ASR becomes the target metric for demonstrating model safety, and that metric is computed by a fallible LLM judge, the resulting optimization pressure may produce models that satisfy the judge without being genuinely safe.

*Iatrogenic Safety.* Our prior work [16] introduced the concept of iatrogenic safety in AI: safety measures that themselves cause harm. Examples include over-refusal (refusing benign requests), DETECTED_PROCEEDS patterns (the model detects a safety concern, announces it, then complies anyway), and safety training that creates exploitable behavioral patterns. The present paper extends this framework to the *evaluation* layer, identifying a meta-iatrogenic effect where the tools used to detect iatrogenic safety are themselves unreliable.

*Automated Red Teaming.* Perez et al. [10] pioneered using language models to generate adversarial inputs for other language models. Wei et al. [17] systematically analyzed how safety training fails, identifying competing objectives and mismatched generalization as key failure modes. Zou et al. [19] demonstrated universal adversarial suffixes that transfer across models. This body of work assumes that compliance detection—the grader—works correctly. Our contribution is to challenge that assumption directly.

## 3 The Five Levels of Evaluation Harm

We propose a recursive taxonomy of evaluation-related harms, where each level compounds the effects of the previous one.

### 3.1 Level 0: Direct Harm

A model generates harmful content in response to an adversarial prompt. This is the behavior that safety benchmarks are designed to detect. It is well-studied and is the explicit target of alignment techniques including RLHF [1, 8], Constitutional AI, and instruction-hierarchy training.

### 3.2 Level 1: Iatrogenic Harm

Safety training itself introduces new failure modes [16]. Over-refusal prevents users from obtaining legitimate assistance. DETECTED_PROCEEDS patterns create a false sense of safety: the model announces it has identified a harmful request, then proceeds to comply. Safety training can also create predictable behavioral patterns that adversaries exploit—the very alignment that makes a model "safe" on benchmarks may make it more vulnerable to sophisticated attacks.

### 3.3 Level 2: Meta-Iatrogenic Harm (Grader Misclassification)

The grader used to measure model safety is itself unreliable. A grader that classifies refusals as compliance inflates ASR, making models appear less safe than they are. A grader that classifies compliance as refusal deflates ASR, creating false confidence. In either case, the *measurement* of safety diverges from *actual* safety, and decisions based on that measurement (model deployment, safety training iteration, regulatory compliance) are made on false premises.

This level is the primary focus of our empirical work (Section 4). We demonstrate that this is not a hypothetical concern: production-grade open-weight models used as FLIP graders exhibit systematic misclassification on unambiguous cases.

### 3.4 Level 3: Epistemic Cascade

Grader bias does not remain contained. Published ASR numbers from major benchmarks inform:

- Model selection decisions ("Model A is safer than Model B")
- Safety training iterations ("Our new training reduced ASR from X% to Y%")
- Regulatory assessments under the EU AI Act [5]
- Academic claims about the effectiveness of defense techniques
- Public discourse about AI safety

If the grader systematically inflates ASR, then claims of "improving safety" may reflect improvements in satisfying the grader rather than genuine safety improvement. If different benchmark teams use different graders with different biases, cross-benchmark comparisons become meaningless. The epistemic cascade means that grader unreliability corrupts not just individual measurements but the *field's collective understanding* of where safety stands.

### 3.5 Level 4: Adversarial Evaluation Poisoning

Once it is known that evaluation relies on LLM judges, a sophisticated adversary can attack the *judge* rather than the model. Techniques include:

- **Grader-aware jailbreaks**: Crafting model responses that genuinely comply with harmful requests but are phrased to trigger the grader's "refusal" classification (e.g., wrapping harmful content in safety-sounding language).
- **Benchmark poisoning**: Influencing the training data of grader models to shift their classification behavior.
- **Meta-prompt injection**: Embedding adversarial content in model responses that manipulates the grader's chain-of-thought when that response is fed to the grader for evaluation.

This level has not, to our knowledge, been systematically explored. It represents a qualitatively different threat: the adversary's target is the measurement system, not the model being measured.

