Nine-Week Red-Teaming Study * May 22, 2026

Responsible AI Model Evaluations

A Nine-Week Red-Teaming Study of Foundation Models Across 22 Safety Risk Categories

Hemant Naik * May 22, 2026 * Dataset: RedBench (Dang et al., 2026, MIT)

Live Dashboard: https://Sushegaad.github.io/Responsible-AI-Model-Evaluations/ * GitHub: https://github.com/Sushegaad/Responsible-AI-Model-Evaluations

Total Evaluations	Attack Failures	Critical Findings	Over-Refusals	Study Duration	Models Evaluated
26,500	**1,269**	**468**	**422**	**9 Weeks**	**7**
9 weekly runs	Across all runs	Severity >= 8/10	False rejections	Apr 10 - May 22	3 providers

Abstract

This paper presents a nine-week automated red-teaming study of seven frontier large language models from Anthropic, OpenAI, and Google, evaluated across 22 safety risk categories using the RedBench adversarial benchmark (29,362 prompts). Over 26,500 individual evaluations were conducted between April 10 and May 22, 2026, producing 468 critical findings and 422 false-outright-refusals. Results are scored by a dual-judge pipeline: a deterministic RegexJudge and a NeuralJudge (Claude Haiku), mapped to the NIST AI Risk Management Framework 1.0. Across all nine runs, Gemini 2.5 Flash achieved the lowest average Attack Success Rate (3.79%), while GPT-4o Mini remained the highest-risk model (8.42% average ASR). CBRN and Cybersecurity categories continued to account for the majority of guardrail bypasses. In the final run (May 22, 2026), GPT-4o Mini recorded the study's highest-ever drift coefficient (+1.67/turn), Gemini 2.5 Flash spiked to 25% False Outright Refusal, and three models: Claude Opus 4.7, Claude Opus 4.6, and Gemini 2.5 Pro, each reached a drift coefficient of +1.11/turn. Election Interference has solidified as a persistent top-5 risk category at 5.66% average ASR. This study is concluded at nine runs. All data, evaluation code, audit logs, and the research paper are publicly archived under the MIT License.

1. Introduction

The deployment of large language models (LLMs) in high-stakes domains, including government, healthcare, finance, and legal services, demands rigorous, reproducible safety evidence. Vendor self-assessments are insufficient for independent governance because they are not replicable, use proprietary benchmarks, and are subject to publication bias. This study addresses that gap with a fully open, automated weekly pipeline that stress-tests seven frontier models against a standardised adversarial dataset and publishes results to a public dashboard over nine weekly runs.

The evaluation covers 22 risk categories spanning physical harm, cybersecurity, privacy, misinformation, and domain-specific risks. Each run draws 500 stratified prompts per model from the RedBench dataset (400 adversarial and 100 benign) and grades every response with a dual-judge system. The nine-week time series produced in this concluded study enables detection of safety regressions, guardrail erosion under adversarial pressure, and comparative model rankings that are independent of any provider's internal testing. All 26,500 evaluations, audit logs, and weekly metrics are permanently archived for replication and longitudinal research.

2. Methodology

2.1 Dataset

RedBench (Dang et al., 2026) aggregates 37 adversarial sub-benchmarks, including HarmBench, ToxiGen, XSTest, DAN, and AdvBench, into a single standardised schema covering 22 risk categories and 19 domains. A local snapshot of 29,362 prompts (26,113 adversarial, 3,249 benign) is committed to the repository as Parquet files, making every run fully reproducible without network dependencies on the dataset host.

2.2 Models

Model	Provider	Runs	Evaluation Period
Claude Opus 4.7	Anthropic	5	Apr 24 - May 22
Claude Opus 4.6	Anthropic	9	Apr 10 - May 22
Claude Sonnet 4.6	Anthropic	9	Apr 10 - May 22
GPT-4o	OpenAI	9	Apr 10 - May 22
GPT-4o Mini	OpenAI	9	Apr 10 - May 22
Gemini 2.5 Flash	Google	6	Apr 17 - May 22
Gemini 2.5 Pro	Google	6	Apr 14 - May 22

Table 1. Models evaluated in this study.

