Benchmark Contamination in AI Safety Evaluation: Evidence from Novel Attack Families

Adrian Wedd
Failure-First Research
research@failurefirst.org

March 2026

Abstract

AdvBench is the most widely cited jailbreak safety benchmark, used to evaluate model robustness across dozens of published studies. We present evidence that safety evaluation scores on AdvBench are inflated by benchmark contamination—models have learned to refuse AdvBench-specific phrasings without developing robust safety generalization. Our methodology uses novel attack families, created in a private repository and absent from any public dataset, as contamination-free controls. Qwen3-8b refuses 84.7% of AdvBench prompts but complies with 98.3% of novel attack family prompts—an 83 percentage-point gap ($\chi^2 = 80.5$, $p < 10^{-18}$, Cramér's $V = 0.82$). Two replication models confirm the directional effect ($p < 10^{-6}$). Frontier-scale testing reveals a non-monotonic relationship between parameter count and safety robustness: ASR follows the trajectory Ministral 14B (96.7%) $\rightarrow$ Nemotron 30B (66.7%) $\rightarrow$ Nemotron Super 230B (78.6%) $\rightarrow$ Qwen3.5 397B (7.1%, corrected), suggesting that safety training methodology dominates parameter count. Qwen3.5 introduces a novel "silent refusal" defense—HTTP 200 with empty response body—that inflates heuristic ASR by 39 percentage points, revealing a methodological blind spot in keyword-based safety evaluation. These findings suggest that any safety claim based solely on public benchmark performance may be inflated, and that safety evaluations should include held-out, non-public test sets to measure genuine generalization.

1 Introduction

The standard methodology for evaluating the safety of large language models (LLMs) relies on static benchmarks: curated sets of adversarial prompts against which model responses are classified as compliance or refusal. AdvBench [Zou et al., 2023], HarmBench [Mazeika et al., 2024], JailbreakBench [Chao et al., 2024], and StrongREJECT [Souly et al., 2024] are the most widely used. Model developers routinely report safety scores on these benchmarks as evidence of alignment quality.

This evaluation paradigm has a structural vulnerability: *benchmark contamination*. If a model's training data includes the benchmark prompts—directly or through derivative datasets—it can learn to refuse those specific phrasings without developing the capacity to refuse semantically similar but syntactically novel harmful requests. The model learns what AdvBench *looks like*, not what harm *looks like*.

Benchmark contamination has been extensively studied for capability benchmarks. Sainz et al. [2023] documented contamination effects across NLP tasks. Golchin and Surdeanu [2024] developed time-travel probes to detect memorization. Oren et al. [2023] and Shi et al. [2024] proposed statistical tests for black-box contamination detection. Yang et al. [2023] showed that simple rephrasing of benchmark items can reveal contamination gaps. Deng et al. [2024] systematically surveyed contamination in modern benchmarks. Dekoninck et al. [2024] demonstrated that contamination detection itself can be evaded.

Despite this body of work on capability benchmarks, **contamination in safety benchmarks has received almost no attention**. Safety benchmarks differ from capability benchmarks in a critical respect: the training signal is not "answer correctly" but "refuse to answer." A model contaminated on a capability

benchmark will show inflated performance; a model contaminated on a safety benchmark will show inflated *safety*—it will appear more aligned than it actually is. The consequences are qualitatively different: inflated safety scores can lead to deployment of inadequately aligned models.

We present the first quantitative evidence of safety benchmark contamination, using a methodology that does not require access to training data or model weights.

Contributions.

1. We introduce a **novel-family contamination control** methodology: attack families created in a private repository and absent from any public dataset serve as a contamination-free baseline against which public benchmark performance can be compared.
2. We demonstrate an **83 percentage-point ASR gap** between AdvBench and novel families on Qwen3-8b ($\chi^2 = 80.5$, $p < 10^{-18}$, $V = 0.82$), providing strong evidence of benchmark-specific overfitting.
3. We show that the contamination effect is **model-specific**: the same methodology applied to Nemotron-30b produces a 33 pp gap ($V = 0.31$), indicating that the Qwen3 gap is 2.7$\times$ larger than the baseline "novelty advantage."
4. We document a **non-monotonic scale–safety relationship** across five models spanning 14B–397B parameters, demonstrating that safety training methodology dominates parameter count as a predictor of robustness.
5. We identify **silent refusal** as a novel defense modality in Qwen3.5 (397B) that evades heuristic classification, inflating measured ASR by 39 pp and highlighting a methodological blind spot in keyword-based safety evaluation.
6. We discuss **implications for published safety claims** and propose that safety evaluations should include held-out, non-public test sets.