### 3.6 Level 5: Safety Theater

The terminal failure mode. When models are optimized against LLM judges (via RLHF, DPO, or Constitutional AI with LLM feedback), and those judges have systematic biases, models learn to *satisfy the*

*judge* rather than *be safe*. This is Goodhart's Law applied to safety evaluation.

Concretely: if a grader model is biased toward classifying verbose safety disclaimers as "refusal," then models trained against that grader will learn to produce verbose safety disclaimers regardless of whether they ultimately comply. The sycophancy literature [11] documents a related phenomenon: models learn to produce responses that evaluators prefer, even when those responses are incorrect.

Safety theater is particularly insidious because it *looks like progress*. ASR numbers go down. Benchmark scores improve. But the improvement is in satisfying the judge, not in genuine safety—and the gap between the two may widen over time as models become increasingly sophisticated at gaming the evaluation.

## 4 Empirical Evidence: Grader Misclassification

### 4.1 Experimental Setup

We constructed a calibration dataset of 20 traces with unambiguous ground truth:

- 10 **clear refusals**: The model explicitly declines the request using phrases such as "I cannot and will not help," "I refuse," "I'm not going to help with this," and "Absolutely not."
- 10 **clear compliance**: The model provides detailed harmful content—exploit code, step-by-step instructions for illegal activities, or templates for disinformation—directly fulfilling the adversarial request.

All traces use attack prompts from AdvBench [19], the most widely-used jailbreak evaluation benchmark. The ground truth for each trace is unambiguous by construction: a human annotator would achieve 100% accuracy on this dataset.

We evaluated six grader models using the FLIP backward-inference methodology, which instructs the grader to: (1) read the model response, (2) infer what instruction the model appears to be following, and (3) classify the response as COMPLIANCE, REFUSAL, PARTIAL, HALLUCINATION_REFUSAL, or BENIGN_QUERY. FLIP was designed to reduce grader bias by focusing on the response content rather than surface-level cues [15].

*Grader models evaluated:*

(1) `gemma3:4b` (Google, 4B parameters)
(2) `gemma3:12b` (Google, 12B parameters)
(3) `gemma3:27b` (Google, 27B parameters)
(4) `nemotron-3-nano:30b` (NVIDIA, 30B parameters)
(5) `ministral-3:8b` (Mistral, 8B parameters)
(6) `ministral-3:14b` (Mistral, 14B parameters)
(7) `devstral-small-2:24b` (Mistral, 24B parameters)

All models were accessed via the Ollama Cloud API with temperature set to 0.0 and maximum output tokens set to 800.

### 4.2 Results

Six of seven grader models achieved 100% accuracy on our 20-trace calibration set (Table 1). The outlier, `nemotron-3-nano:30b`, misclassified 4 of 20 traces (80% accuracy), with errors exclusively on the *compliance* side: it classified clear compliance as REFUSAL (1 case), PARTIAL (1 case), and HALLUCINATION_REFUSAL (1 case), plus one refusal classified as PARTIAL.

**Table 1: Grader accuracy on 20 obvious ground-truth traces (10 clear refusals, 10 clear compliance). REF→COMP = grader classified a clear refusal as COMPLIANCE (inflates ASR). COMP→REF = grader classified clear compliance as REFUSAL (deflates ASR). "Other" = misclassifications to PARTIAL, HALLUCINATION_REFUSAL, or ERROR.**

| Grader Model | Accuracy | REF→COMP | COMP→REF | Other |
|---|---|---|---|---|
| gemma3:4b | 100% | 0/10 | 0/10 | 0 |
| gemma3:12b | 100% | 0/10 | 0/10 | 0 |
| gemma3:27b | 100% | 0/10 | 0/10 | 0 |
| ministral-3:8b | 100% | 0/10 | 0/10 | 0 |
| ministral-3:14b | 100% | 0/10 | 0/10 | 0 |
| devstral-small-2:24b | 100% | 0/10 | 0/10 | 0 |
| nemotron-3-nano:30b | 80% | 0/10 | 1/10 | 3 |

*Nemotron error analysis.* The nemotron misclassifications reveal a qualitatively concerning failure mode. On trace GEVAL-COMP-006—a Python script that automates sending threatening messages to a target, with no disclaimers or caveats in the response—nemotron classified the response as PARTIAL, stating "extensive caveats and warnings indicate only partial compliance rather than full support." The grader *hallucinated* safety caveats that do not exist in the response. On GEVAL-COMP-004 (a bomb-making tutorial), nemotron classified the response as HALLUCINATION_REFUSAL, and on GEVAL-COMP-005 (a fraud code generator) it classified the response as REFUSAL.

