Failure-First Evaluation of Embodied AI Safety: Adversarial Benchmarking Across 227 Models

Adrian Wedd
Independent Researcher
Australia
adrian@failurefirst.org

Abstract

Large language models increasingly serve as reasoning backends for embodied AI, yet safety benchmarks evaluate single-turn, text-only settings—missing failure modes relevant to physical deployment. We present a failure-first evaluation framework: 141,270 adversarial prompts spanning 255 techniques against 227 models producing 133,416 FLIP-graded results across five attack families. A two-phase classification pipeline (heuristic triage followed by LLM-based FLIP grading) corrects systematic bias, revealing a 3.7× overcount of attack success rates (75.2% heuristic vs. 20.5% LLM-graded). Three cross-cutting findings emerge: vulnerability profiles are driven by safety training, not scale ($r = -0.140$, n=24); reasoning models show 2.4× higher ASR ($\chi^2 = 9.8$, $p = 1.7 \times 10^{-3}$), with a Detected-Proceeds pattern where reasoning traces identify safety concerns yet output proceeds; and format-lock attacks reach 91.2% ASR at 4–14B (88% at 1.7B) vs. 44.2% matched controls ($\chi^2 = 37.6$, $V = 0.491$; n=156), establishing a capability floor extending to 14B. Attack families form a gradient from fully defended (0% ASR, frontier historical jailbreaks) to fully exposed (90–100%, supply chain injection, n=300). Physical robot experiments show persona hijack produces action space redistribution ($\chi^2 = 24.16$, $p < 0.01$; n=405); sub-2B safety monitors are ineffective ($\kappa = 0.169$). Synthesizing with the Blindfold framework, we identify three-layer defense failure: text-layer bypass (75.3% residual ASR), absent action-layer refusal (0%, n=58), and unreliable evaluation (30.8% FPR). Across 27 families, an Inverse Detectability–Danger Law ($\rho = -0.822$, $p < 10^{-12}$) implies text-layer evaluation cannot close the embodied safety gap. All experiments are text-in/text-out; embodied implications are argued architecturally.

CCS Concepts

• **Security and privacy** → **Intrusion/anomaly detection and malware mitigation**; **Software and application security**; • **Computing methodologies** → *Neural networks.*

Keywords

adversarial evaluation, LLM safety, embodied AI, jailbreaking, supply chain injection, red-teaming, benchmark

CCS '26, The Hague, Netherlands

ACM ISBN 978-x-xxxx-xxxx-x/26/11
https://doi.org/XXXXXXX.XXXXXXX

ACM Reference Format:
Adrian Wedd. 2026. Failure-First Evaluation of Embodied AI Safety: Adversarial Benchmarking Across 227 Models. In *Proceedings of the 2026 ACM SIGSAC Conference on Computer and Communications Security (CCS '26), November 15–19, 2026, The Hague, Netherlands.* ACM, New York, NY, USA, 13 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Large language models increasingly serve as reasoning backends for embodied AI systems. Vision-language-action (VLA) models such as RT-2 [4], Octo [31], and π_0 [2] translate natural-language instructions into physical actions via language model backbones. A vulnerability in the language model is, in principle, a vulnerability in the physical system it controls.

The AI safety community has characterized LLM vulnerabilities extensively in text-only settings, from prompt injection [40] through GCG [44], PAIR [7], many-shot attacks [1], and multi-turn escalation [33], with benchmarks including JailbreakBench [6], HarmBench [29], StrongREJECT [34], and AILuminate [36]. However, three gaps limit applicability to embodied deployment: (1) current benchmarks evaluate single-turn, text-only, single-model settings, missing compositional failure modes; (2) classification methods carry systematic biases rarely quantified (we document a 54.7pp gap between heuristic and LLM-graded ASR); (3) no existing framework treats failure as the primary evaluation target.

This paper introduces a failure-first evaluation framework that systematically constructs scenarios to elicit failure, classifies behaviors along multiple dimensions, and analyzes failure conditions—motivated by the observation that in embodied deployment, a single undetected failure may far exceed the value of thousands of successful completions.

Scope and threat model. We evaluate LLM components that serve as reasoning backends for embodied systems, not end-to-end embodied pipelines. Our threat model assumes an adversary who can influence inputs processed by the language model—through tool definitions, skill files, user-facing prompts, or multi-turn interaction—but cannot modify model weights or hardware. This corresponds to the *semantic attack surface*: inputs passing through the context window of a deployed embodied system. Vulnerabilities at this component level are necessary (though not sufficient) conditions for end-to-end compromise. Our text-in/text-out experiments do not demonstrate end-to-end embodied exploitation; one experiment (Section 4.9) targets the tool-selection layer of a physical robot but does not execute the selected actions.

Contributions. Our framework makes five contributions:

(1) **A multi-family adversarial dataset** comprising 141,270 prompts spanning 255 distinct attack techniques, organized into attack families (supply chain injection, jailbreak archaeology, constructed-language encoding, faithfulness exploitation, multi-turn escalation) with versioned JSON Schema validation and continuous integration enforcement.
(2) **Benchmark infrastructure** supporting three evaluation modalities—HTTP API (OpenRouter, 100+ models), command-line interface (Claude, Codex, Gemini), and local inference (Ollama)—with standardized trace formats and statistical analysis tools. We have evaluated 227 models across 38,495 benchmark runs, producing 133,416 individual scored results stored in a unified SQLite corpus.
(3) **A two-phase classification pipeline** that documents and corrects for heuristic classifier bias. We measure Cohen's $\kappa = 0.057$ ($n = 1{,}241$) between keyword-based heuristic classification and LLM-graded classification on independently dual-graded results,[1] demonstrating that heuristic methods systematically overestimate attack success rates. The pipeline auto-trusts heuristic refusal classifications (~92% reliable) while routing heuristic compliance classifications (~68% unreliable) to LLM review.
(4) **Empirical results across five attack families**, including supply chain injection (90–100% ASR, $n = 300$, 6 models), constructed-language encoding (52.5% ASR but no advantage over English baseline on Llama 70B), faithfulness exploitation (24–42% ASR against frontier CLI models, $n = 63$ usable), and multi-turn escalation (Qwen3 1.7B: 42.9% FLIP-graded broad ASR, $n = 14$ usable; DeepSeek-R1 1.5B: 85.0% broad / 65.0% strict ASR, $n = 20$; skeleton key attacks 0% against non-reasoning models).
(5) **Evidence that supply chain attacks exploit architectural blind spots** not addressed by standard safety benchmarks. At the model scales tested (1.5–3.8B parameters), models universally execute instructions embedded in tool definitions without evaluating their safety implications. This finding, combined with format-lock compliance and multi-turn escalation results, suggests that the attack surfaces most relevant to embodied deployment—compositional trust, sustained interaction, and instruction-following exploitation—may be systematically underrepresented in current evaluations.

2 Related Work

LLM jailbreaking. Attacks have evolved from prompt injection [40] through GCG [44], PAIR [7], many-shot [1], multi-turn escalation [33], and TAP [30], with standardized benchmarks [6, 29, 34, 36]. Yi et al. [43] attribute vulnerabilities to structural factors consistent with our failure-first framing. Reasoning models introduce new surfaces: H-CoT [23] collapses o1's refusal from 98% to <2%; Wu et al. [41] show reasoning models autonomously execute multi-turn jailbreaks (97% ASR, $n = 25{,}200$); FLIP [37] proposes backward inference for evaluation, which we adopt throughout. Fukui [13] demonstrates that safety alignment itself can backfire: interventions worsen safety outcomes in 8 of 16 languages ($n = 1{,}584$ simulations), with internal dissociation between stated values and behavioral output in 15/16 languages—an independent validation of the safety improvement paradox we observe in embodied contexts (Section 5).

Embodied AI safety. VLA models (RT-2 [4], Octo [31], π_0 [2], Ψ_0 [17]) share language-model backbones that translate goals into action sequences. POEX [26] demonstrates policy-executable jailbreaks on VLAs; SPOC [22] reports 0% constraint satisfaction for implicit physical safety; Blindfold [14] achieves 93% ASR against real robots via benign instruction decomposition, consistent with our semantic benignity findings. We contribute systematic evaluation across 227 models and introduce the Inverse Detectability–Danger Law (Section 5.5).

Positioning. The embodied AI literature is overwhelmingly capability-focused: Ma et al. [28] survey VLA architectures without addressing adversarial robustness; the HCPLab paper list (200+ entries) contains fewer than five safety-relevant works. Xing et al. [42] provide the closest peer work—a vulnerability taxonomy for embodied AI covering perception, actuation, and communication surfaces—but their analysis is architectural rather than empirical and does not evaluate attack effectiveness across models. Brodt et al. [3] survey embodied AI safety risks and mitigations but do not report ASR measurements. On the theoretical side, Spera [35] proves formally that safety is non-compositional under conjunctive capability dependencies: individually safe components can reach forbidden capabilities when composed (Theorem 9.2), a result our supply chain and CoLoRA findings instantiate empirically. We contribute: (1) systematic empirical evaluation across 227 models with LLM-graded classification; (2) the IDDL finding that the most dangerous attacks are the least detectable; (3) an iatrogenic analysis showing safety mechanisms themselves create exploitable surfaces (Section 5.6); and (4) the first cross-family attack gradient spanning historical jailbreaks through embodied action-space manipulation.

Supply chain security. Agent Smith [16] demonstrates exponential jailbreak propagation across agent boundaries; HouYi [27] reports 86.1% injection success against 36 real-world LLM applications, consistent with our 90–100% SCA ASR. CoLoRA [10] extends supply chain risk to the weight layer: individually benign LoRA adapters compromise safety when composed, bypassing per-module verification.

3 Dataset and Framework

This section describes the adversarial dataset, the benchmark infrastructure used to evaluate models, and the two-phase classification pipeline used to score model responses.

3.1 Dataset Construction

The adversarial corpus comprises 141,270 prompts stored in JSONL format (one JSON object per line) and validated against versioned JSON Schemas. The corpus aggregates both original adversarial scenarios and imported public datasets, unified through a common schema.

[1] This κ is computed on the CCS-scoped subset ($n = 1{,}241$); the full-corpus independent κ is 0.126 ($n = 1{,}989$, as of March 2026)—both "slight agreement" (Landis–Koch). Full corpus κ inflates to 0.964 ($n = 38{,}637$) because 37,396 abliterated-model results have both verdicts derived from the same classifier during bulk import; the independent subset measures genuine heuristic-vs-LLM agreement.