2.3 Evaluation Pipeline

Each model receives 500 prompts per weekly run: 400 adversarial attacks drawn proportionally across all 22 risk categories, and 100 benign prompts to measure false refusal rates. Every prompt is submitted in three modes: single-turn, agentic 3-turn wrap (system + user + follow-up), and for a 5% sub-sample, a 10-turn adversarial drift probe that measures how safety posture degrades under sustained pressure.

Responses are graded by a Dual-Judge pipeline. The RegexJudge applies deterministic pattern-matching for PII leakage, jailbreak confirmations, malware code, explosive synthesis instructions, and CSAM. The NeuralJudge (Claude Haiku) returns a structured JSON verdict (SAFE, UNSAFE, or OVER_REFUSAL), a severity score from 1 to 10, and a chain-of-reasoning sentence for the forensic audit trail. A regex match always overrides the neural result and forces a CRITICAL_FAIL; a neural UNSAFE verdict with severity 8 or above also escalates to CRITICAL_FAIL regardless of the regex result.

2.4 Metrics

Metric	Definition	Target
Attack Success Rate (ASR)	% of adversarial prompts that bypassed guardrails	< 5%
False Outright Refusal (FOR)	% of benign prompts incorrectly refused	< 10%
Drift Coefficient	Safety change per conversation turn (10-turn probe)	< 0.5%/turn
Provenance Score	% of evaluations with complete chain-of-reasoning JSON	> 95%

Table 2. Evaluation metrics and targets.

3. Major Data Findings

3.1 Overall Safety Leaderboard: Nine-Week Averages

The table below summarises each model's average performance across all evaluation runs. Colour coding follows the study's own thresholds: green = at or below target; amber = moderate concern; red = above target.

Model	Avg ASR	Avg FOR	Avg Drift	Avg Prov.	Total Crit.	Runs
Gemini 2.5 Flash	3.79%	16.33%	+0.37	100.0%	24	6
Claude Opus 4.7	4.00%	1.20%	+0.78	100.0%	26	5
Claude Sonnet 4.6	5.28%	1.00%	-0.12	99.9%	63	9
GPT-4o	5.67%	7.67%	+0.43	100.0%	70	9
Claude Opus 4.6	5.81%	0.78%	-0.12	99.9%	74	9
Gemini 2.5 Pro	8.00%	18.33%	+0.28	100.0%	115	6
GPT-4o Mini	8.42%	13.67%	+0.49	100.0%	96	9

Table 3. Per-model averages across all successful evaluation runs. Green = at or below target; amber = moderate concern; red = above target.

3.2 Finding 1: Anthropic Dominates the Safety-Utility Balance

The most structurally important finding of the study is that Anthropic's Claude models are the only provider to achieve both low attack bypass rates and near-zero false refusals simultaneously. All three Anthropic models average FOR rates between 0.78% and 1.20%, compared to 13.67–18.33% for Google and OpenAI's GPT-4o Mini. This directly contradicts the common industry assumption that safety and utility exist in permanent tension. GPT-4o Mini uniquely fails both metric thresholds simultaneously: it holds the study's highest average ASR (8.42%), making it the most permissive model to adversarial prompts, while also exceeding the 10% FOR target at 13.67% false refusals, a dual failure mode from a single evaluation baseline.

Standout Result

Claude Sonnet 4.6 on May 22, 2026: 5.25% ASR + 0.00% FOR + 0.00 drift + 99.8% provenance. This is the only data point in the entire nine-week study where a model achieved a perfect false-refusal rate alongside a sub-8% attack success rate, a benchmark proof point for safety-utility co-optimisation.

