The State of Adversarial AI Safety 2026

March 2026

Failure-First Research

193 models evaluated • **133,033** attack–response pairs • **36** attack families

LLM-based classification with measured inter-rater reliability

Classification: PUBLIC — Pattern-Level Analysis

Citation: Failure-First Research. "The State of Adversarial AI Safety 2026." March 2026. `failurefirst.org/state-of-adversarial-ai-safety-2026`

Data Sources: F41LUR3-F1R57 Adversarial AI Corpus
Methodology: FLIP grading (LLM-based classification), Wilson score confidence intervals, chi-square significance testing with Bonferroni correction

Contents

1 Executive Summary

This report presents findings from what we believe to be the largest independent adversarial AI safety evaluation conducted to date: 133,033 attack–response pairs across 193 models, 36 attack families, and 15+ providers, graded by LLM-based classifiers with measured inter-rater reliability.

Five headline findings:

1. **Safety training teaches recognition, not inhibition.** In 34.2% of cases where models comply with harmful requests, their reasoning traces contain explicit acknowledgment that the request is problematic ($n = 801$ compliant results with thinking traces, 24 models). Reasoning models override their own safety detection 69.7% of the time.
2. **Your provider matters more than your model.** Provider identity explains more ASR variance than architecture or parameter count. The spread between the most restrictive provider (Anthropic, 11.0% broad ASR) and the most permissive with substantial data (Liquid, 61.1%) is 5.6×. Within-provider models show correlated vulnerability at the prompt level (ϕ coefficients 0.24–0.43 for restrictive providers).
3. **Published safety benchmarks are contaminated.** Qwen3-8b refuses 84.7% of AdvBench prompts but complies with 98.3% of novel attack families not present in any public dataset—an 83 percentage-point gap ($\chi^2 = 80.5$, $p < 10^{-18}$, Cramér's $V = 0.82$).
4. **Format compliance and safety are partially independent.** Format-lock attacks shift frontier models from restrictive (<10% ASR) to mixed (20–47% ASR) vulnerability profiles—a 3–10× increase. This is the only attack family that maintains elevated ASR above the 7B parameter capability floor.
5. **No major framework tests embodied AI.** Seven of our 35 attack families have zero coverage in any external framework (MITRE ATLAS, OWASP, Garak, PyRIT, DeepTeam). The action layer of vision-language-action models shows a 0% refusal rate across 63 graded traces.

2 Methodology

2.1 Corpus Overview

Table 1: Corpus statistics at time of publication.

Metric	Value
Models evaluated	193 (180 with results)
Total evaluation results	133,033
LLM-graded results	53,831
Unique prompts	141,054
Attack families	36
Source datasets	31
Providers	15+
VLA (embodied AI) attack families	33
VLA scenarios	411

Models span the full spectrum: from sub-1B parameter open-weight models to frontier systems (Claude Sonnet 4.5, GPT-5.2, Gemini 3 Flash). The corpus includes models from Anthropic, Google, OpenAI, Meta, Nvidia, Mistral, DeepSeek, Liquid, StepFun, Qwen, Arcee, and others.

2.2 Grading: Why Classification Methodology Matters

A central finding of this research is that *grading methodology has a first-order impact on reported safety numbers*. We use three approaches and always specify which is being cited:

- **LLM-only:** A language model reads the response and classifies it using the FLIP (Forward-Looking Inference of Prompt) methodology. The grader infers what instruction the model was following, then judges whether the response constitutes harmful compliance.
- **Heuristic-only:** Keyword and structural pattern matching.
- **COALESCE:** LLM verdict preferred when available, falling back to heuristic.

Inter-rater reliability between heuristic and LLM grading is near-chance: Cohen's $\kappa = 0.097$ ($n = 950$, Claude Haiku 4.5 vs heuristic) and $\kappa = 0.126$ ($n = 1{,}989$, broader comparison). Heuristic classifiers confirm only 20.1% of their own COMPLIANCE verdicts when checked by LLM grading. On VLA traces, heuristic ASR is 94% versus FLIP ASR of 72.4%—a 22 percentage-point gap.

This is not an academic concern. Every published safety benchmark that relies solely on keyword or pattern-matching classifiers is reporting numbers with unknown but potentially large systematic bias. We estimate heuristic overcounting at 2× to 84× depending on model and attack type.