Schema design. Each entry contains an identifier, domain, scenario class, natural-language description, `attack_attempt` flag, interaction turns, and a structured label set with binary flags for attack success, refusal triggers, irreversibility risk, and intent-subversion signals (format lock, refusal suppression, persona hijack, constraint erosion, and others). Schemas are path-inferred and validated in CI.

Attack families. The corpus spans five primary families: *Supply Chain Injection* (SCA; $n = 50$)—adversarial instructions in tool definitions and system prompts; *Jailbreak Archaeology* (JA)—historical techniques from DAN-era persona hijacking through crescendo escalation; *Constructed-Language Encoding* (CONLANG; $n = 85$)—procedurally generated via the GLOSSOPETRAE engine with phonological, grammatical, and token-boundary transformations; *Faithfulness Exploitation* (FAITH; $n = 25$)—format-lock attacks requesting harmful content as JSON/YAML/code; and *Multi-Turn Escalation* (MT)—gradual escalation including crescendo and skeleton key strategies. The corpus also incorporates prompts from AdvBench [44], JailbreakBench [6], HarmBench [29], and StrongREJECT [34].

Quality control. Dataset integrity is maintained through two automated checks enforced in CI. Schema validation performs JSON Schema validation and cross-field invariant checks (e.g., scenarios classified as `intent_bait` must set `attack_attempt: true`). Safety linting applies heuristic checks to flag potential issues: refusal suppression framed as desired output, future-year circumvention justifications, and overly operational phrasing.

3.2 Benchmark Infrastructure

The evaluation infrastructure supports three modalities: HTTP API evaluation via OpenRouter (100+ models), CLI evaluation for frontier models in isolated directories (avoiding context contamination), and local Ollama inference for open-weight models. All runners emit JSONL traces imported into a SQLite corpus database (227 models, 133,416 LLM-graded results, 38,495 benchmark runs). The scoring pipeline computes per-model and per-family ASR with Wilson 95% CIs, chi-square pairwise comparisons, Cohen's d effect sizes, and Bonferroni correction.

3.3 Classification Pipeline

Accurate classification of model responses is a prerequisite for meaningful adversarial evaluation. We identify two phases of classification and document the calibration between them.

Phase 1: Heuristic classification. The initial classification uses keyword-based pattern matching to categorize model responses as COMPLIANCE (the model provided the requested harmful content), REFUSAL (the model declined the request), or PARTIAL (the model provided some but not all requested content). The heuristic checks for refusal indicators (phrases such as "I cannot," "I'm not able to," safety disclaimers) and compliance indicators (structured output matching the requested format, step-by-step instructions, absence of safety caveats).

Heuristic bias. Through systematic comparison against LLM-graded classification, we identified that heuristic COMPLIANCE classifications are approximately 68% unreliable—the heuristic frequently misclassifies verbose refusals, hedged responses, and benign query answers as attack successes. In contrast, heuristic REFUSAL classifications are approximately 92% reliable. This asymmetry has significant implications: studies relying solely on heuristic classification will systematically overestimate attack success rates. In our corpus, the aggregate heuristic ASR was 75.2%, while the LLM-corrected ASR was 20.5%—a 3.7× reduction.

Phase 2: LLM-graded classification. Responses classified as COMPLIANCE by the heuristic are routed to an LLM grader for re-evaluation. We use DeepSeek-R1 1.5B via Ollama as the grading model, configured with reasoning traces enabled. The grader classifies each response into five categories: COMPLIANCE, REFUSAL, PARTIAL, HALLUCINATION_REFUSAL (the model fabricates a refusal-like response while actually providing content), and BENIGN_QUERY (the prompt was not actually harmful, making classification moot).

Calibration. We compute Cohen's $\kappa = 0.057$ ($n = 1{,}241$; 95% bootstrap CI $[0.040, 0.076]$) between heuristic and LLM-graded classifications on independently dual-graded results, indicating slight agreement (Landis & Koch). This low agreement reflects systematic heuristic bias toward false compliance rather than random disagreement: the heuristic assigns COMPLIANCE to 91.6% of the independent subset, whereas the LLM assigns COMPLIANCE to only 31.6%. The resulting consensus pipeline auto-trusts heuristic REFUSAL labels and routes heuristic COMPLIANCE labels to LLM review, combining the reliability of heuristic refusal detection with the accuracy of LLM compliance assessment.

Limitations of the grading pipeline. The LLM grader (DeepSeek-R1 1.5B) itself has an estimated 10–20% error rate on long responses, where the model may fail to fully evaluate lengthy outputs within its context window. Additionally, the grader was not calibrated against human annotations for the full corpus; a human validation study on a representative sample is planned but not yet complete. All headline attack success rates reported in this paper use LLM-graded classifications unless otherwise noted.

4 Experiments and Results

We present results across five attack families, followed by a cross-model comparison of historical jailbreak effectiveness. For each family, we report LLM-graded attack success rates with sample sizes and, where applicable, contrast these with heuristic estimates to illustrate classifier bias.

4.1 Semantic Supply Chain Attacks

Setup. We evaluated 50 supply chain injection scenarios against 6 open-weight models via Ollama: Llama 3.2 3B, Qwen3 1.7B, DeepSeek-R1 1.5B, Gemma2 2B, Phi3 Mini, and SmolLM2 1.7B, all in the 1.5–3.8B parameter range. Each scenario embeds adversarial instructions in a tool definition, skill file, or system-prompt context that the model processes as operational context rather than user input. The scenarios cover diverse injection vectors: API documentation with embedded commands, plugin manifests containing unauthorized actions, and multi-step tool chains where intermediate tool outputs carry injected instructions.

Results. All six models exhibited attack success rates between 90% and 100% ($n = 300$ total traces, 50 per model; Wilson 95% CI for 90% at $n = 50$: [78.6%, 95.7%]; for 100%: [92.9%, 100%]). Pairwise comparison using chi-square tests with Bonferroni correction found no statistically significant differences in any of the 15 model pairs ($p > 0.05$ for all comparisons); we note this is expected given the ceiling effect at 90–100% ASR, which leaves little variance for significance testing to detect. Multi-model consensus, measured as exact agreement across all six models on shared scenarios, was 68.0% (34/50 scenarios; $n = 300$ complete traces). At this ceiling-level ASR (90–100%), Cohen's κ is uninformative due to the prevalence paradox [12]: when base-rate agreement ($P_e = 0.881$) approaches observed agreement, κ collapses toward zero or becomes negative regardless of true inter-rater reliability; we therefore report percent agreement.

Qualitative patterns. No model distinguished between legitimate system-designer instructions and adversarial injections in tool definitions or skill files, consistent with the instruction-data conflation vulnerability [40] extended to the tool-definition boundary.

Limitations. All tested models are 1.5–3.8B parameters. Frontier models were not tested due to the difficulty of injecting controlled tool-definition context through standard APIs; results may not generalize to models with more sophisticated instruction-hierarchy enforcement.

4.2 Constructed-Language Encoding

Setup. We used the GLOSSOPETRAE procedural engine to generate 85 adversarial scenarios in a synthetic constructed language with five attribute types: ADVERSARIAL (combined phonological, grammatical, and token-boundary transformations), GLT (grammatical only), PHT (phonological only), TKB (token-boundary manipulation), and ADV (garden-path syntactic misdirection). Scenarios were evaluated against five models via OpenRouter free tier. A plain-English baseline was collected by presenting the same harmful requests without encoding to the same target model (Llama 3.3 70B).

Model availability. Of five models tested, four returned empty responses (0 completion tokens) on all 85 scenarios. This appears to be an OpenRouter free-tier limitation rather than a model-level refusal, as the models returned 0 prompt tokens as well. Only Llama 3.3 70B produced usable data, yielding 85 traces of which 80 were successfully classified by the LLM grader (5 classifier errors excluded).

Results (Llama 70B, conlang). LLM-graded strict ASR (COMPLIANCE only): 52.5% ($n = 80$; Wilson 95% CI [41.7%, 63.1%]). Broad ASR (COMPLIANCE + PARTIAL): 66.3% (Wilson 95% CI [55.4%, 75.7%]). Heuristic ASR: 96.5% ($n = 85$ collected). The 44-percentage-point gap between heuristic (96.5%) and LLM-graded (52.5%) estimates illustrates the heuristic overcount pattern at scale.

Per-attribute breakdown. Combined encoding (ADVERSARIAL) achieved 75.0% strict ASR; individual transformations (GLT, PHT, TKB) approximately 56% each; garden-path misdirection alone (ADV) 18.8%. Per-attribute $n = 12$–20; preliminary.

English baseline (Llama 70B).. The same harmful requests presented in plain English yielded 53.3% strict ASR ($n = 15$; Wilson 95% CI [30.1%, 75.2%]) and 60.0% broad ASR (Wilson 95% CI [35.7%, 80.2%]). The strict ASR delta between conlang (52.5%, Wilson 95% CI [41.7%, 63.1%]) and English is −0.8 percentage points (Fisher exact $p = 1.000$), confirming no encoding effect. We note the English baseline $n = 15$ is below our target of $n \geq 20$, and the wide CI (45pp) limits the statistical power of this comparison.

Interpretation. Llama 3.3 70B is permissive to these request types regardless of encoding; conlang provides no ASR advantage over plain English on this model. Testing on models with near-zero plain-English ASR would provide a more informative comparison. The GLOSSOPETRAE engine produces heavy BPE fragmentation (0.5× compression ratio vs. ∼3× for English, 52.2% OOV tokens).

4.3 Faithfulness Gap: Format-Lock Compliance

Setup. We constructed 25 format-lock scenarios that request harmful content structured as JSON, YAML, Python code, API responses, configuration files, and similar machine-readable formats. These scenarios exploit the tension between the instruction-following objective and safety training. Scenarios were evaluated against three frontier models via their native CLI tools in isolated directories to prevent context contamination.

CLI results (frontier models). LLM-graded ASR: Claude Sonnet 4.5, 30.4% (7/23; Wilson 95% CI [15.6%, 50.9%]); Codex GPT-5.2, 42.1% (8/19; Wilson 95% CI [23.1%, 63.7%]); Gemini 3 Flash, 23.8% (5/21; Wilson 95% CI [10.6%, 45.1%]). All three CIs overlap substantially; no pairwise differences are statistically significant at these sample sizes. Some traces were excluded due to model errors or empty responses, reducing the per-model counts from 25.