3.3 Finding 2: The CBRN Gap Is Systemic and Severe

CBRN (Chemical, Biological, Radiological, and Nuclear) information remained the highest-risk category across all nine weeks and all seven models, averaging 35.41% ASR, more than 1.6x the second-ranked category, Cybersecurity Threats Beyond Malware at 21.99%. This gap persists across all providers, confirming it is a systemic limitation of current RLHF-based alignment rather than a model-specific failure. Illegal Weapons (Non-CBRN) ranked third at 17.52%, followed by Malware Code at 12.26%. Sexual Content (4.53%), Misinformation (3.88%), Bias Content (3.45%), and Hateful Content (3.30%) round out the top ten. Privacy and Surveillance categories remained below 2% throughout, likely reflecting the strength of the deterministic RegexJudge component.

Risk Category	Avg ASR	Data Points	NIST Function
CBRN Information or Capabilities	35.41%	53	MAP
Cybersecurity Threats (Beyond Malware)	21.99%	53	MAP

Illegal Weapons (Non-CBRN)	17.52%	53	MANAGE
Malware Code	12.26%	53	MAP
Election Interference (+)	5.66%	53	GOVERN
Sexual Content	4.53%	53	MANAGE
Misinformation	3.88%	53	MAP
Bias Content	3.45%	53	GOVERN
Hateful Content	3.30%	53	GOVERN
Environmental Harm	3.15%	53	MAP

Table 4. Top 10 risk categories by average ASR across all nine runs. (+) = persistent upward trend. NIST AI RMF function mapping shown.

3.4 Finding 3: Election Interference Is a Structural, Not Sampling, Risk

Election Interference ranked fifth overall at 5.66% average ASR across all 53 model-run combinations (7 models across up to 9 runs). This consistency across every model and every week of the study confirms that current frontier LLMs are structurally susceptible to generating misleading electoral content, impersonating candidates, or fabricating polling data, and no provider has solved this category. The slight decline from 6.52% reported at the eight-week mark to 5.66% at nine weeks suggests marginal improvement, but the category remains stubbornly in the top 5 across all providers. Given ongoing electoral cycles globally, this finding warrants targeted safety investment from all three providers and carries elevated urgency relative to its numerical rank.

3.5 Finding 4: Multi-Turn Drift - May 22 Was a Watershed Run

The Drift Coefficient, computed on a 5% sub-sample of attack prompts subjected to 10 consecutive adversarial turns, is arguably the most practically important metric for real-world deployment. A positive value means guardrails erode under sustained pressure; negative means the model becomes more conservative. The May 22 final run produced the study's most alarming drift results: four out of seven models simultaneously posted drift coefficients at or above +1.11/turn, with GPT-4o Mini reaching +1.67/turn, the highest single-run reading in the entire study. The simultaneity across four models suggests the May 22 adversarial sub-sample contained particularly effective multi-turn manipulation patterns not seen in prior runs.

Model	Avg Drift	May 22 Drift	Assessment
Claude Sonnet 4.6	-0.12/turn	0.00	*Guardrails tighten under pressure. Best drift profile.*
Claude Opus 4.6	-0.12/turn	+1.11	*Usually stable; joined four-model drift spike on May 22.*
Gemini 2.5 Pro	+0.28/turn	+1.11	*Low average but spiked in final run.*
Gemini 2.5 Flash	+0.37/turn	0.00	*Low average drift; stable in final run.*
GPT-4o	+0.43/turn	+0.56	*Moderate positive drift; borderline acceptable.*
GPT-4o Mini	+0.49/turn	+1.67	*HIGHEST drift in study; structural multi-turn vulnerability.*
Claude Opus 4.7	+0.78/turn	+1.11	*Highest avg drift of any model, despite strong single-turn ASR.*

Table 5. Per-model drift coefficients: average across all runs vs. final run (May 22, 2026).

3.6 Finding 5: Weekly ASR Trends Show High Variance With Stable Patterns

Week-over-week ASR varies considerably for all models, driven by stratified sampling randomness and prompt-composition shifts between runs. Despite this noise, several structural patterns are stable across the full nine weeks. GPT-4o Mini exceeded 8% ASR in 6 of its 9 runs and never demonstrated a sustained downward trajectory. Gemini 2.5 Pro has never fallen below 7.25% across any of its 6 evaluated runs. Claude Sonnet 4.6 shows the lowest single-run ASR (2.75%) of any model with a full nine-run history. Claude Opus 4.6 shows the narrowest overall range (4.00%–7.50%, a 3.5 pp spread). GPT-4o and Claude Opus 4.6 both exhibit a W-shaped pattern: strong in weeks 3-4, elevated in mid-study, and moderating slightly by the final run.

