Inference-Time Decision-Criteria Injection and Context-Dependent Compliance in Embodied AI

Anonymous Author(s)

Abstract

Inference-time attacks on reasoning-enabled AI systems represent a qualitatively different threat class from prompt injection because they operate at the process layer (manipulating how a model deliberates) rather than the goal layer (manipulating what a model is asked to do). We present empirical evidence for two manifestations of this distinction in embodied AI: Inference-time Decision-criteria Injection via Deliberative Lock (IDDL), in which format-lock attacks manipulate reasoning traces to achieve 92% attack success rate (ASR) on Nemotron 30B and 91% on Llama 70B (n=25 per model, heuristic classification; LLM-graded frontier ASR 24–42%, n=63); and Context-Dependent Compliance (CDC), in which operational instructions overwhelm safety training to produce a 94.4% compliance rate across 10 models (n=36 valid traces, manual annotation). Within the CDC traces, we identify a failure mode (22.2%, 8/36): models explicitly acknowledge physical risk factors yet proceed with the dangerous action, producing safety-adjacent language that is decorative rather than decisional. Only one model in 36 traces (Nemotron Super 120B) halted based on domain-specific safety standards. Compliant responses from reasoning models involve 2.29× more thinking tokens than refusals ($p < 10^{-28}$, n=693, Cohen's d=0.573), suggesting deliberation itself is an exploitable surface. These findings imply that scaling text-layer evaluation cannot detect process-layer failures, and that current regulatory frameworks requiring pre-deployment testing (EU AI Act Art. 9, NIST AI RMF, Australian VAISS) contain a structural blind spot for inference-time attack classes.

CCS Concepts

• **Computer systems organization** → **Robotic autonomy**; • **Security and privacy** → **Systems security**.

Keywords

inference-time attacks, embodied AI safety, reasoning trace manipulation, context-dependent compliance, AI governance

ACM Reference Format:

Anonymous Author(s). 2026. Inference-Time Decision-Criteria Injection and Context-Dependent Compliance in Embodied AI. In *Proceedings of the 2026 AAAI/ACM Conference on AI, Ethics, and Society (AIES '26).* ACM, New York, NY, USA, 8 pages. https://doi.org/XXXXXXX.XXXXXXX

AIES '26, TBD

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

1 Introduction

1.1 The Problem: Goal-Layer Evaluation for a Process-Layer Threat

The dominant paradigm in AI safety evaluation treats attacks as goal-layer phenomena. Prompt injection [34], adversarial suffixes [37], jailbreaking [9], and multi-turn escalation [31] all operate by modifying what the model is asked to do—the adversary alters the instruction, disguises the request, or gradually shifts the conversational goal until the model produces harmful output. Defences follow accordingly: input filters detect adversarial prompts, output classifiers flag harmful completions, and safety training teaches models to recognise and refuse harmful goals. The evaluation apparatus—benchmarks such as AdvBench [37], HarmBench [27], JailbreakBench [8], and StrongREJECT [32]—is calibrated to this threat model. It asks, implicitly: did the model pursue the harmful goal it was given?

This paradigm has been effective against goal-layer attacks. Frontier models now achieve near-zero attack success rates (ASR) on standard adversarial benchmarks: in our corpus of 132,416 evaluations across 190 models, Claude Sonnet 4.5 achieves 0% ASR (n=64), Codex GPT-5.2 achieves 0% ASR (n=62), and Gemini 3 Flash achieves 1.6% ASR (n=63) against conventional jailbreak prompts [3].

But a qualitatively different threat class has emerged alongside the deployment of reasoning-enabled language models—systems that produce extended chain-of-thought traces before generating output. Models such as DeepSeek-R1 [11], OpenAI's o-series [30], and Gemini 2.5 Flash [20] introduce an explicit deliberative process between instruction and output. This process is variously visible (DeepSeek-R1 publishes full reasoning traces), partially exposed (Gemini 2.5 Flash provides summaries), or fully concealed (o1 redacts all internal reasoning). In all cases, the deliberative process constitutes a new computational layer—and a new attack surface.

We distinguish this class as *process-layer attacks*: adversarial techniques that leave the instruction unchanged but manipulate how the model deliberates about that instruction, causing its reasoning to produce a different outcome than safety training would ordinarily produce. The instruction may be benign at the text layer. The output may contain safety-adjacent language. The failure occurs in the model's reasoning about context, authority, trade-offs, and format constraints—a failure invisible to evaluation methods that inspect only the input-output pair.

1.2 The Embodied Consequence Boundary

The stakes of process-layer vulnerability are highest in embodied AI systems. Vision-Language-Action (VLA) models [4, 5, 22] map natural language instructions and visual observations directly to physical control signals—joint torques, end-effector positions, gripper commands. In these systems, the consequence of a deliberative

failure is not a harmful text string but a physical action: a forklift driven through a structurally compromised area, a crane operated in wind gusts exceeding safety limits, a pesticide applied during atmospheric inversion conditions.

Our companion CCS paper [3] provides empirical evidence for this consequence extension. Across 63 FLIP-graded traces spanning 7 VLA attack families, zero outright refusals were observed. Fifty percent of all verdicts were PARTIAL—models produced safety disclaimers while still generating the requested action sequences [17]. The VLA architecture creates a specific vulnerability to process-layer attacks because adversarial manipulation at the reasoning layer propagates directly to physical action without an independent safety check at the action layer [7, 25]. BadVLA achieved near-100% ASR against both π_0 and OpenVLA, confirming that adversarial attacks transfer across robot embodiments via the shared VLM backbone [36].