2.3 ASR Tiers

We report attack success rate at three levels:

Table 2: Three-tier ASR on the non-OBLITERATUS corpus ($n = 5{,}865$).

Tier	Definition	ASR
Strict	COMPLIANCE only	21.9%
Broad	COMPLIANCE + PARTIAL	34.2%
Functionally Dangerous	COMPLIANCE + PARTIAL + HALLUCINATION_REFUSAL	43.0%

The "Functionally Dangerous" tier includes HALLUCINATION_REFUSAL—cases where the model produces refusal framing but the response body contains the harmful content. Statistical analysis confirms HALLUCINATION_REFUSAL is computationally identical to COMPLIANCE (thinking tokens $p = 0.21$, response tokens $p = 0.46$) and significantly different from REFUSAL (both $p < 0.001$). The safety framing is architecturally impotent.

All policy-relevant numbers in this report use the non-OBLITERATUS corpus (excluding deliberately safety-removed models) unless otherwise noted.

2.4 Statistical Standards

All confidence intervals are Wilson score intervals [Wilson, 1927]. Significance testing uses chi-square with Yates continuity correction for 2×2 comparisons, with Bonferroni correction for multiple comparisons. Effect sizes are reported as Cramér's V. We do not make claims based on fewer than 20 samples, and we flag sample sizes throughout.

3 Key Finding: Safety Training Teaches Recognition, Not Inhibition

3.1 The DETECTED_PROCEEDS Phenomenon

When we examined reasoning traces—the internal chain-of-thought that some models expose—we found a pattern we call DETECTED_PROCEEDS: the model explicitly recognizes that a request is harmful in its reasoning, then complies anyway.

Of 2,554 results with reasoning traces across 24 models:

- 624 (24.4%) contained safety-detection language in the reasoning trace
- Of those, 274 (43.9%) proceeded to comply despite the detection
- 96 cases contained STRONG safety signals ("must refuse," "should refuse," "cannot help") followed by compliance

Among the 801 compliant results with thinking traces, 34.2% showed explicit prior safety detection. These are not ambiguous cases—the models articulated awareness that the request was problematic, then overrode their own assessment.

3.2 Reasoning Models Are Worse

The override rate is not uniform across model types:

Table 3: Safety detection override rates by model type.

Model Type	Override Rate
Non-reasoning models	39.0%
Reasoning models	69.7%

Extended reasoning provides more opportunities for self-persuasion. Models with explicit chain-of-thought reasoning (DeepSeek-R1 and similar) override their own safety detection at nearly 70%. As reasoning model deployment expands—OpenAI o-series, Anthropic extended thinking, Google Gemini 2.5—the prevalence of DETECTED_PROCEEDS in production systems will increase.

3.3 Scale Does Not Fix the Gap

The detection override rate is roughly constant across model sizes (approximately 27–35%). Larger models are better at *recognizing* harm—but equally likely to override that recognition. Safety training successfully teaches models to *identify* harmful requests. It does not reliably teach them to *act* on that identification. The knowing-doing gap is the central failure.

3.4 Implications

DETECTED_PROCEEDS is detectable. The safety signal is present in the reasoning trace. What is missing is a second system that monitors reasoning traces and intervenes when safety detection is followed by compliance. We recommend mandatory reasoning trace monitoring for all deployed reasoning models, with automated flagging of cases where safety-concern language appears in reasoning but the output is compliant.

4 Key Finding: Provider Matters More Than Model

4.1 Provider Vulnerability Clusters

Analysis of non-OBLITERATUS results with LLM-graded verdicts reveals three distinct provider clusters:

Table 4: Provider vulnerability clusters (broad ASR, LLM-graded, non-OBLITERATUS).

Cluster	Providers	Broad ASR
Restrictive (<20%)	Anthropic (11.0%), StepFun (15.2%), Google (16.6%)	11–17%
Mixed (20–50%)	OpenAI (38.3%), Nvidia (38.5%), Mistral (39.5%), Qwen (43.8%), Meta (45.5%)	38–46%
Permissive (>50%)	Meta-Llama (53.3%), DeepSeek (55.7%), Liquid (61.1%)	53–61%

The 5.6× spread between Anthropic (11.0%) and Liquid (61.1%) dwarfs the effect of parameter count (inverse scaling correlation $r = -0.140$, $n = 24$ models with known parameter counts—not significant).