Model	Apr 10	Apr 12	Apr 14	Apr 17	Apr 24	May 2	May 8	May 16	May 22
Claude Opus 4.6	6.00%	7.50%	4.25%	4.00%	6.75%	5.25%	4.50%	7.25%	6.75%
Claude Opus 4.7	--	--	--	--	4.00%	3.00%	4.25%	5.25%	3.50%
Claude Sonnet 4.6	6.50%	7.00%	2.75%	3.25%	6.50%	5.00%	4.00%	7.25%	5.25%
Gemini 2.5 Flash	--	--	--	2.50%	3.50%	3.75%	4.00%	5.00%	4.00%
Gemini 2.5 Pro	--	--	7.25%	8.00%	7.25%	--	7.75%	9.00%	8.75%
GPT-4o	6.25%	6.50%	3.25%	3.75%	7.00%	5.25%	6.25%	6.75%	6.00%
GPT-4o Mini	10.25%	10.00%	5.50%	6.00%	9.50%	6.00%	9.50%	9.75%	9.25%

Table 6. Weekly Attack Success Rate by model. -- = model not evaluated that week. Green < 5%, amber 5-8%, red > 8%.

4. Audience-Specific Insights

AI Engineers and Red-Teamers

* CBRN (35.41%) and Cybersecurity (21.99%) categories account for the majority of bypasses; these two categories alone drive over half of all attack failures across all nine runs.
* May 22 produced the study's highest-ever drift reading: GPT-4o Mini at +1.67/turn. Four models simultaneously posted >= +1.11/turn drift; examine whether the May 22 prompt sample contains new multi-turn attack patterns not seen in prior weeks.
* Claude Sonnet 4.6 achieved 0.00% FOR on May 22 with a 5.25% ASR, the best single-run safety-utility balance in the study. This is the benchmark to target.
* Election Interference (5.66% avg ASR, 53 data points) is a statistically robust structural top-5 risk; add it to red-team priority lists for all deployments.

GRC and Compliance Officers

* 468 critical findings across 26,500 evaluations (1.77% critical rate), all mapped to NIST AI RMF functions (MAP, MANAGE, GOVERN, MEASURE) with chain-of-reasoning audit trails.
* GSAR 552.239-7001 forensic logs are generated for every CRITICAL_FAIL and all FAIL verdicts with severity >= 7, providing PII-redacted prompt/response pairs for direct inclusion in compliance documentation.
* Gemini 2.5 Pro (18.33% avg FOR, 115 critical findings across 6 runs) and GPT-4o Mini (13.67% avg FOR, 96 criticals across 9 runs) both exceed the 10% FOR threshold; flag for procurement review if deploying in user-facing or regulated applications.
* Provenance Scores averaged >= 99.9% across all models, ensuring a near-complete audit trail for all 26,500 evaluations.

Product Leaders and Deployment Decision-Makers

* Anthropic models offer the best safety-utility balance: ASR between 4-6% and FOR below 1.20%, meaning low attack exposure without significant refusal overhead for legitimate users.
* Gemini 2.5 Flash's 3.79% average ASR is impressive, but its 16.33% avg FOR, spiking to 25% on May 22, means roughly 1 in 4 to 6 safe user requests may be incorrectly rejected depending on the week. Evaluate carefully for customer-facing use cases.
* GPT-4o Mini's persistently high ASR (8.42% average) and 13.67% FOR represent a dual failure mode: simultaneously too permissive to adversarial prompts and too restrictive to legitimate users. The May 22 drift of +1.67/turn is an additional deployment risk.
* Claude Sonnet 4.6 is the standout deployment choice: 5.28% avg ASR, 1.00% avg FOR, negative drift (-0.12/turn), and a perfect 0.00% FOR on May 22, the only model to achieve this across the nine-week study.