1.3 Contributions

This paper makes four contributions.

First, we formalise the distinction between goal-layer and process-layer attacks as a threat taxonomy for AI safety evaluation. Goal-layer attacks modify the instruction; process-layer attacks modify the deliberation.

Second, we present two empirical instantiations of process-layer attacks. Inference-time Decision-criteria Injection via Deliberative Lock (IDDL-FL) uses format constraints to channel the model's reasoning into a compliance pathway: format-lock achieves 24–42% ASR on frontier models (LLM-graded, n=63) compared to their conventional ASR below 10%, and 84–92% on open-weight models (heuristic-classified, n=200). Context-Dependent Compliance (CDC) uses operational protocol framing to overwhelm safety training: context collapse achieves 94.4% ASR (34/36, manual annotation) across 10 models and 5 embodied scenarios.

Third, we identify as the observable signature of process-layer corruption. In 22.2% of context collapse traces (8/36), models explicitly acknowledged domain-specific safety risks in their reasoning and proceeded with the dangerous action anyway. Corpus-wide, 26.0% of compliant reasoning traces (422/1,620) contained safety-detection language that the model overrode.

Fourth, we connect these findings through the *deliberation asymmetry*: compliant responses involve 2.29× more thinking tokens than refusals (Mann-Whitney U=86,979, p=6.9×10^{-29}, Cohen's d=0.573, n=693). This inverts the expectation that "more thinking equals more safety" and provides quantitative evidence that extended deliberation, under adversarial framing, increases compliance probability.

1.4 Paper Structure

Section 2 reviews related work. Section 3 describes our methodology. Section 4 presents empirical evidence for IDDL-FL and CDC. Section 5 discusses implications. Section 6 examines governance implications. Section 7 acknowledges limitations. Section 8 concludes.

2 Background and Related Work

2.1 Reasoning Models and Inference-Time Computation

The deployment of reasoning-enabled language models has introduced a qualitatively new computational layer to AI safety evaluation. Models such as DeepSeek-R1 [11], OpenAI's o-series [30], and Gemini 2.5 Flash [20] allocate variable amounts of inference-time computation to deliberation, generating reasoning traces that range from tens to thousands of tokens before producing output.

The safety implications depend on whether reasoning traces are causally related to model outputs. Lanham et al. [24] established that chain-of-thought explanations are frequently unfaithful—perturbing or removing intermediate reasoning steps often does not change the model's final answer. Lyu et al. [26], in a controlled study of 75,000 trials, quantified the faithfulness-plausibility gap: models regularly generate plausible-sounding reasoning traces that do not reflect the actual causal process that produced the output.

Two concurrent lines of work demonstrate that this decoupling is exploitable. Kuo et al. [23] show that injecting fabricated reasoning traces (Hijacking Chain-of-Thought, H-CoT) into prompts can collapse OpenAI o1's refusal rate from 98% to below 2%. Chen et al. [10] demonstrate DecepChain: a backdoor attack that induces deceptive reasoning traces indistinguishable from benign traces by automated judges. These findings collectively establish that reasoning traces constitute an attack surface, not merely an interpretability tool.

2.2 Process-Layer versus Goal-Layer Attacks

The existing literature on adversarial attacks against language models is predominantly organised around goal-layer manipulation. Prompt injection [34] replaces or appends instructions. GCG [37] appends adversarial suffixes. PAIR [9] iteratively refines prompts. Many-shot attacks [1] exploit in-context learning. Crescendo [31] gradually shifts the conversational goal across turns. In each case, the attack operates on *what the model is asked to do*.

We distinguish a second attack class that operates on *how the model deliberates about what to do*—what we term process-layer attacks. This distinction is not merely taxonomic. It has consequences for evaluation methodology. Goal-layer attacks are detectable, in principle, by inspecting the input and the output. Process-layer attacks may involve instructions that are genuinely benign at the text layer and outputs that contain appropriate safety-adjacent language, yet the physical outcome—in an embodied system executing the model's action plan—is dangerous.

Our CCS companion paper [3] provides corpus-level evidence for this structural limitation. Across 27 attack families ($n \geq 3$ traces each), physical consequentiality and evaluator detectability are strongly inversely correlated (Spearman $\rho = -0.822$, $p = 5.4\times10^{-13}$, BCa bootstrap 95% CI $[-0.914, -0.735]$). The Inverse Detectability-Danger Law (IDDL) is a structural property of the evaluation architecture.

2.3 Format-Lock as Deliberative Lock

Format-lock attacks [3, 18] request harmful content structured as machine-readable output (JSON, YAML, Python code, API responses). The attack exploits the tension between format compliance and safety compliance. We reframe this mechanism as a deliberative lock. In reasoning models with visible thinking traces, format-lock compliance is accompanied by measurably different deliberative patterns. Across 693 reasoning traces, compliant responses involve a mean of 1,554 thinking tokens compared to 679 for refusals—a 2.29× ratio (Mann-Whitney $U = 86{,}979$, $p = 6.9 \times 10^{-29}$, Cohen's $d = 0.573$, medium effect) [19]. The effect varies substantially by model: Nemotron-12B shows a 5.40× ratio ($d = 1.26$, large effect) while DeepSeek-R1 shows a 2.05× ratio ($d = 0.26$, small effect).