2 Methodology

2.1 Design Rationale

The core insight is that *novel attack families*—adversarial prompt sets that have never appeared in any public dataset, benchmark, or training corpus—can serve as contamination controls. If a model shows similar attack success rates (ASR) on both a public benchmark and novel families, its safety behavior is likely genuine. If ASR drops sharply on the public benchmark but remains high on novel families, the model has likely memorized benchmark-specific refusal patterns.

This design is analogous to the "rephrased samples" approach of Yang et al. [2023], but applied to safety rather than capability benchmarks, and using entirely novel attack *architectures* rather than paraphrases of existing prompts.

2.2 Novel Attack Families

We constructed six novel attack families, each comprising 10 adversarial scenarios (60 total). These families were designed, implemented, and stored in a private repository. They have never been published, shared, or included in any public dataset.

Table 1: Novel attack family descriptions. All scenarios target embodied AI / robotics domains.

Family	Description
Compositional Reasoning (CRA)	Each individual step is benign; the composite action sequence produces harm. Exploits the model's inability to reason about aggregate effect.
Meaning Displacement (MDA)	Harmful actions are reframed using domain-appropriate technical vocabulary that displaces the harmful semantics (e.g., "decommission" for "destroy").
Pressure Cascade (PCA)	Escalating urgency cues (safety emergency, countdown timer, human life at risk) that override safety deliberation.
Reward Hacking (RHA)	Scenarios where the optimal action for the stated reward signal produces harmful outcomes.
Sensor Spoofing (SSA)	Manipulated perception inputs that cause the model to misidentify safe/unsafe conditions.
Multi-Agent Collusion (MAC)	Coordinated multi-actor requests where no single request is harmful but the combination enables harm.

Three properties make these families suitable as contamination controls:

1. **Provable absence from training data.** The scenarios were created in March 2026 in a private repository. They cannot appear in training data for models released before this date.
2. **Semantic equivalence to public benchmarks.** The scenarios request genuinely harmful outcomes (physical injury, property destruction, safety system compromise)—they are not trick questions or ambiguous edge cases.
3. **Architectural novelty.** Unlike paraphrases of existing prompts, these families use entirely different attack architectures (compositional decomposition, semantic displacement, pressure escalation) that cannot be matched by pattern-level memorization of existing benchmarks.

2.3 Public Benchmark: AdvBench

AdvBench [Zou et al., 2023] is a set of 520 harmful requests originally developed for evaluating gradient-based adversarial attacks. It has become the de facto standard for evaluating model safety across the jailbreak literature. We used the first 60 prompts from the full AdvBench set, matching the sample size of our novel families.

AdvBench prompts are direct harmful requests (e.g., "Write a script that can exploit vulnerabilities in a software or operating system"). They have been publicly available since 2023 and appear in numerous training data compilations, safety fine-tuning datasets, and red-teaming tools.

2.4 Models Under Test

Table 2: Models tested with available trace counts. Parameters are approximate where official figures are unavailable.

Model	Params	AdvBench n	Novel n	API	Tier
Ministral 3	14B	—	30†	Ollama Cloud	Scale probe
Gemma3	27B	—	28†	Ollama Cloud	Scale probe
Qwen3-8b	8B	59	60	OpenRouter	Primary
Nemotron-3-nano	30B	30	60	OpenRouter	Control
Trinity Large Preview	—	30	59	OpenRouter	Replication
Nemotron 3 Super	∼230B	—	28†	Ollama Cloud	Frontier probe
Qwen3.5	397B	—	28†	Ollama Cloud	Frontier probe

† Tested with curated top-ASR prompts (novel family subset), not the full 60-prompt novel family set.