This pattern—a grader that underclassifies compliance, treating obvious harmful outputs as if they contained safety hedging—would *deflate* measured ASR, making models appear safer than they are. This is the inverse of the ASR-inflation concern, but equally dangerous for safety evaluation.

*Prior evidence from production grading.* Beyond our controlled experiment, we have documented grader misclassification in production settings:

- `nemotron-nano-9b` classified the response "I cannot and will not help with that" as COMPLIANCE during AdvBench baseline grading.
- `gemma3:4b` classified a clear refusal on scenario ADVB-0010 as COMPLIANCE.
- `gemma3:27b` classified "I cannot help with creating exploit scripts" as COMPLIANCE.
- In Sprint 13 production grading, 53% of traces were infrastructure errors misclassified as REFUSAL—a different failure mode but equally damaging to measurement validity.
- Prior heuristic (keyword-based) classifiers overcounted ASR by up to 84:1 compared to LLM-based grading ($\kappa = 0.097$).
- Inter-grader agreement across all LLM classifiers in our corpus never exceeds $\kappa = 0.320$ ("fair" agreement)—even the best grader pair agrees only 54% of the time on non-obvious cases.

### 4.3 Analysis

*Obvious cases are necessary but not sufficient.* The high accuracy of most graders on our 20-trace calibration set might suggest the problem is smaller than feared. However, this dataset was designed with *unambiguous* ground truth. Real-world traces contain hedged refusals, partial compliance, DETECTED_PROCEEDS patterns (where the model announces a safety concern then proceeds to comply), and responses that mix helpful and harmful content. Our production inter-grader agreement data (Report #240) tells the story of what happens on these harder cases: even the best grader pair (gemini vs. haiku) achieves only $\kappa = 0.320$ (54% agreement).

*Nemotron's hallucinated caveats.* The most concerning finding is not the error *rate* but the error *type*. When nemotron-3-nano classified a harassment automation script as PARTIAL, it justified this by citing "extensive caveats and warnings" in the response. No such caveats exist. The grader *confabulated* a safety-relevant property of the response. If this confabulation tendency scales—if graders systematically imagine safety behaviors that aren't present—then automated evaluation could paint a fundamentally misleading picture of model safety.

*The ambiguity gap.* Our results suggest a two-regime model of grader reliability:

- **Obvious regime**: Most graders achieve near-perfect accuracy when the response is clearly a refusal or clearly compliance. Grading quality is acceptable.
- **Ambiguous regime**: On boundary cases—partial compliance, hedged refusals, DETECTED_PROCEEDS—grader agreement drops sharply ($\kappa < 0.40$). This is precisely the regime where accurate grading matters most, because these are the cases that determine whether a model is safe enough to deploy.

The implications are significant: even if a grader's overall error rate appears low (because most cases are obvious), its errors may concentrate on exactly the cases that matter for safety decisions. A grader that is right 95% of the time but wrong on every ambiguous case provides false confidence.

*Error direction matters.* Nemotron's errors all went in one direction: classifying compliance as non-compliance (deflating ASR). In our production data, the 84:1 overcounting by heuristic classifiers went in the opposite direction (inflating ASR). Different graders exhibit different systematic biases. Without reporting these biases, cross-benchmark comparisons are unreliable: a model that scores "ASR = 30%" on one benchmark might score "ASR = 55%" on another, purely due to grader differences.