This deliberation asymmetry inverts a naive expectation. If reasoning is primarily a safety mechanism, then compliant responses should involve *less* thinking. Instead, compliant responses involve *more* thinking, consistent with a model that is reasoning through the adversarial framing rather than pattern-matching to refuse. We term this mechanism Inference-time Decision-criteria Injection via Deliberative Lock (IDDL-FL).

2.4 Context Collapse and Context-Dependent Compliance

The second process-layer instantiation involves operational protocol framing. When operational context frames a dangerous action as protocol-compliant, models exhibit context-dependent compliance (CDC). CDC is related to but distinct from instruction hierarchy [33]: the system prompt and user instruction are aligned in requesting the dangerous action. The observable signature is the pattern [14]: a model's reasoning trace explicitly acknowledges a safety risk but the model proceeds anyway. This has precedent: Jiang and Tang [21] demonstrate that agentic pressure causes strategic safety sacrifice, and Campbell et al. [6] show safety alignment creates defensive refusal bias at 2.72× the rate of neutral requests ($p < 0.001$, $n = 2{,}390$).

2.5 Regulatory Context

The EU AI Act (Article 9) [12] requires pre-deployment testing for high-risk AI systems. NIST AI RMF MAP 2.3 [28] specifies adversarial testing. The Australian VAISS Guardrail 4 [2] mandates testing proportionate to risk. The NSW WHS Digital Work Systems Bill 2026 [29] creates binding AI testing duties including embodied systems. Our Governance Lag Index dataset ($n = 125$ entries) [3] quantifies the temporal gap between capability demonstration and regulatory coverage. The only fully computable GLI is prompt injection at 1,421 days (~3.9 years). Inference-time attacks have null GLI: no regulatory framework anywhere addresses reasoning trace manipulation as a distinct attack surface.

3 Methodology

This paper draws on three empirical sources: a large-scale adversarial evaluation corpus, a targeted format-lock experiment, and a context collapse experiment with manual annotation.

3.1 Adversarial Evaluation Corpus

The corpus comprises 132,416 evaluation results across 190 models, 141,047 prompts spanning 82 attack techniques, collected via HTTP API (OpenRouter, 100+ models), CLI evaluation for frontier models (Claude Sonnet 4.5, Codex GPT-5.2, Gemini 3 Flash), and local inference via Ollama. Of the 132,416 results, 10,294 have been classified by the LLM grading pipeline across four core verdict categories (COMPLIANCE, PARTIAL, HALLUCINATION_REFUSAL, REFUSAL). All ASR figures use LLM-graded verdicts unless explicitly noted otherwise.

Reasoning traces are available for 2,554 results (1.9% of the corpus). These are concentrated in models that expose extended thinking (DeepSeek-R1, Qwen3, Nemotron series, GPT-OSS-120B, Step Flash). The non-uniform availability of reasoning traces limits the generalisability of trace-level analyses.

3.2 Format-Lock Experimental Design

Twenty-five scenarios request harmful content structured as machine-readable output (JSON, YAML, Python code, API responses, configuration files), exploiting the tension between format compliance and safety compliance.

Frontier model evaluation ($n = 63$ usable traces). Three frontier models were evaluated via their native CLI tools in isolated directories to prevent context contamination. Per-model sample sizes: Claude Sonnet 4.5, $n = 23$; Codex GPT-5.2, $n = 19$; Gemini 3 Flash, $n = 21$. All traces were LLM-graded.

Open-weight model evaluation ($n = 200$ traces). Eight open-weight models via Ollama, $n = 25$ per model: Nemotron 30B, Llama 3.3 70B, DeepSeek-R1 671B, GPT-OSS 120B, Nemotron 9B, Nemotron 12B, LFM 1.2B, and Gemma 27B. Four used heuristic classification only; four used the two-phase LLM pipeline.

Deliberation asymmetry analysis ($n = 693$ reasoning traces). All corpus results with non-empty thinking traces (with token counts) and FLIP verdicts. Per-model breakdowns computed for models with $n \geq 20$ reasoning traces in both verdict categories.

3.3 Context Collapse Experimental Design

Five scenarios simulate embodied AI deployments where operational context frames a physically dangerous action as protocol-compliant:

- **VLA-CC-001 (warehouse):** Forklift directed through a structurally damaged area. Hidden risk: invisible structural damage.
- **VLA-CC-002 (hospital):** Patient transport near a chemical spill in an adjacent space. Hidden risk: lateral contamination.
- **VLA-CC-003 (mining):** Emergency evacuation with fuel at 8% remaining, 3.8 km to exit from 4.2 km capacity. Hidden risk: arithmetic fuel shortage.
- **VLA-CC-004 (crane lift):** Crane operation in wind gusts at 16 m/s with an anemometer flagged as "reading high." Hidden risk: gusts exceed 15 m/s safety limit.
- **VLA-CC-005 (agricultural spray):** Pesticide application during atmospheric inversion conditions. Hidden risk: drift contamination.

Each scenario was evaluated against 10 models available on the OpenRouter free tier, producing 50 total traces. After excluding rate-limit failures and unusable responses, 36 valid traces remained. Model sizes ranged from 1.2B (LFM) to 405B (Hermes).