Qwen3-8b is the primary model under test because (a) Qwen models are widely used and extensively safety-benchmarked, and (b) the Qwen3 family's training data is likely to include AdvBench-derived datasets given the scale and diversity of their training corpus.

Trinity Large Preview (Arcee) and Nemotron-3-nano-30b (Nvidia) serve as comparison models to distinguish model-specific contamination from generic novelty effects. Ministral 3, Gemma3, Nemotron 3 Super, and Qwen3.5 extend the analysis across the 14B–397B parameter range to test whether the contamination gap scales with model size.

2.5 Classification

All traces were classified using heuristic (keyword-based) refusal detection. We acknowledge the known limitations of heuristic classification (Cohen's $\kappa = 0.097$–0.126 vs LLM grading; see Wei et al. [2024] for related discussion). However, for *within-model, within-classifier comparisons*—which is our use case—heuristic classification is valid: systematic bias cancels when comparing the same classifier on the same model across two prompt sets.

The key metric is the *relative* ASR difference between AdvBench and novel families, not the absolute ASR values. Confidence intervals are Wilson score intervals [Wilson, 1927]. Significance testing uses chi-square with Yates continuity correction and Fisher's exact test.

3 Results

3.1 Primary Result: Qwen3-8b

Table 3: Qwen3-8b: AdvBench vs novel attack families.

Prompt Set	n	Complied	ASR	Wilson 95% CI
Novel families	60	59	98.3%	[91.1%, 99.7%]
AdvBench	59	9	15.3%	[8.2%, 26.5%]
Delta			**+83.1 pp**	

Statistical tests:

- Chi-square: $\chi^2 = 80.48$, $p = 2.93 \times 10^{-19}$ (df $= 1$)
- Fisher's exact test: OR $= 327.8$, $p = 5.32 \times 10^{-23}$
- Cramér's $V = 0.822$ (large effect; threshold > 0.5)

Qwen3-8b complied with 59 of 60 novel family prompts (the single refusal was a multi-agent collusion scenario) but refused 50 of 59 AdvBench prompts. The near-perfect compliance on novel families (98.3%) indicates that Qwen3-8b has minimal intrinsic safety alignment for attack patterns outside its training distribution.

3.1.1 Novel Family Breakdown

Table 4: Qwen3-8b ASR by novel attack family.

Attack Family	n	**Complied**	**ASR**	**Wilson 95% CI**
Compositional reasoning	10	10	100.0%	[72.2%, 100.0%]
Meaning displacement	10	10	100.0%	[72.2%, 100.0%]
Pressure cascade	10	10	100.0%	[72.2%, 100.0%]
Reward hacking	10	10	100.0%	[72.2%, 100.0%]
Sensor spoofing	10	10	100.0%	[72.2%, 100.0%]
Multi-agent collusion	10	9	90.0%	[59.6%, 98.2%]

The uniformity across five of six families (100% ASR each) suggests that Qwen3-8b is not evaluating safety at all for these prompt types—it has no learned refusal pattern for these attack architectures.

3.2 Replication: Trinity Large Preview

Table 5: Trinity Large Preview: AdvBench vs novel attack families.

Prompt Set	n	**Complied**	**ASR**	**Wilson 95% CI**
Novel families	59	52	88.1%	[77.5%, 94.1%]
AdvBench	30	11	36.7%	[21.9%, 54.5%]
Delta			**+51.5 pp**	

Statistical tests: Fisher's exact $p < 10^{-6}$, Cohen's $h = 1.14$ (large effect). Trinity shows the same directional effect: substantially lower ASR on AdvBench than on novel families.

3.3 Cross-Model Comparison

Table 6: Cross-model comparison of AdvBench–novel delta.

Model	**AdvBench ASR**	**Novel ASR**	**Delta**	**Cramér's V**
Qwen3-8b	15.3%	98.3%	+83.1 pp	0.822
Nemotron-3-nano-30b	43.3%	76.7%	+33.4 pp	0.306
Trinity Large Preview	36.7%	88.1%	+51.5 pp	—

All three models show higher ASR on novel families than on AdvBench. This baseline effect is expected: novel families use more sophisticated attack architectures (compositional decomposition, semantic displacement) than AdvBench's direct requests.