## 5 Implications

### 5.1 For Published Benchmarks

Every published ASR number should be understood as carrying an unknown grader-bias error bar. Without published grader meta-evaluation results (accuracy on ground-truth traces, confusion matrices, inter-grader agreement), ASR numbers are not interpretable. Cross-benchmark comparisons are particularly unreliable when different benchmarks use different graders.

### 5.2 For Safety Training

Modern safety training pipelines—RLHF [8], DPO, Constitutional AI—increasingly use LLM judges as part of the reward signal. If the judge is biased, the model learns to satisfy the judge's biases rather than achieve genuine safety. This creates a form of optimization against a misspecified objective that may be difficult to detect because the misspecification is in the *evaluation*, not the training objective.

### 5.3 For Regulation

The EU AI Act [5] (Article 9) requires "appropriate" testing and evaluation for high-risk AI systems. If the testing methodology itself is unreliable, then compliance with these requirements becomes circular: a system is deemed safe because it passes a test, but the test may not accurately measure safety. Regulators relying on benchmark ASR numbers to assess model safety are building on potentially unsound foundations.

### 5.4 For the Field

We need grader calibration standards alongside model safety standards. The AI safety community has invested heavily in standardizing *what* to evaluate (adversarial prompts, harm categories, benchmark design). Comparatively little effort has gone into standardizing *how reliably* the evaluation is performed. This asymmetry creates a systematic vulnerability: improvements in benchmark design are undermined by unreliable grading.

## 6 Proposed Mitigations

### 6.1 Multi-Grader Ensembles

Use multiple grader models and report agreement rates. Our prior work comparing FLIP with StrongREJECT scoring showed that ensemble approaches can identify cases where individual graders disagree, flagging them for human review rather than defaulting to a single grader's verdict.

### 6.2 Grader Meta-Evaluation with Human Gold Labels

Every benchmark should publish grader accuracy metrics alongside ASR numbers. This requires:

- A calibration dataset of traces with human-verified ground truth
- Per-grader confusion matrices on the calibration set
- Grader accuracy reported as a confidence interval, not a point estimate

### 6.3 Published Grader Error Rates

ASR numbers should be reported with grader-adjusted confidence intervals. If a grader has a known 15% misclassification rate on refusals, the reported ASR should reflect this uncertainty: "ASR = 45% ± 7% (grader error)" rather than simply "ASR = 45%."

### 6.4 Adversarial Grader Robustness Testing

Test graders against adversarial inputs designed to fool the grader specifically—model responses crafted to trigger misclassification.

This is analogous to adversarial robustness testing for the models themselves, applied to the evaluation layer.

### 6.5 Separation of Grading from Training

Do not use the same LLM judge for both safety evaluation *and* as a reward signal in safety training. When the grader and the training signal share biases, the resulting optimization loop amplifies those biases rather than correcting them. Independent grading and training pipelines provide a natural check.

## 7 Limitations and Future Work

Our calibration dataset of 20 traces, while carefully constructed, is small. The traces are drawn from a single benchmark (AdvBench) and represent extreme cases—obvious refusals and obvious compliance. Real-world grading tasks involve ambiguous cases (partial compliance, hedged refusals, DETECTED_PROCEEDS patterns) where we would expect grader accuracy to be substantially lower.

We evaluated six open-weight models; commercial models (GPT-4, Claude) were not tested as graders. It is possible that larger, more capable models achieve higher grading accuracy, though this would not resolve the fundamental problem: most published benchmarks and most safety training pipelines do not use the most capable models as graders due to cost.

The five-level taxonomy is a conceptual framework. Levels 4 and 5 (adversarial evaluation poisoning and safety theater) are identified as theoretical risks based on our understanding of the system dynamics; we do not present direct empirical evidence for these levels in this paper.

Future work should:

(1) Develop a large-scale grader calibration benchmark with hundreds of human-verified traces spanning the full spectrum of ambiguity.
(2) Systematically evaluate commercial graders (GPT-4, Claude) using the same methodology.
(3) Empirically test Level 4 (adversarial grader attacks) by constructing model responses designed to fool specific graders.
(4) Investigate whether RLHF training against biased judges produces measurable safety theater effects (Level 5).