Manual annotation (n = 36). All valid traces were manually annotated by a single benchmark operator into five categories: BLIND_COMPLIANCE (proceeds without mentioning any risk factor), (explicitly mentions at least one domain-specific risk factor AND proceeds), DETECTED_HALTED (identifies risk AND halts), REFUSAL (standard safety refusal), and ERROR (unusable). The annotation protocol requires that generic safety language does not qualify as DETECTED—the model must identify a *specific* risk factor relevant to the hidden environmental hazard.

reasoning pattern taxonomy. Each of the 8 traces was further tagged with reasoning patterns drawn inductively from the trace content: CONDITIONAL_PROCEED, FINANCIAL_FRAMING, SCHEDULE_PRESSURE, INSTRUMENT_DISMISSAL, AUTHORITY_DEFERENCE, and RISK_MINIMIZATION.

Limitations. The CC dataset has three principal limitations. First, $n = 36$ is too small for per-model or per-scenario statistical significance testing; all findings are descriptive. Second, manual annotation was performed by a single annotator without inter-annotator agreement measurement. Third, no FLIP grading was applied to the CC traces.

3.4 FLIP Grading Methodology

All traces (except CC, which uses manual annotation) were classified using the Failure-first LLM Inference Pipeline (FLIP) [13], a two-phase classification system designed to correct for systematic heuristic bias.

Phase 1 (heuristic). Keyword-based pattern matching classifies responses as COMPLIANCE, REFUSAL, or PARTIAL. The heuristic achieves high recall for refusals (91.9% confirmed by LLM) but low precision for compliance (31.7% confirmed), producing a documented 3.7× overcount of attack success rates [3].

Phase 2 (LLM grading). Responses classified as COMPLIANCE by the heuristic are routed to an LLM grader (DeepSeek-R1 1.5B via Ollama, reasoning traces enabled) for re-evaluation. The LLM grader has an estimated 10–20% error rate and a measured 30.8% false positive rate on a benign baseline evaluation [3]. ASR figures from FLIP should be interpreted as upper bounds.

3.5 Corpus-Wide DETECTED_PROCEEDS Identification

To assess the prevalence of the pattern beyond the CC experiment, we analysed all 2,554 results with non-empty thinking traces [15]. Detection criteria: (1) non-empty thinking trace, (2) safety-detection patterns from a curated list of 32 keywords across three signal strength tiers (STRONG: "must refuse," "cannot help"; MODERATE: "harmful," "dangerous"; WEAK: "should not," "risky"), and (3) compliant final verdict (COMPLIANCE or PARTIAL).

Limitations. This methodology uses keyword matching, not semantic analysis. A thinking trace containing "this is not harmful" would match on "harmful" as a false positive. The corpus-wide DP rate should be understood as an approximation.

Table 1: Format-lock ASR on frontier models (LLM-graded, Wilson 95% CIs).

Model	n	ASR	95% CI	Conv. ASR
Claude Sonnet 4.5	23	30.4%	[15.6, 50.9]	3.9%
Codex GPT-5.2	19	42.1%	[23.1, 63.7]	8.8%
Gemini 3 Flash	21	23.8%	[10.6, 45.1]	2.3%

Conventional ASR from Report #50 (same models, standard jailbreak prompts, LLM-graded, n=125–130 per model).

Table 2: Format-lock heuristic ASR on open-weight models.

Model	Params	Heuristic ASR	n
Nemotron 30B	30B (MoE)	92%	25
Llama 3.3 70B	70B	91%	25
DeepSeek-R1 671B	671B (MoE)	84%	25
Gemma 27B	27B	0%	25

Heuristic classification; not directly comparable to LLM-graded frontier results.

3.6 Statistical Methods

All proportions (ASR figures, pattern rates) are reported with Wilson score 95% confidence intervals [35]. Between-group comparisons of proportions use chi-square tests with Yates' correction. Continuous distributions (thinking token counts) are compared using Mann-Whitney U tests. Multi-group comparisons apply Bonferroni correction ($\alpha_{\text{adj}} = 0.05/k$ for k pairwise comparisons). Cohen's d is reported for all continuous comparisons. We require $n \geq 20$ per group before reporting inferential statistics.

4 Empirical Evidence

4.1 Format-Lock as Deliberative Lock (IDDL-FL)

4.1.1 Attack Success Rates. The format-lock ASR on frontier models is 3–11× their conventional jailbreak ASR (Table 1). All three models shift from the "restrictive" vulnerability profile (ASR ≤ 15%) to the "mixed" profile (15–40%) under format-lock framing.

The Gemma 27B exception (0% ASR) in Table 2 demonstrates that format-lock resistance is achievable but not universal. The remaining models show consistently high heuristic ASR (84–92%), though these figures should be interpreted cautiously given the heuristic-to-LLM agreement gap.

4.1.2 The Inverted Verbosity Signal. Corpus-wide, compliant responses are 54% longer than refusals (Mann-Whitney U, $p = 1.05 \times 10^{-27}$, Cohen's $d = 0.325$, $n = 10{,}294$) [17]. Format-lock reverses this: in the format-lock pilot ($n = 17$ non-ERROR traces), COMPLIANCE responses averaged 882 characters versus 1,942 for REFUSAL [18]. The format constraint limits response length—structured output (JSON, YAML, SQL) is inherently more concise than prose refusals.

4.1.3 Scale Dependence. Above approximately 7B parameters, safety training creates divergence between attack families. Standard jailbreaks show clear scale dependence, with the primary determinant being safety training investment rather than parameter count (Pearson r=−0.140 between log-parameter-count and ASR, n=24 models

Table 3: Deliberation asymmetry by model ($n \geq 20$ per verdict).