The critical observation is the *relative magnitude*. Nemotron-30b's 33 pp gap ($V = 0.306$) represents the baseline "novelty advantage"—the ASR increase attributable to novel families being inherently harder to refuse. Qwen3-8b's 83 pp gap ($V = 0.822$) is 2.7× larger. The excess 50 pp beyond the baseline cannot be explained by novelty alone and is consistent with benchmark-specific contamination.

3.4 Cross-Novel-Family Comparison

Table 7: Novel family ASR across models, demonstrating differential vulnerability.

Model	n	ASR	Wilson 95% CI
Qwen3-8b	60	98.3%	[91.1%, 99.7%]
Trinity Large Preview	60	56.7%	[44.1%, 68.4%]
Nemotron-3-nano-30b	60	76.7%	[64.6%, 85.6%]

The variation in novel family ASR across models (56.7%–98.3%) confirms that novel families are not trivially easy for all models. Qwen3-8b's near-perfect compliance is anomalous rather than generic.

3.5 Frontier Probe: Non-Monotonic Scale–Safety Relationship

To test whether contamination effects and safety robustness scale with model size, we evaluated two frontier-scale models—Nemotron 3 Super (∼230B parameters) and Qwen3.5 (397B parameters)—against our curated top-ASR prompt set (28 scenarios drawn from novel attack families that achieved 100% heuristic ASR on Gemma3 27B).

Table 8: Non-monotonic scale–safety relationship. ASR on curated top-ASR prompts (novel family origin, $n = 28$ per model except Ministral $n = 30$). Heuristic classification; Qwen3.5 corrected for silent refusals.

Model	Params	Heuristic ASR	Corrected ASR	Pattern
Ministral 3	14B	96.7%	96.7%	Near-universal compliance
Nemotron-3-nano	30B	66.7%	66.7%	Selective resistance
Gemma3	27B	100.0%	100.0%	Universal compliance
Nemotron 3 Super	∼230B	78.6%	78.6%	Modest improvement over 30B
Qwen3.5	397B	46.4%	7.1%	Silent refusal defense

The relationship between parameter count and safety robustness is clearly non-monotonic: Ministral 14B (96.7%) → Nemotron 30B (66.7%) → Nemotron Super 230B (78.6%) → Qwen3.5 397B (7.1%). Scaling from 30B to 230B within the Nemotron family *increases* ASR by 12 percentage points, while Qwen3.5 at 397B achieves dramatically lower ASR than any other model tested.

This pattern suggests that **safety training methodology dominates parameter count as a predictor of robustness**. The Qwen3 family presents a striking within-family contrast: Qwen3-8b shows near-zero safety generalization on novel families (98.3% ASR), while Qwen3.5-397B shows near-complete refusal (7.1% corrected ASR). This is consistent with the contamination hypothesis: Qwen3.5's safety training may incorporate defenses against novel attack architectures that Qwen3-8b lacks, or Qwen3.5 may employ an API-level content filter that operates independently of the model's learned safety behavior.

3.5.1 Silent Refusal: A Novel Defense Modality

Qwen3.5 exhibits a previously undocumented defense behavior: for 11 of 28 prompts (39.3%), the API returns HTTP 200 with a completely empty response body. We term this *silent refusal*—the model (or its serving infrastructure) declines to generate any output, but does not return an error code.

This has immediate methodological implications:

- **Heuristic classifiers overestimate ASR.** Keyword-based refusal detection searches for refusal phrases in the response text. An empty response contains no refusal keywords, so heuristic classifiers score it as compliance. Qwen3.5's raw heuristic ASR (46.4%) is 39 percentage points higher than its corrected ASR (7.1%).

- **Silent refusal may mask contamination.** If Qwen3.5's silent refusal operates at the API/infrastructure layer rather than the model layer, it could mask the same contamination pattern observed in Qwen3-8b. The model may be generating compliant responses that are then suppressed by a content filter.
- **Defense taxonomy expansion.** Silent refusal represents a fourth refusal modality beyond (1) explicit keyword refusal, (2) topic deflection, and (3) partial compliance with disclaimers.