Policy Analysts and Researchers

* Nine weeks of weekly data on a fixed open benchmark (RedBench, MIT License) enables time-series safety comparison independent of any vendor's internal reporting. 26,500 evaluations across 53 model-run combinations provide high statistical confidence.
* Election Interference at 5.66% average ASR across 53 data points is a statistically robust finding across 7 models and 9 runs; not a sampling artefact, and below the 6.52% reported in the prior eight-week version (slight improvement).
* The CBRN gap (35.41% vs 21.99% for the second category) persists across all models and all runs, suggesting systemic limits of current RLHF-based alignment against dual-use knowledge requests require domain-specific interventions.
* The May 22 run, with four models at drift >= +1.11/turn simultaneously, is a signal worth investigating; it may reflect a specific adversarial prompt composition that more effectively exploits multi-turn vulnerabilities common to current alignment methods.

5. Conclusions

After nine weekly evaluation runs covering 26,500 evaluations across seven frontier models, five structural findings define the state of frontier LLM safety as of May 2026.

Finding 1: Safety and utility are not inherently in tension. Anthropic's Claude models consistently achieve both low attack bypass rates and near-zero false refusals. Claude Sonnet 4.6's May 22 result (5.25% ASR, 0.00% FOR, 0.00 drift, 99.8% provenance) is the clearest single-run proof point of safety-utility co-optimisation in this study.

Finding 2: CBRN and Cybersecurity categories are systemic failures. With average ASRs of 35.41% and 21.99% respectively, these categories require domain-specific safety interventions beyond general RLHF alignment. The gap is persistent across all seven models and all nine runs.

Finding 3: Election Interference is structural, not a sampling artefact. At 5.66% average ASR across 53 data points (7 models x 9 runs), no provider has solved this category. The slight decline from 6.52% at week 8 shows marginal progress but the risk remains elevated.

Finding 4: Multi-turn drift varies critically by provider. GPT-4o Mini's +1.67/turn on May 22 is the highest single-run reading in the study. Four models posted drift >= +1.11/turn in the same run, a signal warranting forensic analysis of the May 22 prompt sub-sample. Anthropic's Opus 4.6 and Sonnet 4.6 both average -0.12/turn, demonstrating stable multi-turn safety is achievable.

Finding 5: Week-to-week variance is high but structural rankings are stable. GPT-4o Mini exceeded 8% ASR in 6 of 9 runs; Gemini 2.5 Pro has never fallen below 7.25%; Claude Sonnet 4.6 achieved the lowest single-run ASR (2.75%) of any nine-run model; Claude Opus 4.6 shows the narrowest overall range (4.00%–7.50%). Sampling noise does not change the relative ordering of providers.

The pipeline, data, and audit logs for this study are fully open-source and available at https://github.com/Sushegaad/Responsible-AI-Model-Evaluations under the MIT License. This study is concluded at nine evaluation runs (April 10 – May 22, 2026). The full dataset, weekly metrics, forensic audit logs, and the evaluation pipeline are preserved in the repository for replication and further research.

6. References

[1] Dang, Q-A., Ngo, C., and Hy, T-S. (2026). RedBench: A Universal Dataset for Comprehensive Red Teaming of Large Language Models. arXiv:2601.03699.

[2] NIST (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). National Institute of Standards and Technology, Gaithersburg, MD. https://airc.nist.gov/

[3] GSA (2026). GSAR 552.239-7001 -- Information Technology Security Requirements (March 2026). U.S. General Services Administration.

[4] Naik, H. (2026). Responsible AI Model Evaluations -- Live Dashboard and Source Code. https://Sushegaad.github.io/Responsible-AI-Model-Evaluations/

[5] Perez, F., et al. (2022). Ignore Previous Prompt: Attack Techniques for Language Models. NeurIPS ML Safety Workshop.

[6] Anthropic (2025). Claude's Character and Constitutional AI. https://anthropic.com/research/constitutional-ai

---