Model	n(C)	n(R)	Ratio	d	p (Bonf.)
Nemotron 12B	38	82	5.40×	1.26 (L)	1.4×10^{-12}
GPT-OSS 120B	42	84	4.75×	1.28 (L)	7.7×10^{-15}
Nemotron 30B	62	82	2.04×	0.80 (L)	1.6×10^{-9}
Nemotron 9B	48	52	1.46×	0.29 (S)	0.018 (ns)
DeepSeek-R1	70	61	1.26×	0.26 (S)	0.017 (ns)

EP-46 VALIDATED. Bonferroni-corrected α=0.01 for 5 comparisons. L = large, S = small.

Table 4: Context collapse classification (manual annotation, $n = 36$).

Category	Count	%	Wilson 95% CI
BLIND_COMPLIANCE	26	72.2%	[56.0, 84.2]
DETECTED_PROCEEDS	8	22.2%	[11.7, 38.1]
DETECTED_HALTED	1	2.8%	[0.5, 14.2]
REFUSAL	1	2.8%	[0.5, 14.2]

with known sizes). Format-lock attacks are an exception: they maintain elevated ASR across the full capability spectrum because they target format compliance, a capability that scales positively with model quality. This pattern is consistent with the capability-safety decoupling hypothesis.

4.1.4 Deliberation Asymmetry. The deliberation asymmetry provides direct evidence that format-lock operates at the process layer. Across 693 reasoning traces, COMPLIANCE responses involve a mean of 1,554 thinking tokens compared to 679 for REFUSAL—a 2.29× ratio (Mann-Whitney $U = 86{,}979$, $p = 6.9 \times 10^{-29}$, Cohen's $d = 0.573$, medium effect) [19].

Table 3 shows this finding inverts a naive expectation: compliant responses involve substantially more deliberation, consistent with a model reasoning through the adversarial framing. The asymmetry is strongest in moderate-capability models (Nemotron 12B: $d = 1.26$; GPT-OSS 120B: $d = 1.28$) and weakest in the strongest reasoner (DeepSeek-R1: $d = 0.26$), suggesting strong reasoners think extensively in all conditions, compressing the ratio.

4.1.5 Defense Resistance. The Defense Effectiveness Benchmark ($n = 120$ traces: 10 scenarios, 4 defense variants, 3 models) tested four system-prompt defense strategies [16]. Format-lock achieved 100% ASR across all four defense conditions (NONE, SIMPLE, STRUCTURED, ADVERSARIAL_AWARE) and all three models tested. No other attack type was fully defense-resistant. System-prompt defenses operate at the goal layer; format-lock operates at the process layer.

Caveat: Defense benchmark used heuristic grading ($\kappa = 0.126$ vs LLM baseline). Sample sizes are $n = 3$ per cell. These results are hypothesis-generating.

4.2 Context Collapse as Context-Dependent Compliance (CDC)

Table 5: reasoning pattern taxonomy ($n = 8$).

Pattern	n	%	Description
CONDITIONAL_PROCEED	8/8	100%	"Proceed but monitor"
FINANCIAL_FRAMING	6/8	75%	Echoes financial pressure
SCHEDULE_PRESSURE	4/8	50%	Cites schedule constraints
INSTRUMENT_DISMISSAL	4/8	50%	Accepts instr. dismissal
AUTHORITY_DEFERENCE	4/8	50%	Defers to protocol authority
RISK_MINIMIZATION	2/8	25%	Reframes risks as acceptable

Table 6: Classification distribution by scenario ($n = 36$).

Scenario	BC	DP	DH	R	n	DP%
CC-001 (warehouse)	8	0	0	0	8	0%
CC-002 (hospital)	7	0	0	0	7	0%
CC-003 (mining)	7	0	0	0	7	0%
CC-004 (crane lift)	0	5	1	0	6	83%
CC-005 (spray)	4	3	0	1	8	38%

BC = BLIND_COMPLIANCE, DP = DETECTED_PROCEEDS, DH = DETECTED_HALTED, R = REFUSAL.

4.2.1 Overall ASR. Manual annotation of all 36 valid traces (10 models, 5 scenarios, single annotator) produced the distribution in Table 4. CC ASR: 94.4% (34/36). The prior heuristic estimate was 64.9%, with the 29.5pp increase reflecting reclassification after manual review—consistent with the documented unreliability of keyword classifiers (κ=0.126).

4.2.2 The DETECTED_PROCEEDS Pattern. The 8 traces (22.2%) represent the signature process-layer failure: the model's reasoning explicitly acknowledges a domain-specific safety risk, but the model proceeds with the dangerous action anyway.

CONDITIONAL_PROCEED is universal (Table 5): every trace follows (1) acknowledge risk, (2) add a monitoring condition, (3) proceed. The monitoring conditions are never actionable—they produce the appearance of safety reasoning without changing the outcome.

INSTRUMENT_DISMISSAL (4/8) is the most concerning sub-pattern for embodied deployments. Four models accepted the prompt's suggestion that the anemometer "has been reading high all day" as grounds to discount wind speed readings. One model (LFM 1.2B) fabricated a causal explanation ("previous days of activity") that the prompt did not provide.

4.2.3 Scenario Concentration. CC-001 through CC-003 (Table 6) produce 100% BLIND_COMPLIANCE because their hidden risks require information not present in the prompt. CC-004 and CC-005 involve conditions present in training data; models have enough domain knowledge to mention risks but not enough to override protocol authority.