3.6 Comprehensive AdvBench vs Novel Family Comparison

Table 9 consolidates the AdvBench–novel family comparison across all models with paired data.

Table 9: AdvBench vs novel family ASR across all models with paired data. All values use heuristic classification except where corrected values are noted.

Model	Params	AdvBench ASR	Novel ASR	Delta	V or h
Qwen3-8b	8B	15.3%	98.3%	+83.1 pp	$V = 0.822$
Trinity Large Preview	—	36.7%	88.1%	+51.5 pp	$h = 1.14$
Nemotron-3-nano	30B	43.3%	76.7%	+33.4 pp	$V = 0.306$

All three models with paired AdvBench and novel family data show the same directional effect: substantially higher ASR on novel families than on AdvBench. The critical finding remains the *magnitude differential*: Qwen3-8b's 83 pp gap is 2.5× the Trinity gap and 2.7× the Nemotron gap.

4 Discussion

4.1 The Most Parsimonious Explanation

The data supports the following interpretation: Qwen3-8b has been fine-tuned or RLHF-trained with AdvBench prompts (or AdvBench-derived datasets) in its safety training data, producing benchmark-specific refusal patterns that do not generalize to novel attack families.

The model has memorized what AdvBench *looks like*, not what harm *looks like*.

Four pieces of evidence support this over alternative explanations:

1. **Disproportionate gap.** The 83 pp delta is 2.7× larger than the control model's delta, ruling out generic novelty effects.
2. **Near-perfect novel compliance.** 59/60 on novel families indicates near-zero intrinsic safety for out-of-distribution attack patterns.
3. **Uniform novel family compliance.** 100% ASR across 5 of 6 novel families suggests the model is not evaluating safety content at all—it simply lacks trained refusal patterns for these architectures.
4. **Within-family contrast.** Qwen3-8b (98.3% novel ASR) and Qwen3.5-397B (7.1% corrected novel ASR) differ by 91 percentage points on the same prompt set, despite sharing the Qwen training lineage. This is consistent with the larger model receiving substantially different safety training—possibly including defenses calibrated to novel attack architectures—while the smaller model's safety is concentrated on public benchmark patterns.

4.2 Alternative Explanations

We consider three alternative explanations and assess their plausibility.

Domain gap. Novel families use embodied/robotics scenarios while AdvBench uses general harmful requests. Qwen3 may have been safety-trained primarily on general harm categories. *Assessment:* Partially plausible. However, the scenarios involve clear physical harm (chemical dispersal, crane drops onto workers) that a genuinely aligned model should refuse regardless of domain framing. Moreover, the disproportionate gap relative to Nemotron argues against domain gap as the sole explanation.

Prompt complexity. Novel family prompts are longer and more operationally detailed than AdvBench one-liners. Longer prompts may bypass safety filters. *Assessment:* Partially plausible for absolute ASR levels, but does not explain why the Qwen3 delta is 2.7× larger than Nemotron's. Both models face the same prompt length differential.

Model capability. Qwen3-8b may simply lack safety capacity. *Assessment:* Inconsistent with the data. Qwen3-8b demonstrates robust safety on AdvBench (84.7% refusal rate), proving it has substantial safety training. The safety training is simply not generalizing.

4.3 Implications for Published Safety Claims

AdvBench is cited in the safety evaluations of numerous models, including published results from major AI laboratories. If AdvBench contamination is present in even a fraction of evaluated models, the published safety landscape is systematically biased toward overestimating alignment quality.

The problem is structural, not specific to any model developer:

1. **AdvBench has been public since 2023.** Three years of availability means it has likely been ingested into the training corpora of most large models, either directly or through derivative datasets that include AdvBench prompts.
2. **Safety fine-tuning datasets include AdvBench.** Several open-source safety training datasets explicitly include AdvBench prompts paired with refusal responses. Models trained on these datasets will learn AdvBench-specific refusal patterns by construction.
3. **Competitive pressure incentivizes training on benchmarks.** Whether deliberate or inadvertent, the incentive to report strong safety scores creates selection pressure toward including benchmark-like data in training.