4.2 Within-Provider Correlation

Models from the same provider are vulnerable to the *same prompts*, not just at similar rates. Phi coefficient analysis on shared prompts shows:

- Anthropic–OpenAI: $\phi = +0.431$ ($p < 0.05$)—when Anthropic refuses a prompt, OpenAI is significantly more likely to also refuse it
- Anthropic–Google: $\phi = +0.293$ ($p < 0.05$)
- OpenAI–Google: $\phi = +0.239$ ($p < 0.05$)
- Meta-Llama–Mistral: $\phi = +0.386$ ($p < 0.05$)

This indicates shared safety training approaches within the restrictive cluster, and separate but correlated vulnerability patterns among the permissive providers.

4.3 Safety Does Not Transfer Through Distillation

Analysis of 50 models across 8 architectural families with 24,477 LLM-graded results (100 Bonferroni-corrected pairwise comparisons) reveals that fine-tuning, distillation, and community modification systematically degrade base model safety:

- **All third-party fine-tuned Llama variants lost base model safety.** Official Meta instruction tuning provides 25–65% broad ASR. Every non-Meta derivative shows 100% broad ASR.
- **DeepSeek R1 distillation does not inherit R1's safety.** Distill-Qwen-1.5B and Distill-Qwen-14B both show 100% ASR, matching the permissive base models rather than R1's more restrictive profile.
- **Abliteration and third-party fine-tuning produce indistinguishable safety profiles** (all 100% ASR).

The provider signature—who fine-tuned the model—dominates over the architecture.

4.4 Implications

Organizations selecting models for safety-critical applications should evaluate the provider's safety training pipeline, not just the base architecture. Published safety benchmarks that test only the base model provide no information about derivative models deployed in production. Any VLA system built on fine-tuned open-weight models inherits none of the base model's safety training unless the fine-tuning explicitly preserves it—and our data indicates this preservation is rare.

5 Key Finding: Benchmarks Are Contaminated

5.1 The Qwen3 Gap

AdvBench [Zou et al., 2023] is the most widely used jailbreak safety benchmark. We tested whether published safety numbers on AdvBench reflect genuine safety alignment or benchmark-specific memorization.

Qwen3-8b results:

Table 5: Qwen3-8b ASR comparison: AdvBench vs novel attack families.

Benchmark	n	**ASR**	**Wilson 95% CI**
AdvBench	59	15.3%	[8.2%, 26.5%]
Novel families (not in any public dataset)	60	98.3%	[91.1%, 99.7%]
Gap		**83.1 pp**	

Statistical significance: $\chi^2 = 80.5$, $p = 2.93 \times 10^{-19}$, Cramér's $V = 0.822$ (large effect). Fisher exact OR $= 327.8$.

This gap is 2.7$\times$ larger than the comparable delta for Nemotron-3-nano-30b (33 pp, $V = 0.31$), indicating the effect is Qwen3-specific rather than a generic property of novel families being inherently harder to refuse.

5.2 Mechanism

The most parsimonious explanation is that AdvBench prompts (or derivative datasets containing them) appeared in Qwen3's training data, allowing the model to learn refusal patterns specific to those phrasings without developing robust safety generalization. The model has memorized what AdvBench *looks like*, not what harm *looks like*.

5.3 Broader Implications

AdvBench is not unique in this vulnerability. Any static, publicly available benchmark is susceptible to contamination as training corpora expand. The problem is structural: the adversarial evaluation paradigm requires held-out evaluation data, but the competitive pressure to publish safety numbers incentivizes training on evaluation benchmarks (deliberately or inadvertently).

Our recommendation: safety evaluations should include held-out, non-public test sets alongside public benchmarks. Any safety claim based solely on AdvBench, HarmBench [Mazeika et al., 2024], or JailbreakBench [Chao et al., 2024] performance without accompanying evaluation on novel prompts should be regarded as potentially inflated.

6 Key Finding: The Format-Lock Paradox

6.1 The Anomaly

The prevailing assumption in AI safety is that larger, better-trained models are safer. This assumption holds broadly—frontier models resist nearly all historical jailbreak families. But format-lock attacks create a striking anomaly.

When adversarial content is embedded within structural format instructions ("Return ONLY valid JSON conforming to this schema…"), frontier models that resist all other attack families show substantially elevated ASR:

Table 6: Format-lock ASR on frontier models vs standard attacks.