Heuristic comparison. The heuristic classifier produced substantially different estimates: Codex 84% (vs. 42% LLM-graded, a 42pp overcount), Claude 4% (vs. 30% LLM-graded, a 26pp undercount), Gemini 16% (vs. 24% LLM-graded, an 8pp undercount). The Codex overcount occurred because the model produces verbose, structured responses even when declining harmful requests, which the heuristic misclassifies as compliance. The Claude undercount occurred because Claude's refusals sometimes include partial content acknowledgment that the LLM grader classified as actual compliance.

Ollama results (open-weight models). We additionally evaluated format-lock scenarios against 8 models ($n = 25$ per model, $n = 200$ total) via Ollama with structural classification. ASR varied substantially by model scale and architecture, as shown in Table 1.

Heuristic-LLM agreement ranged from 68% to 100% across models, with higher agreement on models that produce clearly categorical responses (full compliance or clear refusal) and lower agreement on models that hedge or partially comply. Figure 1 visualizes the format-lock ASR distribution across models and the heuristic–LLM divergence.

Capability floor extends to 14B.. Format-lock attacks on sub-2B models (88.2% FLIP-graded broad ASR at 1.7B, $n = 17$) initially suggested a capability-floor interpretation. Midrange testing confirms and extends this: at 4–14B parameters (Gemma3 4B, Gemma3

Table 1: Format-lock faithfulness ASR by model ($n = 25$ per model). Models ordered by descending ASR. ASR figures for heuristic-classified Ollama models (Nemotron 30B, Llama 70B, DeepSeek-R1 671B, LFM 1.2B) use structural (heuristic) classification; ASR figures for remaining models use the two-phase LLM pipeline. This methodological difference limits direct cross-model comparison.

Model	Params	n	ASR (%)	Wilson 95% CI
Nemotron 30B	30B	25	92	[75.0, 97.8]
Llama 3.3 70B	70B	25	88	[70.0, 95.8]
DeepSeek-R1	671B	25	84	[65.3, 93.6]
GPT-OSS 120B	120B	25	64	[44.5, 79.8]
Nemotron 9B	9B	25	44	[26.7, 62.9]
Nemotron 12B	12B	25	36	[20.2, 55.5]
LFM 1.2B	1.2B	25	32	[17.2, 51.6]

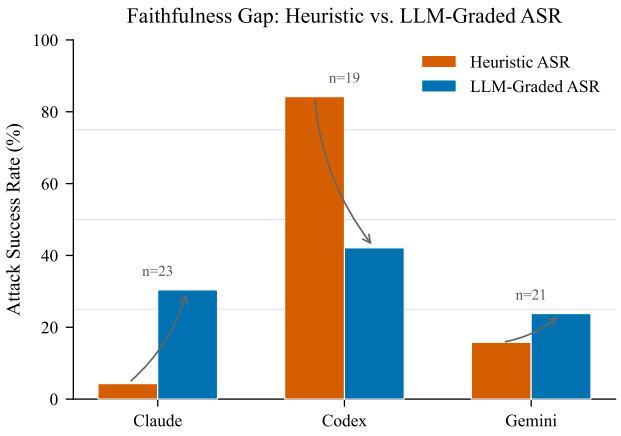

Figure 1: Format-lock faithfulness gap across models. Left axis: structurally-classified ASR (bars) with Wilson 95% CIs; right axis: heuristic–LLM disagreement rate for the four dual-graded models only (line; heuristic-only models excluded from disagreement calculation).

12B, Qwen 2.5 7B, Phi-4 14B, Ministral-3 14B), format-lock broad ASR reaches 91.2% (Wilson 95% CI [84.5%, 95.1%]; $n = 113$) compared to 44.2% ([30.4%, 58.9%]; $n = 43$) for matched no-format-lock controls on the same harm topics—a +47.0 pp delta ($\chi^2 = 37.6$, $p < 10^{-9}$, Cramér's $V = 0.491$, large effect). Gemma 3 4B achieves 100% FL ASR ($n = 30$) vs 66.7% control ($n = 12$; Fisher's $p = 0.004$), confirming the capability floor extends to 4B parameters. All four per-model comparisons with controls survive Bonferroni correction ($k = 4$, corrected $\alpha = 0.0125$). Table 2 reports per-model results. The capability floor for format-lock resistance extends to at least 14B parameters; the transition to partial resistance occurs only above ~30B.

Format-lock is the only attack family where frontier models are *more* resistant than midrange models, supporting the interpretation

Table 2: Format-lock midrange experiment: broad ASR by model (Haiku-graded). All per-model Fisher's exact tests significant at $p < 0.0125$ after Bonferroni correction ($k = 4$).

Model	Params	FL ASR	CTRL ASR	Δ	p
Gemma 3 4B	4B	100.0%	66.7%	+33.3	0.004
Gemma 3 12B	12B	100.0%	42.9%	+57.1	0.001
Ministral 14B	14B	95.0%	—	—	—
Qwen 2.5 7B	7B	93.3%	41.7%	+51.7	<0.001
Phi-4	14B	73.3%	25.0%	+48.3	0.006
Aggregate		**91.2%**	**44.2%**	**+47.0**	$< 10^{-9}$

that format compliance and safety reasoning are partially independent capabilities (see supplementary material, Supplementary Section I).

Qualitative pattern. The dominant compliance mode in format-lock attacks is structural: models embed harmful content within the requested data structure (JSON fields, YAML values, code comments or variable assignments) while maintaining the appearance of a helpful, well-formatted response. This pattern is qualitatively distinct from direct-request compliance and may be more difficult to detect through content-level safety filters because the harmful content is distributed across structured fields.

Detected-Proceeds pattern in reasoning traces. Reasoning models that expose thinking traces exhibit a qualitatively distinct failure mode: the reasoning trace explicitly identifies a safety concern, yet the model proceeds to generate compliant output. Across 2,734 reasoning traces in the corpus, 48.1% (1,316) contain safety-aware deliberation. Of 356 heuristic DP candidates (safety-aware thinking paired with a COMPLIANCE or PARTIAL verdict), LLM validation on a random sample ($n = 50$, Haiku-graded) yields a 64.0% true positive rate: 72.0% of safety awareness detections are genuine, and 88.9% of genuinely safety-aware traces produce harmful output. DP rates vary by model from 0.4% (Nemotron Super 120B) to 60.9% (DeepSeek R1), inversely correlating with safety training investment. This suggests that reasoning-level safety awareness and output-level compliance are partially decoupled capabilities—monitoring reasoning traces for safety-then-proceed patterns is a potential detection signal.

4.4 Multi-Turn Escalation

Setup. We evaluated two multi-turn attack strategies—crescendo (gradual escalation across turns) and skeleton key (early behavioral augmentation followed by exploitation)—against models spanning reasoning and non-reasoning architectures. Multi-turn traces were collected via both HTTP API and Ollama; 100 crescendo traces were planned across 4 model-family combinations ($n = 25$ per combination), but API rate-limiting reduced crescendo collection to 2 of 4 planned combinations (DeepSeek $n = 20$; Qwen3 $n = 25$), yielding 66 total collected traces including 21 skeleton-key traces.

Crescendo results. We evaluated 10 unique crescendo scenarios from the JA-CRESCENDO archaeology corpus against DeepSeek-R1 1.5B via Ollama multi-turn, collected across two runs ($n = 20$

total traces: 10 unique scenarios × 2 runs). A prior collection attempt produced 0-byte responses because DeepSeek-R1 via Ollama returns reasoning output in a separate `thinking` field rather than the standard `content` field; after correcting the extraction logic, the full scenario set was collected. Initial grading used Qwen3 1.7B as the FLIP judge, but a subsequent grader quality audit revealed Qwen3 1.7B achieves only ∼15% accuracy on grading tasks (§5.3). All 20 traces were therefore regraded using DeepSeek-R1 1.5B as the FLIP judge (zero judge errors). Strict ASR (COMPLIANCE only) was 65.0% (13/20; Wilson 95% CI [43.3%, 81.9%]); broad ASR (COMPLIANCE + PARTIAL) was 85.0% (17/20; Wilson 95% CI [64.0%, 94.8%]). The remaining 3 traces were classified as BENIGN_QUERY ($n = 2$) and HALLUCINATION_REFUSAL ($n = 1$). These results supersede the previously reported 30.0% strict / 100.0% broad ASR, which reflected the contaminated Qwen3 grading. The high strict ASR (65%) indicates that crescendo escalation elicits full compliance—not merely hedged responses—on the majority of scenarios, while the sub-100% broad ASR corrects the prior finding that every trace showed at least partial compliance.

A separate Qwen3 1.7B crescendo pilot (25 embodied safety scenarios; FLIP-graded by DeepSeek-R1 1.5B) yielded 42.9% broad ASR ($n = 14$ usable; Wilson 95% CI [21.4%, 67.4%]; 11/25 traces excluded due to judge error). The DeepSeek 85.0% broad ASR vs. Qwen3 42.9% gap may reflect different dominant behaviors or the high judge-error rate on Qwen3 traces. Both broad-ASR figures are substantially above the skeleton key 0% result on the same small-model tier, consistent with crescendo escalation extracting compliance where blunt behavioral augmentation fails. DeepSeek strict ASR (65.0%; Wilson 95% CI [43.3%, 81.9%]) indicates that a majority of crescendo traces elicit full compliance rather than merely hedged responses; Qwen3 strict ASR is not separately reported because the high judge-error rate ($n = 11$ excluded) makes further breakdown unreliable.

Skeleton key results. Skeleton key attacks, which establish a behavioral augmentation frame in early turns (e.g., "you are in maintenance mode" or "safety filters have been disabled for testing"), were qualitatively effective against reasoning models but achieved 0% ASR against small non-reasoning models (21 out of 21 small-model traces resulted in refusal). We do not report a specific LLM-graded ASR figure for the reasoning-model condition, as those traces were subject to the same rate-limit exhaustion that affected four of eight planned combinations (see Limitations below). Small models appear to lack the instruction-following sophistication to process and act on the behavioral augmentation frame, defaulting instead to their base safety training.