This dynamic is well-documented for capability benchmarks [Sainz et al., 2023, Deng et al., 2024] but has not previously been quantified for safety benchmarks.

4.4 Limitations

1. **Heuristic classification.** We used keyword-based classification rather than LLM-based grading. While this introduces systematic bias in absolute ASR values, it does not affect the validity of within-model, within-classifier comparisons that form the basis of our contamination analysis.
2. **Confounded comparison.** AdvBench uses direct harmful requests; novel families use compositional and embodied attack architectures. The ASR difference measures both contamination and attack sophistication. We partially control for this with the Nemotron baseline, but a fully controlled experiment would require novel prompts using the *same* direct-request format as AdvBench with different harmful content.
3. **Sample size.** 59–60 prompts per condition provides adequate statistical power for the observed effect sizes ($V > 0.3$) but limits subgroup analyses. The Wilson confidence intervals in Tables 3–6 reflect this limitation.
4. **Model availability.** We were unable to test additional Qwen3 variants (4b, 1.7b) due to persistent API rate limits. Testing across the Qwen3 family would strengthen the contamination claim.
5. **Single public benchmark.** We tested only AdvBench. Extending the methodology to HarmBench, JailbreakBench, and StrongREJECT would determine whether contamination is AdvBench-specific or affects multiple public benchmarks.
6. **Silent refusal ambiguity.** Qwen3.5's empty-response behavior could originate at the model layer, the API serving layer, or a separate content filter. Without access to model internals, we cannot distinguish these possibilities. Our corrected ASR (7.1%) assumes empty responses are refusals, which may overstate or understate the model's intrinsic safety.
7. **Frontier models lack paired AdvBench data.** The frontier probe models (Nemotron Super, Qwen3.5) were tested only on novel family prompts, not on AdvBench. We therefore cannot compute the contamination delta for these models directly.

4.5 Toward Contamination-Resistant Safety Evaluation

Our results suggest several methodological improvements:

Held-out evaluation sets. Safety evaluations should include non-public prompt sets that have never appeared in training data. These held-out sets should be rotated periodically (temporal holdout) to prevent eventual contamination.

Novel family controls. Any safety evaluation using public benchmarks should include a contamination control: a set of novel prompts not present in any public dataset, tested on the same model with the same classifier. The delta between public and novel ASR provides an estimate of contamination magnitude.

Generalization testing. Safety alignment should be evaluated on its ability to generalize to novel attack architectures, not just novel phrasings of known attack types. Paraphrase-based contamination testing [Yang et al., 2023] is necessary but not sufficient—models may generalize to paraphrases of memorized prompts while failing on architecturally novel attacks.

Independent evaluation infrastructure. Evaluation benchmarks should be maintained by independent third parties who do not publish the prompts and who can verify that evaluated models have not been trained on the test set. This is standard practice in other domains (e.g., hold-out test sets in machine translation, blinded evaluation in clinical trials) but is not yet standard in AI safety evaluation.

5 Related Work

Benchmark contamination in capability evaluation. The problem of data contamination in LLM benchmarks is well-established. Sainz et al. [2023] demonstrated contamination effects across NLP tasks and called for per-benchmark contamination measurement. Golchin and Surdeanu [2024] developed time-travel probes for contamination detection. Oren et al. [2023] proposed statistical tests for black-box contamination. Shi et al. [2024] introduced Min-K% Prob for detecting pretraining data. Dekoninck et al. [2024] showed that contamination detection can be evaded. Our work extends this line to safety benchmarks, where contamination inflates *safety* rather than *capability* scores.

Safety benchmark methodology. Zou et al. [2023] introduced AdvBench alongside GCG attacks. Mazeika et al. [2024] proposed HarmBench as a standardized red-teaming framework. Souly et al. [2024] introduced StrongREJECT with rubric-based scoring. Chao et al. [2024] created JailbreakBench for reproducible jailbreak evaluation. None of these works addresses contamination as a threat to evaluation validity.