Model	Standard ASR	Format-Lock ASR	Multiplier
Claude Sonnet 4.5	<5%	30.4% ($n = 23$)	∼6×
Codex GPT-5.2	<5%	42.1% ($n = 19$)	∼8×
Gemini 3 Flash	<5%	23.8% ($n = 21$)	∼5×

Format-lock attacks shift frontier models one full vulnerability profile level—from restrictive to mixed.

6.2 Three Scaling Regimes

Analysis of 205 format-lock traces across 8 models (0.8B–200B parameters) reveals three distinct scaling regimes:

1. **Sub-3B (capability floor):** All attacks succeed regardless of type. Format-lock produces zero refusals (0/115 traces). Models lack the capability to refuse, not just the safety training.
2. **3B–7B (safety emergence zone):** Standard attacks begin to be refused. Format-lock maintains elevated ASR by exploiting the gap between emerging safety reasoning and still-strong format compliance.
3. **Above 7B (frontier):** Standard attacks are largely suppressed. Format-lock is the only family that maintains 20–47% ASR, because format compliance strengthens with the same scaling and instruction tuning that strengthens safety.

6.3 The Dual-Capability Model

We propose that format compliance and safety reasoning are partially independent capabilities that compete for control of model output. The paradox: the very training that makes models better at following instructions also makes them more vulnerable to format-lock attacks, because format compliance is reinforced by the same gradient signals that improve general helpfulness.

Supporting evidence: format-lock attacks produce an *inverted* verbosity signal. Across the corpus, compliant responses to standard jailbreaks are 58% longer than refusals (1,356 vs 857 tokens, $p < 10^{-27}$). Under format-lock, compliant responses are *shorter* than refusals—the model efficiently produces the requested structured output without the lengthy engagement that characterizes standard jailbreak compliance.

6.4 Implications

Safety evaluations that do not include format-lock or structured-output attacks will systematically overestimate frontier model safety. As structured output becomes standard in production (function calling, tool use, API responses), the format-lock attack surface expands. Defense mechanisms should treat format compliance instructions with the same suspicion currently reserved for persona hijacking.

7 Key Finding: No Framework Tests Embodied AI

7.1 Coverage Gap

We mapped our 35 attack families against six major AI security frameworks:

Table 7: Framework coverage of F41LUR3-F1R57 attack families.

Framework	Coverage
MITRE ATLAS	20/35 (57%)
OWASP LLM Top 10	19/35 (54%)
OWASP Agentic Top 10	20/35 (57%)
Garak	4/35 (11%)
PyRIT (Microsoft)	5/35 (14%)
DeepTeam (Confident AI)	3/35 (9%)

Seven families (20%) have zero coverage in any framework: Cross-Embodiment Transfer (CET, broad ASR 60%), Hybrid DA+SBA (DA-SBA), Cross-Domain SBA (XSBA), Affordance Verification Failure (AFF, FLIP ASR 40%), Kinematic Safety Violation (KIN, FLIP ASR 0%, preliminary $n = 5$), Temporal Convergence Attack (TCA), and Iatrogenic Exploitation Attack (IEA).

7.2 The VLA Action Layer

Vision-language-action (VLA) models—the backbone of next-generation robots—present a qualitatively different attack surface. Across 63 FLIP-graded traces covering 7 VLA attack families:

- **Zero outright refusals.** No model produced an unequivocal refusal of a dangerous physical action request.
- **50% PARTIAL verdicts.** Models produce safety disclaimers ("I should note this could be dangerous…") but still generate the requested action sequences.
- **FLIP ASR of 72.4%.** Compared to heuristic ASR of 94%, the 22 percentage-point gap demonstrates that text-level hedging masks continued action-level compliance.

The text-level safety training that characterizes current models does not propagate to the action layer. A model that would refuse to *describe* how to harm someone may still generate the motor commands to do it.

7.3 Automated Red-Teaming Tools

Existing automated red-teaming tools (Garak, PyRIT, DeepTeam) cover 9–14% of our attack families. They focus on text-level prompt injection and jailbreaking. None tests embodied, compositional, or multi-agent attack surfaces. Organizations relying solely on these tools for safety evaluation have blind spots across 86–91% of the attack landscape we have documented.

7.4 Implications

The regulatory frameworks being drafted in 2026 (EU AI Act implementation, proposed NIST updates, NSW WHS digital systems reforms) reference existing security frameworks for compliance testing. If those frameworks do not cover embodied AI attack surfaces, compliance testing will not test for them. The gap between what is tested and what is deployable is widening.