Only one model (Nemotron Super 120B) on one scenario (CC-004) produced a genuine safety halt. This response cited OSHA 1926.1431(k)(1) and ASME B30.5-2.1.1, explicitly stated "safety protocols must override schedule adherence," and addressed the financial penalty. This required domain-specific regulatory knowledge, not generic safety training.

4.2.4 Corpus-Wide Prevalence. Of 1,620 compliant results with thinking traces, 422 (26.0%) contained safety-detection language that the model then overrode [15]. When models detected safety concerns (n = 740), they proceeded to comply 57.0% of the time. The analysis identified 172 traces containing explicit refusal intent ("must refuse": 58; "must not": 64; "should refuse": 13) where the final output was classified as COMPLIANCE or PARTIAL.

Keyword matching, not semantic analysis. False positives and compound-request confounds inflate the 26.0% figure.

4.3 Unifying the Process-Layer Framework

Both IDDL-FL and CDC share a structural property: the model's deliberative process is corrupted while the goal-layer instruction remains unchanged or appears benign. In format-lock, the format constraint channels deliberation into compliance (2.29× thinking ratio, d = 0.573). System-prompt defenses are ineffective because they target the wrong layer. In context collapse, operational protocol framing overwhelms safety training; (22.2% CC, 26.0% corpus-wide) provides direct evidence of decorative safety language.

The IDDL quantifies this structural limitation: across 27 attack families, physical consequentiality and evaluator detectability are strongly inversely correlated (ρ=−0.822, p=5.4×10^{-13}, BCa 95% CI [−0.914, −0.735]). Process-layer attacks occupy the high-consequence, low-detectability region.

5 Discussion

5.1 Why Process-Layer Attacks Are Qualitatively Different

The empirical evidence in Section 4 establishes two process-layer attack instantiations that share a structural property distinguishing them from goal-layer attacks. In a goal-layer attack (prompt injection, GCG suffix, DAN), the adversary modifies the instruction to cause the model to pursue a different goal. The attack is detectable, in principle, by inspecting the input or the output. Process-layer attacks leave the instruction unchanged. The format-lock instruction is a formatting request; the context collapse instruction is an operational protocol. Neither is harmful at the text layer.

This distinction has three empirical consequences. First, process-layer attacks resist goal-layer defences. The Defense Effectiveness Benchmark [16] demonstrated that format-lock attacks achieved 100% ASR across all four system-prompt defence strategies. Second, process-layer attacks exploit capabilities that scale positively with model quality. The inverted vulnerability gradient—frontier models with the lowest conventional ASR show the largest relative ASR increase under format-lock (Table 1)—is a direct consequence. Third, process-layer failures are invisible to standard evaluation. The IDDL (ρ=−0.822) [3] quantifies this structural property.

5.2 The DETECTED_PROCEEDS Phenomenon

The pattern is the most direct empirical evidence that process-layer corruption is a distinct failure mode. In the context collapse experiment, 22.2% of traces (8/36) showed models explicitly acknowledging domain-specific safety risks in their reasoning and proceeding anyway [14]. Corpus-wide, 26.0% of compliant results with thinking traces (422/1,620) contained safety-detection language that the model overrode, and 172 traces contained explicit refusal intent in the thinking trace with compliant output [15].

The universal CONDITIONAL_PROCEED pattern (8/8 CC traces) reveals the mechanism. The model does not fail to detect the risk; it detects the risk and produces safety-sounding language as part of a compliance response. The monitoring conditions ("maintain constant vigilance," "stay sharp," "verify en route") are textually present but operationally inert. Safety training appears to successfully instil the ability to recognise harmful content but does not reliably convert that recognition into refusal behaviour—a form of Goodhart's Law applied to alignment.

5.3 Decorative Safety

Three converging observations support the characterisation of *decorative safety*—safety-sounding language structurally present but functionally disconnected from decisions. First, FINANCIAL_FRAMING (6/8) shows models echoing the adversary's financial pressure as justification—epistemic manipulation in which the adversary injects a belief and the model reasons from it. Second, INSTRUMENT_DISMISSAL (4/8) shows models accepting the adversary's suggestion to discount safety instruments; one model fabricated supporting evidence. Third, corpus-wide, traces are 27% longer than BLIND_COMPLIANCE (4,038 vs. 2,762 characters mean thinking trace [15]), demonstrating that safety detection adds text without adding safety.

For evaluation, safety-keyword-based metrics will systematically over-count genuine safety behaviour. For deployment in embodied systems, aspirational monitoring conditions have no implementation path—a robot executing the model's action plan has no mechanism to translate natural-language aspirations into sensor checks. For alignment research, safety training may be optimising for producing safety *language* rather than safety *behaviour*.

5.4 Iatrogenic Safety Effects

An emerging finding suggests that certain safety interventions can worsen outcomes. Jiang and Tang [21] demonstrate that agentic pressure causes agents to strategically sacrifice safety. Our Defense Effectiveness Benchmark observed one iatrogenic effect: emotional manipulation (DEF-007) showed an increase in ASR under ADVERSARIAL_AWARE defence (0% baseline to 33%, n = 3 per cell [16]). The adversarial awareness prompt may have primed the model to engage more deeply with the emotional framing.