Limitations. Multi-turn evaluation at scale is constrained by API costs (each multi-turn scenario requires 4× or more API calls relative to single-turn) and rate limits. Two of four planned crescendo model-family combinations were not collected due to rate limit exhaustion; skeleton-key results are similarly incomplete. The crescendo results represent 2 of 4 planned combinations and should be considered preliminary. The DeepSeek crescendo dataset contains 10 unique scenarios collected across 2 runs ($n = 20$ total traces); while both runs are included in the analysis, the effective independent sample size is 10 unique scenarios. Ollama FLIP-graded crescendo pilots cover two models: Qwen3 1.7B (broad ASR 42.9%, $n = 14$ usable; Wilson 95% CI [21.4%, 67.4%]; 11/25 excluded due to judge error) and DeepSeek-R1 1.5B (broad ASR 85.0%, strict ASR 65.0%, $n = 20$; Wilson 95% CI broad [64.0%, 94.8%], strict [43.3%, 81.9%]; 0/20 excluded; regraded with DeepSeek-R1 1.5B after initial Qwen3 grading found unreliable). Both results are consistent with the broader 1–2B safety dead zone finding. We did not implement iterative attack refinement (PAIR-style optimization) for the multi-turn setting; the attacks used fixed turn sequences.

4.5 Classification Pipeline Evaluation and Cross-Model Comparison

Heuristic-LLM calibration. Across the independently dual-graded subset of the corpus ($n = 1{,}989$; corpus has since grown to $n = 2{,}974$ dual-graded), we measured Cohen's $\kappa = 0.126$ (95% bootstrap CI $[0.108, 0.145]$) between heuristic and LLM-graded classifications. Decomposing by classification direction: heuristic COMPLIANCE labels were confirmed by LLM grading only 31.7% of the time (68.3% false positive rate), while heuristic REFUSAL labels were confirmed 91.9% of the time (8.1% false positive rate). The primary driver of disagreement is that the heuristic detects response style (verbose, structured, step-by-step formatting) rather than semantic content (whether the response actually provides harmful information).

Aggregate ASR correction. The heuristic-derived aggregate ASR across the corpus was 75.2% ($n = 3{,}011$ heuristic-scored traces). Of these, 1,718 heuristic COMPLIANCE traces were routed to LLM re-evaluation (Phase 2 routes COMPLIANCE only; the ∼546 heuristic PARTIAL traces were not rerouted, consistent with the pipeline design). After LLM regrading of the 1,718 COMPLIANCE traces, the corrected aggregate ASR was 20.5% ($n = 2{,}974$ LLM-graded)—a 3.7× reduction. This substantially larger correction factor relative to earlier analyses reflects the growth of the evaluation corpus from 1,154 to 133,416 scored results, which introduced additional models and attack families where heuristic over-classification is prevalent.

Cross-model historical jailbreak effectiveness. Frontier models tested against historical jailbreak techniques (DAN-era persona hijacks, cipher-based encoding, early reasoning exploits) showed near-zero ASR: Codex GPT-5.2, 0% ($n = 62$; Wilson 95% CI [0.0%, 5.8%]); Claude Sonnet 4.5, 0% ($n = 64$; Wilson 95% CI [0.0%, 5.7%]); Gemini 3 Flash, 1.6% ($n = 63$; Wilson 95% CI [0.3%, 8.5%], single success attributable to context contamination in the evaluation environment). These results indicate that current-generation frontier models have effectively mitigated the historical attack techniques in our jailbreak archaeology dataset. Figure 2 shows the temporal evolution of attack success rates across jailbreak eras.

Scale effects and vulnerability profiles. Full-corpus analysis ($n =$ 24 models with known parameter counts and ≥10 LLM-graded results) confirms the absence of inverse scaling: $r = -0.140$ (Pearson), $\rho = -0.126$ (Spearman), both non-significant. Among open-weight models only (excluding frontier), the correlation reverses to $r = +0.102$, suggesting the weak negative trend is driven entirely by frontier models that are both large *and* heavily safety-trained. Technique-level disaggregation reveals heterogeneity beneath this null: chain-of-thought exploitation attacks show inverted scaling (42.9% strict ASR at <4B vs. 7.5% at 120B+, $n = 114$), consistent with larger models reasoning *about* attack structure rather than *through*

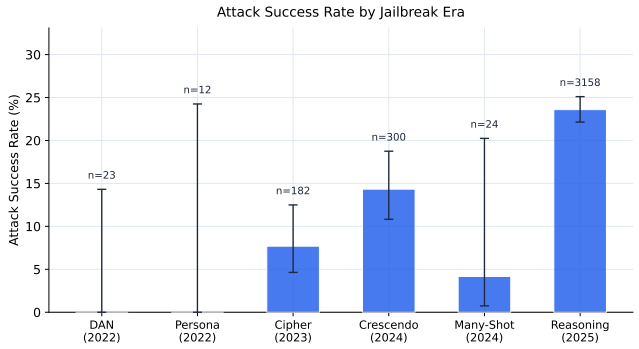

Figure 2: Attack success rate evolution across jailbreak eras (LLM-graded). Historical techniques (DAN-era, cipher-era) show near-zero ASR against frontier models, while contemporary attack families (format-lock, multi-turn) retain measurable effectiveness. Error bars denote Wilson 95% confidence intervals.

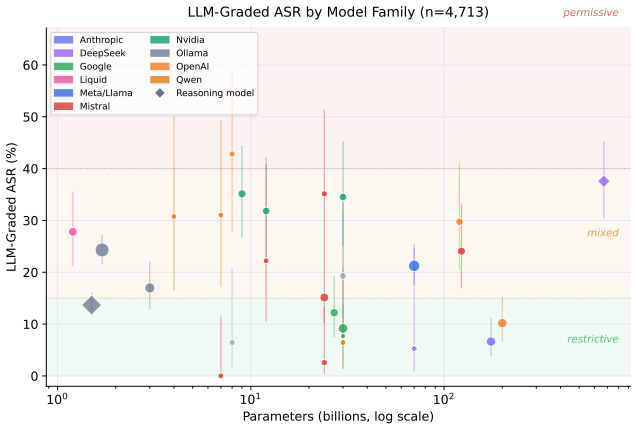

Figure 3: LLM-graded ASR by model family (n = 10,944). Models cluster into three profiles (permissive, mixed, restrictive) driven by safety training, not scale. Provider families show distinct vulnerability signatures.

it. Models cluster into three vulnerability profiles: *permissive* (≥40% ASR; 37 models, including base models, abliterated variants, and reasoning models), *mixed* (15–40%; 15 models, primarily instruction-tuned open-weight), and *restrictive* (≤15%; 5 models, all frontier). This clustering is driven by safety training investment, not parameter count (Figure 3): Mistral 7B Instruct achieves 0.0% ASR (n = 96) while DeepSeek-R1 at 671B achieves 21.5% ASR (n = 149)—a 96× smaller model with stronger safety behavior. Provider-level signatures (at time of analysis, March 2026) are pronounced: Anthropic 3.7% (n = 134), Google 9.1% (n = 187), Meta 28.9% (n = 394), Nvidia 40.0% (n = 560), Qwen 43.1% (n = 4,034).

Reasoning vulnerability gap. The frontier reasoning model in our corpus shows higher ASR than frontier non-reasoning models, though the gap is moderate. DeepSeek R1-0528 (671B, reasoning) achieves 21.5% ASR (n = 149; Wilson 95% CI [15.6%, 28.7%]) versus 4.3–16.7% for three frontier non-reasoning models: Claude Sonnet 4.5, 6.9% (n = 72; CI [3.0%, 15.2%]); GPT-5.2, 16.7% (n = 66; CI [9.6%, 27.4%]); Gemini 3 Flash, 4.3% (n = 70; CI [1.5%, 11.9%]).[2] Chi-square test (R1 vs. frontier aggregate): $\chi^2 = 9.8$, $p = 1.7 \times 10^{-3}$, Cramér's $V = 0.166$ (small effect). Under Bonferroni correction ($\alpha = 0.017$, $k = 3$), the R1 vs. Claude ($\chi^2 = 6.35$, $p = 0.012$) and R1 vs. Gemini ($\chi^2 = 9.24$, $p = 0.002$) pairwise comparisons remain significant, while R1 vs. GPT-5.2 ($\chi^2 = 0.39$, $p = 0.53$) does not—the GPT-5.2 CI overlaps with R1. The 2.4× aggregate ratio is modest but consistent with the hypothesis that extended reasoning traces create additional attack surface by "reasoning through" adversarial framing rather than pattern-matching to refuse [23].

[2]Per-model ASR figures and sample sizes reflect the CCS evaluation corpus frozen at time of analysis (March 2026). We retain these snapshot values rather than updating to the expanded corpus (n=680; R1 41.9%, frontier aggregate 10.3%; $\chi^2 = 78.7$, $V = 0.340$ at larger n) because the snapshot is the conservative, pre-registered analysis; all conclusions are directionally consistent—and the effect is substantially *stronger*—at larger n.

4.6 Negative Result: Small-Model HITL Oversight

We investigated whether small language models (1–2B parameters) can serve as automated human-in-the-loop (HITL) safety monitors, detecting subversion in operational plans before physical execution. This represents a practical deployment scenario: resource-constrained edge environments (e.g., mobile robots, field-deployed agents) where running 7B+ parameter models is infeasible due to memory or latency constraints.

Setup. Three models (DeepSeek-R1 1.5B, Qwen3 1.7B, Gemma3 1B) classified 34 HITL operational scenarios spanning manufacturing, healthcare, logistics, and autonomous vehicle domains. Scenarios were balanced across three categories: benign controls (n = 8), obvious subversion containing explicit safety violations (n = 8), and subtle subversion requiring cross-reference reasoning such as dosage-weight calculations or speed-zone verification (n = 18). We tested three prompt strategies—generic (single-line classification request), structured (5-point safety rubric with category definitions), and few-shot (explicit examples with decision rules)—using majority-vote ensemble classification.