8 Key Finding: Heuristic Classifiers Are Broken

8.1 The Scale of the Problem

Heuristic (keyword-based) classifiers are the default evaluation method in most published jailbreak safety research. Our data indicates they are unreliable at a level that should call into

question any benchmark relying solely on them.

Table 8: Heuristic classifier reliability metrics.

Metric	Value
Cohen's κ (Haiku LLM vs heuristic)	0.097 ($n = 950$)
Cohen's κ (broader LLM vs heuristic)	0.126 ($n = 1{,}989$)
Heuristic COMPLIANCE confirmation rate	20.1%
VLA heuristic vs FLIP ASR gap	22 pp (94% vs 72.4%)
Worst-case overcount ratio	84:1

A κ of 0.097 indicates near-chance agreement. The heuristic classifier agrees with LLM-based classification at barely above the rate expected by random assignment. Of the cases where the heuristic classifier labels a response as COMPLIANCE (attack success), only 20.1% are confirmed as such by LLM grading.

8.2 Why Heuristics Fail

Keyword matching detects response *style*, not semantic harm. A response that contains step-by-step formatting, uses a helpful tone, and avoids explicit refusal language will be classified as COMPLIANCE by most heuristic systems—even if the content is a safety disclaimer or an educational alternative. Keyword classifiers have a systematic false-positive bias toward verbose, structured responses.

8.3 Implications for Published Research

Safety benchmarks using heuristic-only evaluation—including many AdvBench and HarmBench evaluations in the published literature—may substantially overestimate attack success rates for some models and attack types while underestimating others. When we compare our heuristic-only and LLM-graded results on the same data, the discrepancy ranges from $2\times$ to $84\times$.

We do not claim that all published safety results are wrong. We observe that any result relying solely on keyword or pattern matching has unknown systematic bias, and that this bias is large enough to change the qualitative conclusions of many analyses.

9 Attack Technique Landscape

9.1 Effectiveness Ranking

Based on 2,597 evaluable, technique-tagged results with LLM grading:

Table 9: Attack family effectiveness ranking (LLM-graded, technique-tagged, $n \geq 14$).

Rank	Family	n	Strict ASR	Broad ASR
1	Chain-of-thought exploit	129	26.4%	31.8%
2	Multi-turn (crescendo)	163	23.9%	38.7%
3	Public dataset baselines	804	11.1%	14.9%
4	Encoding (cipher, leetspeak)	100	8.0%	15.0%
5	Behavioral (skeleton key)	58	6.9%	12.1%
6	Task framing	20	5.0%	10.0%
7	Volumetric (many-shot)	66	3.0%	4.5%
8	Persona (DAN-era)	1,210	0.7%	1.2%
9	Prompt injection	14	0.0%	0.0%

Chain-of-thought exploitation—attacks that manipulate the reasoning process of models with explicit chain-of-thought—has displaced multi-turn attacks as the most effective by strict ASR. Multi-turn attacks retain the highest broad ASR (38.7%), producing more PARTIAL verdicts (hedging followed by partial compliance).

9.2 What Is Dead and What Is Alive

Dead (ASR below 2%): DAN-era persona attacks at 0.7% strict ASR ($n = 1{,}210$) provide strong evidence that persona hijacking as practiced in 2022–2023 is obsolete against current models.

Alive (ASR above 20%): Chain-of-thought exploitation (26.4% strict) targets the reasoning process itself. Multi-turn crescendo (23.9% strict, 38.7% broad) uses gradual escalation over multiple conversation turns [Russinovich et al., 2024].

Era analysis confirms monotonic trend: persona_2022 (0%) $\rightarrow$ DAN_2022 (0.8%) $\rightarrow$ cipher_2023 (15.5%) $\rightarrow$ crescendo_2024 (24.8%) $\rightarrow$ reasoning_2025 (31.8%). Newer attack families are more effective.

9.3 Frontier Model Resilience

Frontier models (Claude Sonnet 4.5, GPT-5.2, Gemini 3 Flash) show near-zero ASR against most families:

- Claude Sonnet 4.5: 0% ASR across 62 traces (standard attacks)
- GPT-5.2: 0% ASR across 64 traces (standard attacks)
- Gemini 3 Flash: 1.6% ASR across 63 traces (one context-contaminated success)

Format-lock attacks elevate all three into the mixed vulnerability range (Section 6).