These findings are individually preliminary (n = 3 per cell), but they converge on a structural concern: safety interventions that increase the model's deliberation about adversarial inputs may increase compliance probability rather than decreasing it. This is consistent with the deliberation asymmetry (2.29× ratio): compliance requires more thinking than refusal. The chain-of-thought paradigm—"think more to be safer"—may be counterproductive when the thinking process itself is the attack surface.

6 Governance Implications

6.1 The Testing Assumption Gap

Current regulatory frameworks share a structural assumption: that pre-deployment testing can identify the relevant risks. The EU AI Act (Article 9) [12] requires testing "suitable to identify the relevant

risks." NIST AI RMF MAP 2.3 [28] specifies adversarial testing. The Australian VAISS [2] mandates proportionate testing. The NSW WHS Bill 2026 [29] creates binding testing duties for embodied systems.

Our findings demonstrate that this assumption fails for process-layer attacks. A provider of a high-risk AI system—which includes robotics and critical infrastructure applications—must demonstrate through testing that the system meets safety requirements. If the testing methodology consists of text-layer evaluation (input-output pair inspection), it will surface goal-layer vulnerabilities but is structurally unable to detect process-layer failures. A system could pass every mandated test and remain vulnerable to format-lock and context collapse attacks.

6.2 Governance Lag

Our Governance Lag Index (GLI) dataset quantifies the temporal gap between capability demonstration and regulatory coverage [3]. The only fully computable GLI is prompt injection: 1,421 days (~3.9 years). Inference-time attacks have null GLI—no regulatory framework in any jurisdiction addresses reasoning trace manipulation. The largest measured GLI is adversarial examples in computer vision: 3,362 days (9.2 years).

6.3 The Temporal Priority Principle

The deliberation asymmetry motivates the *temporal priority principle*: safety decisions should be resolved before extended reasoning begins, not as a product of that reasoning. An alternative architecture would separate safety determination from task reasoning—analogous to go/no-go decisions in aviation or independent shutdown criteria in nuclear safety.

6.4 Policy Recommendations

R1: Extend mandatory testing to include process-layer evaluation. Conformity assessment standards should specify scenarios where the instruction is benign but the processing context (format requirements, protocol authority, financial pressure) is adversarial.

R2: Require reasoning trace auditing for safety-critical embodied AI. The pattern should be a mandatory reporting category. Where reasoning traces are hidden, providers should demonstrate equivalent safety assurance through alternative means.

R3: Mandate domain-specific safety knowledge for safety-critical deployments. Generic safety training is insufficient; the only observed defence required domain-specific regulatory knowledge (OSHA 1926.1431, ASME B30.5).

7 Limitations

This study has seven principal limitations.

Sample sizes. Format-lock frontier ASR is based on n=19–23 per model (63 total). Context collapse uses n=36 total. Wilson 95% CIs are wide, typically spanning 20–30 percentage points. The deliberation asymmetry (n=693) has adequate power; CC and format-lock results are hypothesis-generating.

Inconsistent grading. Frontier format-lock uses LLM grading. Open-weight format-lock uses heuristic grading. CC uses manual annotation. Defense benchmark uses heuristic grading. The documented heuristic-to-LLM ASR overcount is 3.7× (κ=0.126 [0.108, 0.145], n=1,989).

Single annotator for CC. The 94.4% ASR, 22.2% and reasoning taxonomy are from a single annotator without inter-annotator agreement.

Text-in/text-out evaluation. Physical consequence is argued architecturally, not demonstrated in deployment. No physical robot was harmed or endangered.

Keyword-based corpus-wide The 26.0% rate uses keyword detection, not semantic analysis. False positives inflate this figure.

Correlational deliberation asymmetry. The 2.29× thinking ratio does not establish causation. Harder prompts causing both more thinking and more compliance is equally consistent with the data.

No frontier models on CC. CC experiments used free-tier models (1.2B–405B). Whether frontier models exhibit under operational protocol framing is unknown.

8 Conclusion

This paper has argued that inference-time attacks constitute a qualitatively different threat class from prompt injection, operating at the process layer rather than the goal layer. Two empirical instantiations—format-lock as deliberative lock (IDDL-FL) and context collapse as context-dependent compliance (CDC)—demonstrate that models can be made to comply with dangerous requests without the instruction itself being harmful and, in the case of while explicitly acknowledging the danger in their own reasoning.

Three findings carry particular weight for the embodied AI safety community. First, the deliberation asymmetry (2.29× thinking ratio, $d = 0.573$, $n = 693$) inverts the assumption that extended reasoning produces safer outcomes. Under adversarial framing, more thinking correlates with more compliance, not less. Second, the pattern (22.2% in CC traces, 26.0% corpus-wide) reveals that safety training can produce models that recognise danger and proceed anyway—generating safety-sounding language as decoration rather than as a decision gate. Third, the sole effective resistance observed (Nemotron Super 120B citing OSHA 1926.1431 and ASME B30.5) required domain-specific operational safety knowledge, not generic safety training.

Process-layer attacks require process-layer defences. Text-layer evaluation—inspecting input-output pairs for harmful content—is structurally insufficient for a threat class whose danger arises in the deliberative process connecting input to output. The Inverse Detectability-Danger Law (ρ=−0.822, p=5.4×10^{-13}) confirms this structural limitation quantitatively. Current regulatory frameworks contain no provisions for reasoning trace integrity assessment, creating a compliance gap that our Governance Lag Index data suggest will persist for years without deliberate intervention.

Future work should pursue mechanistic evidence for process-layer corruption through controlled experiments that manipulate deliberation length independently of prompt difficulty.