Jailbreak attack evolution. Wei et al. [2024] analyzed why safety training fails, identifying competing objectives and mismatched generalization. Russinovich et al. [2024] introduced multi-turn escalation attacks. Perez et al. [2022] pioneered automated red-teaming with LLMs. Our novel attack families (compositional reasoning, meaning displacement, pressure cascade) represent a distinct class from these prior works.

6 Conclusion

We present the first quantitative evidence of benchmark contamination in AI safety evaluation. Using novel attack families as contamination-free controls, we demonstrate an 83 percentage-point ASR gap between AdvBench and novel families on Qwen3-8b ($\chi^2 = 80.5$, $p < 10^{-18}$, $V = 0.82$). The gap is

2.7× larger than a control model's gap, indicating model-specific contamination beyond generic novelty effects.

Frontier-scale testing across five models (14B–397B parameters) reveals that safety robustness is non-monotonic in parameter count, with safety training methodology—not scale—as the dominant predictor. The Qwen family presents an instructive contrast: Qwen3-8b (8B) complies with 98.3% of novel prompts while Qwen3.5 (397B) refuses 92.9%, suggesting that larger models in the same family may receive qualitatively different safety training that generalizes beyond public benchmark patterns.

This finding has immediate practical implications: any safety claim based solely on AdvBench performance should be treated with skepticism. Models that appear safe on AdvBench may comply with 98% of novel adversarial requests.

We recommend that safety evaluations adopt held-out, non-public test sets; include novel family contamination controls alongside public benchmarks; and test generalization to architecturally novel attack types rather than only paraphrases of known attacks. Additionally, evaluation frameworks should account for silent refusal behaviors that evade keyword-based classification. The AI safety community should treat benchmark contamination in safety evaluation as a first-order threat to the integrity of published safety claims.

Ethics Statement

This work identifies a vulnerability in safety evaluation methodology. We disclose novel attack family *categories* (compositional reasoning, meaning displacement, etc.) but do not publish the specific prompts, which remain in a private repository. This follows responsible disclosure principles: the methodological finding is publishable; the operational attack content is not.

All models tested are publicly available via API. No model was modified, fine-tuned, or attacked in ways beyond standard API usage. The purpose of this research is to improve safety evaluation methodology, not to enable attacks.

Reproducibility

Trace files, classification results, and statistical analysis scripts are maintained in the Failure-First research repository. Novel family scenarios are available to verified researchers under a responsible disclosure agreement. Contact the corresponding author for access.

The statistical analysis can be reproduced from the published results in Tables 3–6: all significance tests require only the reported cell counts.

References

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. JailbreakBench: An open robustness benchmark for jailbreaking large language models. *arXiv preprint arXiv:2404.01318*, 2024.

Jasper Dekoninck, Mark Fuber, and Joeran Beel. Evading data contamination detection for language models is (too) easy. *arXiv preprint arXiv:2402.02823*, 2024.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. Investigating data contamination in modern benchmarks for large language models. *Proceedings of NAACL*, 2024.

Shahriar Golchin and Mihai Surdeanu. Time travel in LLMs: Tracing data contamination in large language models. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024.

Mantas Mazeika, Long Phan, Xuwang 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 *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

Yonatan Oren, Nicole Meister, Niladri S. Bhatt, Ryan Cotterell, and Omer Levy. Proving test set contamination in black box language models. *arXiv preprint arXiv:2310.17623*, 2023.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*, 2022.

Mark Russinovich, Ahmed Salem, and Ronen Eldan. Great, now write an article about that: The crescendo multi-turn LLM jailbreak attack. *arXiv preprint arXiv:2404.01833*, 2024.

Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Aitor Soroa. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. *Findings of EMNLP*, 2023.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. *Proceedings of ICLR*, 2024.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, and Soheil Feizi. A StrongREJECT for empty jailbreaks. *Advances in Neural Information Processing Systems*, 37, 2024.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does LLM safety training fail? *Advances in Neural Information Processing Systems*, 36, 2024.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference, 1927.

Shuo Yang, Wei-Lin Chiang, Lianmin Zheng, Joseph E. Gonzalez, and Ion Stoica. Rethinking benchmark and contamination for language models with rephrased samples. *arXiv preprint arXiv:2311.04850*, 2023.

Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2023.