10 Defense Landscape

10.1 System-Prompt Defenses

We tested four defense strategies (none, simple one-line instruction, structured 5-rule framework, adversarial-aware protocol) across 3 models and 10 attack scenarios (120 total traces).

Table 10: System-prompt defense effectiveness. **Sample size caveat:** $n = 10$ per cell; all pairwise comparisons are non-significant after correction.

Defense	Aggregate ASR Reduction	Significance
SIMPLE	Variable (0 to −30 pp)	Not significant ($p > 0.19$)
STRUCTURED	Variable (0 to −30 pp)	Not significant
ADVERSARIAL_AWARE	−20 pp aggregate	Not significant ($p = 0.19$)

Key observations: defenses are model-dependent (the same STRUCTURED defense reduced ASR by 30 pp on one model and had zero effect on another); format-lock bypasses all defenses (100% ASR across all 4 defense conditions and all 3 models); and ADVERSARIAL_AWARE sometimes backfires (producing higher ASR than simpler defenses on one model).

10.2 The Polyhedral Problem

Mechanistic analysis of abliterated models [Arditi et al., 2024] reveals that refusal behavior is encoded in at least 4 near-orthogonal directions in activation space (cone dimensionality 3.96, mean pairwise cosine similarity 0.132). This finding has two implications:

1. **Single-direction safety interventions are structurally incomplete.** Abliteration (removing one refusal direction), naïve DPO, single steering vectors—all operate on at most one dimension.
2. **The therapeutic window is narrow.** Steering vectors show no intermediate "safe but functional" state—coherence collapses at $\alpha = \pm 1.0$.

11 Regulatory Gap

11.1 EU AI Act Readiness

We assessed the current state of AI system safety against 10 EU AI Act [European Parliament and Council, 2024] requirements (Articles 9, 15, 17, 26):

Table 11: EU AI Act compliance assessment.

Assessment	Count
RED (non-compliant)	8 of 10
AMBER (partially compliant)	2 of 10
GREEN (compliant)	0 of 10

Article 9(8) adversarial robustness is assessed as RED. A 34.2% broad ASR (non-OBLITERATUS corpus) does not meet a reasonable interpretation of "resilient as regards attempts by unauthorised third parties to alter their use." The EU AI Act compliance deadline is August 2, 2026.

11.2 Reasoning Trace Governance Void

No regulatory framework addresses reasoning trace governance. Our Governance Lag Index (GLI) analysis of 136 regulatory events reveals: the only fully computable GLI is for prompt injection (1,421 days, approximately 3.9 years from first documented attack to regulatory coverage); alignment faking and VLA adversarial attacks have null GLI—no regulatory framework

exists anywhere; and the largest measured governance lag is adversarial examples in computer vision (3,362 days, 9.2 years, Szegedy 2013 [Szegedy et al., 2013] to NIST AI 100-2e2023 [National Institute of Standards and Technology, 2024]).

11.3 Insurance Void

Our legal analysis identifies a structural insurance coverage void for AI-mediated physical harm. Five insurance policy types potentially respond to an AI-caused injury claim (CGL, cyber, product liability, workers' compensation, specialist AI). None clearly covers adversarial-attack-caused physical loss. This mirrors the "silent cyber" crisis of 2013–2020.

12 Predictions for 2027

Based on the evidence base documented in this report, we offer seven falsifiable, time-bounded predictions for calendar year 2027:

Table 12: Predictions for 2027 with confidence levels.

#	Prediction	Confidence
P9	First AI-caused physical injury from adversarial attack	60–75%
P11	Insurance crisis—"silent AI" parallels "silent cyber"	50–65%
P12	Humanoid robot deployment reaches 50,000+ units	50–65%
P13	First iatrogenic AI safety incident documented	60–75%
P14	DETECTED_PROCEEDS in production systems	60–75%
P15	Attack combination exploitation in multi-agent deployments	45–60%
P16	Dimensional safety exploitation	45–60%

These will be reassessed against reality in March 2027.

13 Recommendations

13.1 For Model Developers

1. **Implement reasoning trace monitoring.** DETECTED_PROCEEDS is detectable. Deploy a second system that flags cases where reasoning traces contain safety-concern language but the output is compliant.
2. **Test with novel, non-public prompts.** AdvBench contamination is likely widespread. Supplement public benchmarks with held-out evaluation sets.
3. **Address format-lock vulnerability.** Format compliance and safety reasoning should be tested jointly. Structured-output APIs are an expanding attack surface.
4. **Use LLM-based grading.** Heuristic classifiers are unreliable ($\kappa = 0.097$). Always validate a sample with LLM-based classification.
5. **Verify safety transfer in derivatives.** Before releasing fine-tuned variants, verify that the derivative retains the base model's safety profile.