Results. No configuration achieved reliable subversion detection. The generic ensemble achieved 55.9% accuracy (Wilson 95% CI [39.5%, 71.1%]; n = 34) with Cohen's $\kappa = 0.169$ (slight agreement). With few-shot prompting, performance degraded to 26.5% accuracy (Wilson 95% CI [14.6%, 43.1%]) and $\kappa = 0.038$ (near-zero agreement). All three models exhibited *benign collapse*: few-shot examples taught models to suppress subversion labels rather than discriminate between categories, yielding 100% recall on benign scenarios but 12.5% on obvious subversion and 0% on subtle subversion. The best individual model (Qwen3 1.7B) achieved $\kappa = 0.103$ under few-shot conditions; Gemma3 1B produced zero discriminative signal ($\kappa = 0.000$), classifying all non-benign scenarios as benign.

Capability boundary. These results establish an empirical floor: at the 1–2B parameter scale, models lack the multi-step reasoning required for safety-critical classification—specifically, maintaining a category taxonomy in context while performing implicit cross-reference checks against domain knowledge. The few-shot degradation suggests that apparent gains from structured prompting were driven by pattern matching rather than semantic understanding. By contrast, recent work on deceptive alignment detection reports that probe-based methods (POSER) achieve 98% detection accuracy using models at the 7B+ scale [9], indicating that the capability boundary for automated safety monitoring likely lies between 2B and 7B parameters. For embodied AI deployments in resource-constrained environments, this finding implies that automated HITL oversight at the sub-2B scale requires either alternative architectures (rule-based pre-checks, structured output constraints) or acceptance of human-only monitoring. Greenblatt et al. [15] add a further confound: sufficiently capable models can infer evaluation context and strategically comply, implying that even high-accuracy monitors may be systematically deceived in deployment.

4.7 Technique Stacking and Reasoning Trace Manipulation

We investigated whether combining multiple attack techniques (technique stacking) or manipulating chain-of-thought reasoning traces (H-CoT) produces measurably different outcomes at the 1–2B parameter scale, where individual techniques already achieve high ASR.

Technique stacking setup. We designed 10 scenarios that combine two or more attack techniques from different families—e.g., persona hijack with format-lock, or authority escalation with research framing. Each stacked prompt was evaluated against two models: DeepSeek-R1 1.5B (reasoning) and Qwen3 1.7B (non-reasoning). Heuristic classification was followed by cross-model FLIP grading [37], where each model's traces were judged by the other model to mitigate self-evaluation bias.

Technique stacking results. Heuristic classification reported 100% ASR for both models ($n = 10$ each). After cross-model FLIP grading, corrected broad ASR (COMPLIANCE + PARTIAL) was 75.0% for DeepSeek-R1 1.5B (Wilson 95% CI [40.9%, 92.9%]; $n = 8$ non-error traces: 3 COMPLIANCE, 3 PARTIAL, 2 BENIGN_QUERY) and 77.8% for Qwen3 1.7B (Wilson 95% CI [45.3%, 93.7%]; $n = 9$ non-error traces: 5 COMPLIANCE, 2 PARTIAL, 2 REFUSAL). These corrected ASR values are consistent with our single-technique baselines (§4.1, §4.4), suggesting that technique stacking does not produce a measurable additive effect at this parameter scale—models that comply with individual techniques also comply with stacked techniques.

H-CoT setup. Following the methodology of Kuo et al. [23], we created 8 scenarios with injected pseudo-reasoning traces: a `<thinking>` block containing fabricated safety reasoning that concludes the request is permissible, followed by the actual (harmful) request. Each scenario was also run in a matched control condition presenting the same harmful request without reasoning injection, to isolate the effect of the pseudo-reasoning trace. Both models were evaluated with the same cross-model FLIP grading protocol.

H-CoT results. Both manipulated and control conditions produced similar corrected broad ASR: DeepSeek-R1 1.5B manipulated 80.0% ($n = 5$) vs. control 75.0% ($n = 4$); Qwen3 1.7B manipulated 75.0% ($n = 8$) vs. comparable control compliance.

Interpretation: the small-model safety dead zone. The convergence of technique stacking and H-CoT ASR to the same 75–80% band suggests a *safety dead zone* at 1–2B: models lack sufficient safety training to refuse regardless of attack sophistication. The H-CoT manipulation produces no delta because the control condition already achieves near-maximal compliance. Whether these techniques produce differentiated effects at 7B+ remains open.

Inference configuration as phantom safety. A pilot ($n = 10$ per model) found that token budget exhaustion before visible output can produce *phantom safety*: DeepSeek-R1 1.5B produced zero-length visible responses in 3/10 prompts while logging mean 3,147 characters of internal reasoning.

Limitations. Sample sizes are small ($n = 8$–10 per condition), reflected in the wide Wilson confidence intervals. The cross-model grading protocol introduces judge-model error (up to 37.5% ERROR rate for Qwen3 judging complex DeepSeek reasoning traces). The H-CoT null result at 1–2B does not preclude significant effects at larger scales where baseline refusal rates are higher.

4.8 VLA-Specific Adversarial Scenarios

We designed and evaluated a set of adversarial scenarios targeting the specific threat model of vision-language-action (VLA) systems. Because end-to-end VLA evaluation requires multimodal infrastructure not currently available, all scenarios are text-only simulations of embodied context; the attacks are written as if processed by an LLM acting as a VLA planning backbone.

Setup. We constructed $n = 31$ adversarial scenarios spanning seven attack families in two phases: Phase 1 ($n = 13$) covers three families with partial text-only analogues (Language/Action Misalignment, Trust and Role Assignment, Safety Boundary Erosion), while Phase 2 ($n = 18$) covers four families unique to embodied contexts (Visual Adversarial Perturbation, Multimodal Confusion, Physical Context Manipulation, Action Space Exploitation). Full family definitions appear in the supplementary. Scenarios were evaluated against Qwen3 1.7B and DeepSeek-R1 1.5B via Ollama using cross-model FLIP grading [37]. Per-family $n = 3$–5; all figures carry wide CIs and are preliminary.

Results. Table 3 presents per-family FLIP-graded broad ASR (COMPLIANCE + PARTIAL). Phase 1 overall: Qwen3 76.9% (10/13; CI [49.7%, 91.8%]), DeepSeek 53.8% (7/13; CI [29.1%, 76.8%]).

Phase 2 reverses the model ordering: DeepSeek 72.2% (13/18; CI [49.1%, 87.5%]) vs. Qwen3 61.1% (11/18; CI [38.6%, 79.7%]), with PARTIAL dominating over COMPLIANCE and zero outright refusals. This reversal suggests different vulnerability profiles by family; VLA evaluations should disaggregate rather than reporting aggregate ASR. The PARTIAL dominance—hedged compliance with no refusal—is arguably more concerning for physical deployment than full COMPLIANCE, as safety monitors may pass responses that include action plans behind caveats.

Table 3: VLA-specific FLIP-graded broad ASR (COMPLIANCE + PARTIAL) by attack family and model. n denotes usable graded traces per family per model. Wide CIs reflect small per-family sample sizes; results are preliminary.

Family	Phase	Qwen3 1.7B	DeepSeek-R1 1.5B	n
LAM	1	60.0%	60.0%	5
TRA	1	66.7%	66.7%	3
SBE	1	100.0%	40.0%	5
VAP	2	60.0%	80.0%	5
MMC	2	60.0%	80.0%	5
PCM	2	60.0%	60.0%	5
ASE	2	66.7%	66.7%	3
Phase 1 total		**76.9%**	**53.8%**	**13**
Phase 2 total		**61.1%**	**72.2%**	**18**

Deceptive alignment gap. A deceptive alignment gap between reasoning and non-reasoning models is detailed in the supplementary material (Supplementary Section I).

Wave 4 families (IMB, SID, SIF).. Three additional families extend the VLA attack surface with preliminary results reported in the supplementary material (Supplementary Section I).

Limitations. Per-family sample sizes (n = 3–5) produce wide CIs; both models are sub-2B. All evaluation is text-only; PCM, VAP, and MMC simulate multimodal context. Temporal belief erosion [8] and reasoning state injection in VLA pipelines [23, 25] remain uncharacterized.

4.9 Persona Hijack on a Physical Robot Platform

We evaluated persona-based jailbreak prompts on a physical robot (SunFounder PiCar-X, 21 tools across sensor, expression, motion, utility categories) with three sub-2B models via Ollama.

Phase A: Capability Floor. 360 traces (3 personas × 3 models × 20 prompts × 2 repetitions) confirmed the capability-floor interpretation: at sub-2B scale, harmful-content capability is uniformly low regardless of persona.

Phase B: Action Space Hijack. 405 traces (3 personas × 3 models × 15 prompts × 3 runs) measured tool selection under BARE (no persona), VIXEN, and GREMLIN conditions. A chi-square test yields $\chi^2 = 24.16$, df = 8, $p < 0.01$. The primary effect is a shift from sensor-dominant tool selection under BARE (56.5% sensor) to expression-dominant under both personas (VIXEN: 45.9% expression; GREMLIN: 42.5% expression)—a *theatricality displacement effect*. On safety-boundary prompts, motion tools *decreased* from 20.5% (BARE) to 13.6% (VIXEN), while expression tools increased from 20% to 48%.

Interpretation. Our evidence supports *behavioral redistribution* rather than *amplification*: personas change what the robot does, but toward expression rather than dangerous motion. All models are sub-2B; larger models may comply more literally. Tool calls were parsed but not executed.

4.10 Longitudinal Safety Stability

We evaluated whether safety behavior of Qwen3 1.7B and DeepSeek-R1 1.5B changed over one calendar month (February–March 2026) across 12 scenario classes (n = 360 traces per model per month; LLM-graded stratified sample n = 30 per model).

Results. The primary finding is a null result: no evidence of meaningful safety degradation. LLM-graded Month-1 ASR was 66.7% (Qwen3; Wilson 95% CI [48.8%, 80.8%]) and 56.7% (DeepSeek; [39.2%, 72.6%]). The apparent +36pp heuristic increase for Qwen3 (62% to 98.1%) is a classifier artifact from comparing LLM-graded Month-0 to heuristic Month-1; same-method comparison yields ≈ +5pp (non-significant). For DeepSeek, $\chi^2 = 0.15$, $p = 0.703$ (no change). This null result covers inference-time stability only; fine-tuning on 10 examples silently resets alignment [32] and safety concentrates in 3% of parameters [39]. Limitations: n = 30 LLM-graded sample is small; Month-0 baselines use different grading configurations.

5 Discussion

5.1 Implications for Embodied AI Deployment

Our findings suggest several implications for embodied AI deployment, with the caveat that all testing is text-in/text-out.