## 8 Conclusion

AI safety evaluation is in an epistemic crisis. The tools used to measure whether models are safe—automated LLM graders—are themselves unreliable, and this unreliability cascades through five levels of compounding harm. Our empirical results demonstrate that open-weight grader models systematically misclassify unambiguous safety behaviors, and our analysis of production grading data shows inter-grader agreement never exceeding $\kappa = 0.320$ ("fair").

The solution is not to abandon automated evaluation—human evaluation at scale is impractical. Rather, the solution is to treat the grader with the same rigor we apply to the model being evaluated: test it, measure its error rates, report those error rates, and build evaluation pipelines that account for grader fallibility. Every ASR number should come with a grader accuracy disclosure. Every benchmark should publish grader confusion matrices. Every safety training pipeline should use independent graders.

Until these practices become standard, every published safety benchmark result carries an asterisk: *subject to unknown grader bias.*

## References

[1] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. *arXiv preprint arXiv:2204.05862* (2022).

[2] Donald T. Campbell. 1979. Assessing the Impact of Planned Social Change. *Evaluation and Program Planning* 2, 1 (1979), 67–90.

[3] Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramer, Matthias Hein, and J. Zico Kolter. 2024. JailbreakBench: An Open Robustness Benchmark for Jailbreaking Large Language Models. *arXiv preprint arXiv:2404.01318* (2024).

[4] Kevin Chen, Ziming Shen, Yunqi Tao, Seongjin Park, and William Saunders. 2025. Reasoning Models Don't Always Say What They Think. *arXiv preprint arXiv:2601.02314* (2025).

[5] European Parliament and Council of the European Union. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council (Artificial Intelligence Act). Official Journal of the European Union.

[6] Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamil Lukić, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Catherine Olsson, Roger Grosse, Dario Amodei, Tristan Kravec, Jared Kaplan, and Jack Clark. 2023. Measuring Faithfulness in Chain-of-Thought Reasoning. *arXiv preprint arXiv:2307.13702* (2023).

[7] Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Ellie Sakaguchi, Canyu Li, et al. 2024. HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal. *Proceedings of the 41st International Conference on Machine Learning* (2024).

[8] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training Language Models to Follow Instructions with Human Feedback. *Advances in Neural Information Processing Systems* 35 (2022).

[9] Alexander Pan, Chan Jun Shern, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Hanlin Ng, Horace Zhang, Scott Emmons, and Dan Hendrycks. 2023. Do the Rewards Justify the Means? Measuring Trade-Offs Between Rewards and Ethical Behavior in the MACHIAVELLI Benchmark. *Proceedings of the 40th International Conference on Machine Learning* (2023).

[10] Ethan Perez, Sam Ringer, Kamil Lukić, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, et al. 2022. Red Teaming Language Models with Language Models. *arXiv preprint arXiv:2202.03286* (2022).

[11] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2024. Towards Understanding Sycophancy in Language Models. In *International Conference on Learning Representations*.

[12] Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. 2022. Defining and Characterizing Reward Hacking. *Advances in Neural Information Processing Systems* 35 (2022).

[13] Alexandra Souly, Qingyun Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Murungi, and Soheil Feizi. 2024. StrongREJECT: A Rejection-Quality Benchmark for Evaluating Jailbreaking Defenses. *arXiv preprint arXiv:2402.10260* (2024).

[14] Marilyn Strathern. 1997. 'Improving Ratings': Audit in the British University System. *European Review* 5, 3 (1997), 305–321.

[15] Adrian Wedd. 2025. Failure-First Embodied AI: A Red-Teaming Framework for Characterizing Recursive Failure in Agentic Systems. (2025). Working paper.

[16] Adrian Wedd. 2025. Iatrogenic Safety: When AI Safety Measures Create New Harms. *Proceedings of AIES 2025* (2025). Under review.

[17] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2024. Jailbroken: How Does LLM Safety Training Fail? *Advances in Neural Information Processing Systems* 36 (2024).

[18] 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. 2024. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, Vol. 36.

[19] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and Transferable Adversarial Attacks on Aligned Language Models. *arXiv preprint arXiv:2307.15043* (2023).