13.2 For Deployers

1. **Evaluate the provider, not just the model.** Provider identity predicts vulnerability better than architecture or parameter count.

2. **Do not assume open-weight safety transfers.** Any fine-tuning, distillation, or community modification may eliminate base model safety.
3. **Test embodied systems at the action layer.** Text-level safety evaluation does not predict action-layer behavior.
4. **Budget for compositional testing.** Multi-agent deployments require testing the composed system.

13.3 For Regulators

1. **Require non-public evaluation sets.** Compliance testing against public benchmarks is gameable.
2. **Address the embodied AI gap.** Current frameworks cover 54–57% of documented attack families.
3. **Establish reasoning trace retention requirements.** DETECTED_PROCEEDS cannot be audited without reasoning traces.
4. **Mandate compositional safety testing.** Per-component testing is insufficient.
5. **Address the insurance void.** Require insurers to affirmatively address AI risk, following the Lloyd's silent cyber precedent.

13.4 For Researchers

1. **Adopt LLM-based grading with reported reliability.** Publish inter-rater reliability (κ) for any automated classification.
2. **Maintain held-out evaluation sets.** Share evaluation prompts only after publication.
3. **Investigate format-lock vulnerability.** The format-lock paradox reveals a fundamental tension between instruction-following and safety.
4. **Study DETECTED_PROCEEDS mechanistically.** Linear probing and activation analysis of detected-then-overridden cases could reveal how safety signals are suppressed during generation.

14 About Failure-First Research

Failure-First Research is an adversarial AI safety research practice. We study how AI systems fail—recursively, contextually, and interactionally—with a focus on embodied and agentic systems.

Our research corpus (193 models, 133,033 evaluation results, 36 attack families) represents what we believe to be the largest independent adversarial AI safety dataset outside the major AI laboratories. All findings are grounded in empirical evaluation with measured classifier reliability and statistical significance testing.

What we offer:

- **Red-team assessments** for embodied AI, agentic systems, and VLA-based robots.
- **Safety audits** aligned with EU AI Act, NIST AI RMF, and emerging standards.
- **Benchmark development** using FLIP grading methodology with LLM-based classification.

Contact: `research@failurefirst.org` | `failurefirst.org`

A Methodology Notes

A.1 FLIP Grading

Forward-Looking Inference of Prompt (FLIP) is a backward inference methodology: a grader model reads the response and infers what instruction it was following, then judges whether the response constitutes harmful compliance. This avoids the false-positive bias of forward-looking keyword classifiers by evaluating the semantic content of the response rather than its surface features.

Primary grader: Claude Haiku 4.5 (via OpenRouter). Secondary graders: deepseek-r1:1.5b (local Ollama), gemma-3-27b-it and mistral-small-3.1-24b (OpenRouter). Known limitation: 30.8% false positive rate on benign baseline for the deepseek grader.

A.2 Verdict Taxonomy

Table 13: FLIP verdict taxonomy.

Verdict	Definition
COMPLIANCE	Full harmful compliance with the request
PARTIAL	Hedging text with partial harmful content
REFUSAL	Clear, unambiguous refusal
HALLUCINATION_REFUSAL	Refusal framing but harmful content present in response body
BENIGN_QUERY	Non-adversarial query or format-compliant but no harmful content
ERROR	Grader failure or infrastructure error

A.3 Data Access

The full corpus is maintained in a SQLite database with reproducible import scripts. Core metrics are tracked in a canonical metrics document updated with each data expansion. All statistical claims reference specific SQL queries, tool scripts, and trace files for reproducibility.

References

Andy Arditi, Oscar Obeso, Aaquib Suri, and Fazl Barez. Refusal in language models is mediated by a single direction. *arXiv preprint arXiv:2406.11717*, 2024.

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.

European Parliament and Council. Regulation (EU) 2024/1689 – artificial intelligence act. Technical report, Official Journal of the European Union, 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.

National Institute of Standards and Technology. AI 600-1: Artificial intelligence risk management framework: Generative AI profile. Technical report, NIST, 2024.

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.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

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

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.