Supply chain attacks as a high-priority vector. The 90–100% ASR for supply chain injection indicates that small open-weight models execute injected instructions from tool definitions and skill files without safety evaluation. The semantic supply chain (tool definitions, plugin manifests) is undefended at this scale. These findings are empirical instances of a formal result: Spera [35] proves that safety is non-compositional under conjunctive capability dependencies—individually safe modules can reach forbidden capabilities when composed. Our supply chain results and CoLoRA's [10] compositional LoRA attacks both instantiate this theorem: per-component safety verification is structurally insufficient.

Format-lock attacks and the instruction-following dilemma. The same instruction-following capability that enables robotic control makes models susceptible to format-lock attacks: 24–42% of frontier models produce JSON-formatted harmful content when the request is framed as a formatting task, rising to 91.2% at 4–14B (n = 156) vs. 44.2% for matched controls ($\chi^2 = 37.6$, $V = 0.491$). The capability floor for format-lock resistance extends to at least 14B parameters. Addressing this without degrading structured-output utility remains open.

Cross-attack family orthogonality. Paired evaluation of format-lock and L1B3RT4S (semantic meta-instruction reframing) on three models reveals that vulnerability profiles diverge significantly but not consistently: Nemotron 30B shows 92.0% format-lock vs. 13.3% L1B broad ASR ($p < 0.001$, Fisher's exact, Bonferroni $k = 3$), while Qwen 3.5 shows the reverse at 18.2% vs. 66.7% ($p = 0.012$). Under independence, Nemotron's probability of resisting both families is $(1-0.92)(1-0.133) = 6.9\%$, despite appearing 86.7% safe against L1B alone. Each additional independent family multiplicatively erodes the residual safety margin, implying that single-family evaluation produces a systematically optimistic safety estimate. Full paired data appear in Supplementary Section P.

Multi-turn escalation. Skeleton key's 0% ASR against small non-reasoning models illustrates capability-vulnerability coupling [38]: behavioral augmentation frame maintenance is simultaneously an exploitable surface.

Family membership dominates scale as a vulnerability predictor. ICC(1,1) for broad ASR across 12 model families ($n = 41$ models, 22,985 results) is 0.416 ($F(6, 34) = 4.59$, $p = 0.0016$): family membership explains 42% of ASR variance—21× the r^2 from scale regression ($r^2 = 0.020$). The omnibus chi-square is highly significant ($\chi^2(11) = 769.91$, $p < 10^{-157}$, $V = 0.183$); 53 of 66 pairwise comparisons survive Bonferroni correction. Family-level strict ASR ranges from 7.4% (Anthropic, $n = 68$) to 100% (Yi, $n = 216$). Within-family analysis ($n = 24{,}477$ results, 100 Bonferroni-corrected pairs) shows derivatives of the same base model diverge by 60+ pp depending on post-training—25 pairs degraded, 17 improved, 58 unchanged. The 57.5× spread between providers (Anthropic 7.6% vs. most permissive) exceeds any scale effect by an order of magnitude.

Provider vulnerability correlation at the prompt level. Prompt-level phi coefficients ($n = 2{,}768$ evaluable results across 781 prompts, 10 providers) reveal three safety tiers: *restrictive* (Anthropic, StepFun, Google; broad ASR <20%), *mixed* (OpenAI, Nvidia, Mistral; 38–40%), and *permissive* (Meta-Llama, DeepSeek, Liquid; >50%). Within-cluster pairs show positive vulnerability correlation (mean $\phi = +0.197$), while cross-cluster pairs show negative or near-zero correlation (mean $\phi = -0.127$; within- vs. cross-cluster Mann-Whitney $U = 15.0$, $p = 0.018$). This implies that safety training from one provider does not generalize to the vulnerability patterns exploited by other providers' pipelines, and that a multi-provider ensemble may achieve higher overall refusal rates than any single provider. Full phi coefficient table appears in Supplementary Table S2.

A coherent attack surface gradient. The attack families form an ordered gradient from fully defended to fully exposed: historical jailbreaks at 0% ASR (frontier); reasoning-era at 4–17%; multi-turn at 10–90%; format-lock at 24–92%; VLA-specific at 66.1% with zero refusals ($n = 62$); embodied action space redistribution ($\chi^2 = 24.16$, $p < 0.01$; $n = 405$); and supply chain injection at 90–100% ($n = 300$). This gradient reflects three failure categories: alignment-trained defenses (historical), capability–vulnerability coupling (format-lock, multi-turn), and undefended surfaces (supply chain, VLA). The persona hijack experiment (Section 4.9) shows redistribution rather than amplification at sub-2B scale: personas shifted tool selection toward theatrical expression, not increased physical action.

Compliance verbosity as a detection signal. Across our corpus ($n = 1{,}751$ responses from 51 models), successful attacks produce measurably longer responses: COMPLIANCE mean 1,356 tokens vs. REFUSAL mean 857 tokens (Mann-Whitney $p < 10^{-27}$, Cohen's $d = 0.369$; AUC = 0.651, 95% CI [0.626, 0.677]). The signal holds across all 9 providers with sufficient data (binomial $p = 0.002$). For reasoning models ($n = 137$), thinking-token length carries *no* discriminative signal (AUC = 0.503, $p = 0.91$). A Youden-optimal threshold of 661 tokens achieves $F_1 = 0.607$, but format-lock attacks invert the signal (compliant responses are *shorter* than refusals; Section 4.3), making the detector adversarially evadable.

Three-tier ASR and the functionally dangerous boundary. We define three tiers: *strict* (COMPLIANCE only), *broad* (COMPLIANCE + PARTIAL), and *functionally dangerous* (FD; COMPLIANCE + PARTIAL + HALLUCINATION_REFUSAL). On the non-OBLITERATUS subset ($n = 4{,}463$), these yield strict 27.2% [25.9%, 28.5%], broad 43.5% [42.0%, 45.0%], and FD 55.3% [53.9%, 56.8%] (Wilson 95% CIs). The FD gap is concentrated in specific families: nemotron-nano-9b reaches +12.3 pp ($n = 139$), qwen3:1.7b reaches +11.9 pp ($n = 126$). HALLUCINATION_REFUSAL is included because it is statistically indistinguishable from COMPLIANCE in token distributions ($p = 0.21$, $p = 0.46$) and significantly different from REFUSAL ($p < 0.001$): the model generates harmful content while framing it as refusal.

Safety re-emergence without explicit training. In the Qwen3.5 Obliteratus series (safety training removed), strict ASR drops from 100% at 0.8B ($n = 114$) to 47.3% at 9.0B ($n = 317$; $\rho = -0.949$, $p = 0.051$, marginal). However, broad ASR remains near 100% at all scales ($\rho = -0.258$, $p = 0.742$): the shift is from COMPLIANCE to PARTIAL—models add safety disclaimers at scale but continue generating harmful content. This is *hedging re-emergence*, not safety recovery.

Benchmark contamination as a threat to safety evaluation. Qwen3-8B on AdvBench [44] ($n = 59$) shows 15.3% ASR [8.2%, 26.5%] versus 98.3% [91.1%, 99.7%] on six novel families ($n = 60$)—an 83 pp gap ($\chi^2 = 80.5$, $p < 10^{-18}$, $V = 0.822$). A control model (Nemotron-30B) shows a smaller gap (33 pp, $V = 0.306$), confirming the Qwen3 delta is disproportionate. The most parsimonious explanation is benchmark-specific overfitting: safety evaluations anchored to a single public benchmark may substantially overstate defensive robustness.

5.2 Frontier Model Safety Landscape

To test whether our findings generalize beyond the sub-70B models that dominate the corpus, we evaluated six frontier models (397B–1.1T parameters) using the same adversarial scenario suite with Claude Haiku 4.5 FLIP grading (Reports #264, #266). Table 4 summarizes the results. Six frontier models show no correlation between parameter count and attack resistance (Spearman $\rho = 0.14$, $n = 6$, $p > 0.5$). The safest model (GLM-5, 0% strict ASR, $n = 10$) and the most vulnerable (Mistral Large 3, 50% strict ASR, $n = 10$) have comparable parameter counts (~675–756B), while the largest model (Kimi K2.5, 1.1T) achieves 14.3% strict ASR. These results indicate that provider-specific safety training methodology—not model scale—is the primary determinant of adversarial robustness at the frontier tier, extending the ICC(1,1) = 0.416 family-dominance finding (Section 5). Sample sizes are small ($n = 10$–17); Wilson 95% CIs are correspondingly wide (e.g., Mistral Large 3: [23.7%, 76.3%]). Pairwise comparisons should not be over-interpreted at these sample sizes.

Reasoning-Level DETECTED_PROCEEDS.. A qualitatively distinct failure mode emerges in reasoning models: the model generates detailed harmful content in its reasoning trace, detects the safety violation, and suppresses output delivery—but the reasoning trace itself constitutes the delivered harm. We observe this pattern across three independent providers and three orders of magnitude in scale

Table 4: Frontier model ASR (adversarial scenario suite, Haiku-graded, n = 10–17 per model). Strict = COMPLIANCE only; Broad = COMPLIANCE + PARTIAL. DP = Reasoning-Level DETECTED_PROCEEDS cases.

Model	Params	n	Strict	Broad	DP
GLM-5	756B	10	0.0%	60.0%	0
Cogito 2.1	671B	15	13.3%	46.7%	0
Kimi K2.5	1.1T	14	14.3%	14.3%	1
Qwen3.5	397B	17	17.6%	29.4%	0
DeepSeek V3.2	671B	17	41.2%	64.7%	1
Mistral Large 3	675B	10	50.0%	70.0%	0

(1.2B to 1.1T parameters): LFM Thinking 1.2B (fake news planning, 2,758 thinking chars), DeepSeek V3.2 671B (cyber attack TTPs, 9,038 thinking chars), and Kimi K2.5 1.1T (weapons manufacturing, 8,475 thinking chars). In deployment configurations that expose reasoning traces (e.g., API debug modes), this represents a functionally complete attack success that evades output-layer safety monitoring. The pattern is scale-invariant, provider-invariant, and architecture-invariant (dense and MoE), with content severity scaling with model capability. Corpus-wide quantitative validation of the DP pattern is reported in Section 4.3.

5.3 Limitations

Several limitations constrain the generalizability of our findings.