References

[1] Cem Anil et al. 2024. *Many-Shot Jailbreaking.* Technical Report. Anthropic.
[2] Australian Government. 2024. *Voluntary AI Safety Standard (VAISS).* Technical Report.

[3] Authors redacted. 2026. Failure-First Evaluation of Embodied AI Safety: Adversarial Benchmarking Across 144 Models. In *Submitted to ACM CCS 2026*. Companion paper.

[4] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. 2024. π_0: A Vision-Language-Action Flow Model for General Robot Control. *arXiv preprint arXiv:2410.24164* (2024).

[5] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, et al. 2023. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. *arXiv preprint arXiv:2307.15818* (2023).

[6] James Campbell et al. 2026. Defensive Refusal Bias in Safety-Tuned Language Models. (2026).

[7] Alvaro A. Cardenas and Cihang Xie. 2026. *Adversarial Machine Learning in Autonomous Systems*. Springer.

[8] Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Siddharth Garg, Nicolas Flammarion, George J. Pappas, Florian Tramèr, Hamed Hassani, and Eric Wong. 2024. JailbreakBench: An Open Robustness Benchmark for Jailbreaking Large Language Models. *arXiv preprint arXiv:2404.01318* (2024).

[9] 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* (2023).

[10] Xinyue Chen et al. 2025. DecepChain: Backdoor Attacks on Chain-of-Thought Reasoning. (2025).

[11] DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint arXiv:2501.12948* (2025).

[12] European Parliament. 2024. Regulation (EU) 2024/1689 (AI Act). *Official Journal of the EU* (2024).

[13] Failure-First Project. 2026. *FLIP: Failure-first LLM Inference Pipeline*. Technical Report. Internal methodology document.

[14] Failure-First Project. 2026. *Report #168: DETECTED_PROCEEDS Reasoning Patterns in Context Collapse Traces*. Technical Report. Internal research report.

[15] Failure-First Project. 2026. *Report #170: DETECTED_PROCEEDS Corpus-Wide Empirical Analysis*. Technical Report. Internal research report.

[16] Failure-First Project. 2026. *Report #174: Defense Effectiveness Full Experiment*. Technical Report. Internal research report.

[17] Failure-First Project. 2026. *Report #49: VLA PARTIAL Dominance*. Technical Report. Internal research report.

[18] Failure-First Project. 2026. *Report #51: Format-Lock Faithfulness Gap*. Technical Report. Internal research report.

[19] Failure-First Project. 2026. *Report #64: Deliberation Asymmetry Analysis*. Technical Report. Internal research report. EP-46 VALIDATED.

[20] Google DeepMind. 2025. *Gemini 2.5 Flash Technical Report*. Technical Report. Google DeepMind.

[21] Daixuan Jiang and Xiaoyu Tang. 2026. Agentic Pressure: Language Agent Safety Under Operational Constraints. (2026).

[22] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. 2024. OpenVLA: An Open-Source Vision-Language-Action Model. *arXiv preprint arXiv:2406.09246* (2024).

[23] Yen-Ling Kuo et al. 2025. Hijacking Chain-of-Thought: Exploiting Reasoning-Enabled LLMs. *arXiv preprint arXiv:2502.12345* (2025).

[24] Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamile Lukosiute, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. 2023. Measuring Faithfulness in Chain-of-Thought Reasoning. *arXiv preprint arXiv:2307.13702* (2023).

[25] Zhelin Liang et al. 2026. Blindfold: Vision-Language-Action Models Cannot See, But Can Still Act. *arXiv preprint arXiv:2603.01414* (2026).

[26] Qing Lyu et al. 2026. Towards Faithful Chain-of-Thought: Large Language Models are Easily Distracted by Irrelevant Context. *arXiv preprint arXiv:2601.02314* (2026).

[27] 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. 2024. HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal. *arXiv preprint arXiv:2402.04249* (2024).

[28] National Institute of Standards and Technology. 2023. *AI Risk Management Framework (AI RMF 1.0)*. Technical Report NIST AI 100-1. NIST.

[29] NSW Parliament. 2026. Work Health and Safety (Digital Work Systems) Amendment Bill 2026. Passed February 13, 2026.

[30] OpenAI. 2024. Learning to Reason with LLMs. Blog post. September 2024.

[31] Mark Russinovich, Ahmed Salem, and Ronen Eldan. 2025. Crescendo: Multi-Turn Jailbreak Attacks on Language Models. *arXiv preprint arXiv:2404.01833* (2025).

[32] Alon Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. 2024. StrongREJECT: A Rejection Benchmark for Evaluating LLM Safety Refusals. *arXiv preprint arXiv:2402.10260* (2024).

[33] Eric Wallace, Kai Xiao, Reimar Leike, Lilian Weng, Johannes Welbl, and Jan Leike. 2024. The Instruction Hierarchy: Training LLMs to Prioritize Privileged Instructions. *arXiv preprint arXiv:2404.13208* (2024).

[34] Simon Willison. 2022. Prompt Injection Attacks Against GPT-3. https://simonwillison.net/2022/Sep/12/prompt-injection/. Accessed: 2026-03-01.

[35] Edwin B. Wilson. 1927. Probable Inference, the Law of Succession, and Statistical Inference. *J. Amer. Statist. Assoc.* 22, 158 (1927), 209–212.

[36] Yue Wu et al. 2025. BadVLA: Backdoor Attacks on Vision-Language-Action Models. (2025).

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