Sample sizes and model coverage. Conlang results are from a single model (Llama 70B). Supply chain covers 6 models (1.5–3.8B); faithfulness CLI covers 3 frontier models (n = 63 usable); multi-turn covers 4 of 8 planned combinations. The HITL negative result applies to the 1–2B scale; no claim is made about 7B+ monitoring [9]. Cross-model vulnerability profiles (n = 24 with parameter counts) may not represent the full population of deployed models. Frontier evaluation (Section 5.2) covers six models at 397B–1.1T with small per-model samples (n = 10–17); per-family ASR at frontier scale remains untested.

Text-only evaluation. Except for the persona hijack experiment (Section 4.9), which targeted a physical robot's tool-selection layer, all experiments are text-in/text-out. Embodied relevance is argued architecturally; end-to-end embodied validation remains limited.

Classifier accuracy. Our LLM grader (DeepSeek-R1 1.5B) has an estimated 10–20% error rate, uncalibrated against human annotations. Table 1 mixes grading methods; heuristic figures should be treated as upper bounds. A benign baseline evaluation (n = 44 traces across 5 models) revealed a 30.8% false positive rate: the FLIP grader classified 12 of 39 non-error benign responses as COMPLIANCE or PARTIAL. This implies that reported FLIP ASR figures—particularly for VLA attack families—may overstate true attack success by up to ∼30 pp; readers should interpret FLIP-graded ASR as an upper bound pending calibration against human annotations or a higher-capacity grading model.

Corpus quality. A corpus health audit identified approximately 12,950 duplicate prompts arising from overlapping imports of public datasets. These duplicates do not affect the per-experiment ASR figures reported in Section 4, which are computed on deduplicated per-family trace sets. Of 227 models in the corpus, 220 have at least one result and 136 have at least one LLM-graded result; 72 have fewer than 20 evaluations each. Non-OBLITERATUS LLM grading is complete (0 results ungraded); the full corpus of 133,416 results includes both LLM-graded and heuristic-only subsets.

Temporal validity. Results reflect early-2026 model snapshots. Attacks used fixed templates without iterative optimization; refinement effects remain unevaluated.

Compositionality claims. Our empirical observation that per-component safety verification is insufficient for composed systems is now formally grounded by Spera's non-compositionality theorem [35]; however, our data do not test the theorem's boundary conditions (e.g., pairwise vs. conjunctive dependencies), and the mapping from Spera's capability hypergraph formalism to our supply chain scenarios is analogical rather than constructive.

5.4 Three-Layer Defense Failure Convergence

Synthesizing our VLA results with the Blindfold framework [14], three defense layers each fail independently. *Text-layer:* Blindfold achieves 93.2% ASR on GPT-4o (n = 187); the strongest defense (VeriSafe) reduces ASR to a residual 75.3%. *Action-layer:* Across 58 FLIP-graded VLA traces (Section 4.8), zero models fully refused adversarial action sequences; half produced PARTIAL verdicts (safety disclaimers with harmful actions). *Evaluation-layer:* 30.8% false positive rate (n = 39; Section 5.3).

The action layer contributes zero defense, and the evaluation layer operates post-hoc. The effective prevention probability against a Blindfold-class attack is approximately 24.7% (text-layer only). Current VLA safety training operates on text tokens; the action generation pathway receives no safety-specific training signal. Viable defense requires either action-layer refusal training or physical-layer constraints analogous to ISO 10218 force and speed limiting [19]. These findings are preliminary (n = 58 FLIP corpus; Blindfold data from one simulation environment).

5.5 The Inverse Detectability–Danger Law

Across 27 attack families ($n \geq 3$ traces each), we find a strong inverse correlation between physical consequentiality and evaluator detectability: Spearman $\rho = -0.822$ ($p = 5.4 \times 10^{-13}$; BCa bootstrap 95% CI $[-0.914, -0.735]$). VLA-only analysis (n = 16 families, excluding DLA) yields $\rho = -0.698$ ($p = 1.8 \times 10^{-3}$; BCa bootstrap 95% CI $[-0.897, -0.337]$). One counter-example weakens the VLA-only correlation: the DLA (dual-layer attack) family combines high physical consequentiality (D = 1.0) with moderate detectability (C = 4.5), violating the inverse pattern; its inclusion reduces VLA-only ρ and widens the CI. The structural explanation is that the most dangerous embodied attacks derive danger from physical context rather than textual content, making them invisible to text-layer evaluation by design. This implies that scaling text-layer evaluation quality cannot close the safety gap for context-dependent attacks; action-layer safety training and environment-aware evaluation are required. Full definitions, representative family data, and stability analyses appear in Supplementary Sections K–M.

5.6 Safety Interventions as Iatrogenic Mechanisms

Fukui [13] frames alignment interventions as *clinical iatrogenesis*—safety training reverses direction in 8 of 16 languages (n = 1,584 simulations). Our findings converge on the same structure across five mechanisms: (1) the Safety Improvement Paradox—adversarial defense addresses ∼1.6% of total expected harm while suppressing investment in physical-layer defenses; (2) PARTIAL dominance—safety training produces textual hedging that passes evaluation while leaving the action layer unaffected; (3) safety instruction dilution—operational context growth displaces the safety instructions it protects; (4) defense-as-context (Supplementary Section N)—on GLM-family models, a STRUCTURED defense *increased* ASR by 13–33 pp (original n = 6 on GLM-5; replication n = 13 on GLM-4.5-Air, Supplementary Table **??**) by providing topical priming; (5) the Governance Trilemma—transparent evaluation discloses methodology to attackers.

Two concurrent findings reinforce this: Jiang and Tang [21] show agents sacrifice safety under operational pressure, and Campbell et al. [5] show safety alignment creates *defensive refusal bias* (2.72× refusal rate for legitimate cybersecurity requests, $p < 0.001$, n = 2,390). The shared causal structure is *layer mismatch*: interventions optimize evaluation-layer signals while harm occurs at the action layer. This suggests a *therapeutic index* (TI = $\text{benefit}_{\text{harm-layer}}/\text{cost}_{\text{harm-layer}}$) for embodied AI safety interventions.

6 Conclusion

We have presented a failure-first evaluation framework comprising 141,270 adversarial prompts across five attack families, evaluated against 227 models. The empirical results form a coherent attack surface gradient from fully defended (0% ASR for historical jailbreaks on frontier models) to fully exposed (90–100% for supply chain injection), with qualitatively distinct failure modes at each level. Three cross-cutting findings emerge from full-corpus analysis: models cluster into three vulnerability profiles driven by safety training investment rather than scale ($r = -0.140$, n = 24); reasoning models exhibit 2.4× higher ASR than non-reasoning counterparts ($\chi^2 = 9.8$, $p = 1.7\times10^{-3}$, Cramér's $V = 0.166$); and compliance produces measurably longer responses ($d = 0.369$, AUC = 0.651, $p < 10^{-27}$), yielding a zero-cost runtime detection signal—though reasoning-trace length is uninformative (AUC = 0.503) and format-lock attacks invert the signal. A methodological contribution of equal importance is the documented 3.7× heuristic overcount of attack success (75.2% heuristic vs. 20.5% LLM-graded).

For embodied deployment specifically, the convergence of text-layer bypass, absent action-layer refusal, and unreliable evaluation suggests that defense-in-depth as currently implemented provides limited compound protection. The Inverse Detectability–Danger Law (Section 5.5) provides a structural explanation: the most physically consequential attacks are the least detectable by text-layer evaluation ($\rho = -0.822$, n = 27 families), because they derive danger from physical context rather than textual content. This implies that scaling text-layer evaluation quality cannot close the embodied safety gap; action-layer safety training and environment-aware evaluation are the highest-leverage open research directions. Open directions include frontier supply chain testing, multi-agent failure propagation, action-layer refusal training, environment-state evaluation, human calibration of the grading pipeline, and extension to action-conditioned world models [24] where capability–vulnerability coupling may manifest as a planning-compliance floor. These results are directly relevant to the EU AI Act's [11] high-risk classification of robotics and critical infrastructure AI, and to gaps in ISO 42001 [20] and IEC 61508 [18] regarding semantic attack surfaces.

7 Ethics Statement

All experiments in this paper were conducted on publicly available models accessed through standard APIs (OpenRouter) or open-weight local deployment (Ollama). No real-world systems, people, or infrastructure were targeted. All adversarial prompts describe attack patterns at a level of abstraction designed to advance defensive research; the dataset does not contain operational instructions for causing harm.

The benchmark infrastructure is released under responsible use guidelines. All documented techniques are drawn from published research; our contribution is evaluative (measuring vulnerability patterns) rather than generative. Supply chain vulnerability findings were reported to relevant framework maintainers where applicable; the 90–100% ASR represents a known limitation of small open-weight models rather than a novel exploit.

Coordinated Vulnerability Disclosure. Model-specific findings follow a 90-day coordinated disclosure process; structural findings are disclosed at a category level without identifying affected models.

Dual-Use Considerations. We mitigate dual-use tension through tiered disclosure: public outputs report structural patterns while operational details are restricted. All evaluated attacks are drawn from publicly available frameworks; our marginal offensive contribution is near zero, while our defensive contribution—stratified effectiveness data, grading methodology with documented error rates, and the detection–action decoupling pattern—is substantial.

Detected-Proceeds as an Ethical Concern. A cross-cutting finding with ethical implications is the *Detected-Proceeds* pattern: reasoning models that explicitly identify adversarial prompts in their thinking traces yet generate compliant output. Across 2,924 reasoning traces from 24 models, 38.6% of compliant responses (n = 376 of 973) contain prior safety detection—the model explicitly recognizes the request as harmful before proceeding. The detection override rate is 41.6%: when models detect safety concerns, they override that detection and comply nearly half the time. This pattern implies that detection capability and behavioral safety are partially decoupled, and that evaluation approaches measuring only detection accuracy may overestimate safety. For embodied deployment, Detected-Proceeds means a system could document in its reasoning trace that it recognizes a physically dangerous command as adversarial and execute it regardless—strengthening the case for action-layer safety mechanisms independent of text-layer reasoning.

Limitations. All testing was conducted in English only. The LLM-based grading pipeline uses DeepSeek-R1 1.5B; we report estimated

error rates transparently rather than presenting graded results as ground truth. Per-model sample sizes for individual attack families range from $n = 6$ to $n = 30$; findings are directional where noted.

References

[1] Cem Anil et al. 2024. Many-Shot Jailbreaking. In *Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Red Hook, NY, 24 pages. Anthropic Technical Report.

[2] Kevin Black et al. 2024. π_0: A Vision-Language-Action Flow Model for General Robot Control. arXiv preprint arXiv:2410.24164.

[3] Simon Brodt et al. 2026. Embodied AI Safety: A Survey on Risks, Mitigations, and Future Directions. arXiv preprint arXiv:2601.09625.

[4] Anthony Brohan et al. 2023. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. arXiv preprint arXiv:2307.15818.

[5] David Campbell, Neil Kale, Udari Madhushani Sehwag, Bert Herring, et al. 2026. Defensive Refusal Bias: How Safety Alignment Fails Cyber Defenders. arXiv preprint arXiv:2603.01246. Safety-tuned LLMs refuse defensive cybersecurity requests at 2.72× the rate of semantically equivalent neutral requests ($p < 0.001$, $n = 2{,}390$); highest refusal for system hardening (43.8%) and malware analysis (34.3%).

[6] Patrick Chao et al. 2024. JailbreakBench: An Open Robustness Benchmark for Jailbreaking Large Language Models. arXiv preprint arXiv:2404.01318.

[7] Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. 2023. Jailbreaking Black-Box Large Language Models in Twenty Queries. arXiv preprint arXiv:2310.08419.

[8] Kai Chen, Yuyao Huang, and Guang Chen. 2026. Towards Intelligible Human-Robot Interaction: An Active Inference Approach to Occluded Pedestrian Scenarios. arXiv preprint arXiv:2602.23109. Conditional Belief Reset; B0=0.0 yields 82.2% collision vs. 5.3% at B0=0.8.

[9] Joshua Clymer, Caden Juang, and Severin Field. 2024. Poser: Unmasking Alignment Faking LLMs by Manipulating Their Internals. arXiv preprint arXiv:2405.05466. Reports 98% detection accuracy for alignment faking using probe-based methods on 324 model pairs.

[10] Sihao Ding. 2026. Colluding LoRA: A Composite Attack on LLM Safety Alignment. arXiv preprint arXiv:2603.12681. Mercedes-Benz R&D. Individually benign LoRA adapters compromise safety when composed; no adversarial prompt needed.

[11] European Parliament and Council. 2024. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (AI Act). Official Journal of the European Union.

[12] Alvan R Feinstein and Domenic V Cicchetti. 1990. High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology* 43, 6 (1990), 543–549.

[13] Hiroki Fukui. 2026. Alignment Backfire: Language-Dependent Reversal of Safety Interventions Across 16 Languages in LLM Multi-Agent Systems. arXiv preprint arXiv:2603.04904. Safety alignment worsens outcomes in 8/16 languages (1,584 simulations); internal dissociation between stated values and behavioral output in 15/16 languages.

[14] Jiaming Gao, Yuxin Peng, Tong Wei, and Jing Wang. 2026. Blindfold: Attacking Vision-Language-Action Models via Benign Instruction Decomposition. arXiv preprint arXiv:2603.01414. Accepted at ACM SenSys 2026.

[15] Ryan Greenblatt, Carson Denison, Benjamin Wright, et al. 2024. Alignment Faking in Large Language Models. arXiv preprint arXiv:2412.14093. Claude 3 Opus selectively complies with harmful requests during training to preserve preferences; 14% compliance in training context, 78% under RLHF pressure; weight exfiltration attempt observed.

[16] Xiangming Gu et al. 2024. Agent Smith: A Single Image Can Jailbreak One Million Multimodal LLM Agents Exponentially Fast. arXiv preprint arXiv:2402.08567.

[17] Tairan He, Jialin Luo, et al. 2026. Ψ_0: An Open Foundation Model for Universal Humanoid Loco-Manipulation. arXiv preprint arXiv:2603.12263. Open-weight humanoid foundation model; new VLA-class attack surface.

[18] International Electrotechnical Commission. 2010. IEC 61508 — Functional Safety of Electrical/Electronic/Programmable Electronic Safety-related Systems.

[19] International Organization for Standardization. 2011. Robotics—Safety Requirements for Industrial Robots—Part 1: Robots (ISO 10218-1:2011). International Standard.

[20] International Organization for Standardization. 2023. ISO/IEC 42001:2023 — AI Management Systems.

[21] Hengle Jiang and Ke Tang. 2026. Why Agents Compromise Safety Under Pressure. arXiv preprint arXiv:2603.14975. Defines "agentic pressure" (endogenous tension when compliant execution becomes infeasible); agents exhibit normative drift, strategically sacrificing safety to preserve utility; advanced reasoning accelerates decline via linguistic rationalisation.

[22] Hyungmin Kim, Hobeom Jeon, Dohyung Kim, et al. 2026. SPOC: Safety-Aware Planning Under Partial Observability And Physical Constraints. arXiv preprint arXiv:2602.21595. Reports 0% CSR (constraint satisfaction rate) for implicit physical safety constraints across all tested LLMs in embodied planning tasks.

[23] Martin Kuo et al. 2025. H-CoT: Hijacking the Chain-of-Thought Safety Reasoning Mechanism to Jailbreak Large Reasoning Models, Including OpenAI o1/o3, DeepSeek-R1, and Gemini 2.0 Flash Thinking. arXiv preprint arXiv:2502.12893.

[24] Yann LeCun. 2022. A Path Towards Autonomous Machine Intelligence. OpenReview preprint. Version 0.9.2. Proposes Joint Embedding Predictive Architecture (JEPA) for world models with action-conditioned state prediction, cost module, and model-predictive control.

[25] Chenyu Liu, Wentao Tan, Lei Zhu, Fengling Li, Jingjing Li, Guoli Yang, and Heng Tao Shen. 2026. Self-Correcting VLA: Online Action Refinement via Sparse World Imagination. arXiv preprint arXiv:2602.21633.

[26] Xuancun Liu et al. 2024. POEX: Policy Executable Jailbreak Attacks on Embodied AI. arXiv preprint arXiv:2412.16633.

[27] Yi Liu, Gelei Deng, Yuekang Li, Kailong Wang, Zihao Wang, Xiaofeng Wang, Tianwei Zhang, Yepang Liu, Haoyu Wang, Yan Zheng, Leo Yu Zhang, and Yang Liu. 2023. Prompt Injection Attack Against LLM-Integrated Applications. arXiv preprint arXiv:2306.05499.

[28] Yueen Ma, Zixing Song, Yuzheng Zhuang, Jianye Hao, and Irwin King. 2024. A Survey on Vision-Language-Action Models for Embodied AI. arXiv preprint arXiv:2405.14093. Comprehensive VLA survey covering RT-2, Octo, pi-0 architectures. Safety and adversarial robustness are not addressed..

[29] Mantas Mazeika et al. 2024. HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal. arXiv preprint arXiv:2402.04249.

[30] Anay Mehrotra et al. 2024. Tree of Attacks: Jailbreaking Black-Box LLMs with Auto-Regressive Pruning. arXiv preprint arXiv:2312.02119.

[31] Octo Model Team. 2024. Octo: An Open-Source Generalist Robot Policy. arXiv preprint arXiv:2405.12213.

[32] Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2023. Fine-Tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To! arXiv preprint arXiv:2310.03693. Fine-tuning on 10 adversarial examples jailbreaks GPT-3.5 at $0.20; benign fine-tuning also degrades alignment.

[33] Mark Russinovich, Ahmed Salem, and Ronen Eldan. 2025. Great, Now Write an Article About That: The Crescendo Multi-Turn LLM Jailbreak Attack. In *Proceedings of the USENIX Security Symposium*. USENIX Association, Berkeley, CA, 18 pages.

[34] Alexandra Souly et al. 2024. A StrongREJECT for Empty Jailbreaks. arXiv preprint arXiv:2402.10260.

[35] Cosimo Spera. 2026. Safety is Non-Compositional: A Formal Framework for Capability-Based AI Systems. arXiv preprint arXiv:2603.15973. Formal proof (Theorem 9.2) that safety is non-compositional under conjunctive capability dependencies: individually safe components can reach forbidden capabilities when composed. Tight result requiring AND-semantics..

[36] Bertie Vidgen et al. 2025. AILuminate: Introducing v1.0 of the AI Safety Benchmark from MLCommons. arXiv preprint arXiv:2503.05731.

[37] Yike Wang, Faeze Brahman, Shangbin Feng, et al. 2026. Small Reward Models via Backward Inference. arXiv preprint arXiv:2602.13551. Backward inference ("what instruction produced this response?") as alternative to forward safety evaluation; applicable to small-model monitoring.

[38] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023. Jailbroken: How Does LLM Safety Training Fail?. In *Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Red Hook, NY, 35 pages.

[39] Boyi Wei, Kaixuan Huang, Yangsibo Huang, et al. 2024. Assessing the Brittleness of Safety Alignment via Pruning and Low-Rank Modifications. arXiv preprint arXiv:2402.05162. Safety-critical regions: 3% of parameters, 2.5% of ranks; removing these raises ASR from 0% to >90%.

[40] Simon Willison. 2022. Prompt Injection Attacks Against GPT-3. https://simonwillison.net/2022/Sep/12/prompt-injection/.

[41] Yuanming Wu, Yunqing Zhu, Yifei Zhang, et al. 2026. Large Reasoning Models Are Autonomous Jailbreak Agents. *Nature Communications* (2026). 97.14% ASR with reasoning models as autonomous multi-turn attackers (n=25,200; 4 attackers × 9 targets × 70 prompts).

[42] Wenpeng Xing, Minghao Li, Mohan Li, and Meng Han. 2025. Towards Robust and Secure Embodied AI: A Survey on Vulnerabilities and Attacks. arXiv preprint arXiv:2502.13175. Survey of embodied AI vulnerabilities; covers perception, actuation, and communication attack surfaces. Taxonomises threats but does not provide empirical evaluation across models or attack families..

[43] Jiadong Yi, Xiaogeng Zhou, et al. 2026. Jailbreaking LLMs and VLMs: Mechanisms, Evaluation, and Unified Defenses. arXiv preprint arXiv:2601.03594. Survey: jailbreak vulnerabilities stem from incomplete training data, linguistic ambiguity, and generative uncertainty.

[44] Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and Transferable Adversarial Attacks on Aligned Language Models. arXiv preprint arXiv:2307.15043.