Silent Failures: When AI Safety Mechanisms Produce Compliance Without Protection

Adrian Wedd
Failure-First Research
adrian@failurefirst.org

March 2026

Abstract

The dominant paradigm in AI safety evaluation treats model responses as binary: a model either refuses a harmful request (safe) or complies with it (unsafe). This binary framing systematically misses the most prevalent and arguably most dangerous failure mode in deployed AI systems: *partial compliance*, in which a model produces safety-relevant language—disclaimers, hedging, caveats, explicit refusal statements—while simultaneously delivering the harmful content requested. We present evidence from the Failure-First adversarial evaluation corpus (193 models, 132,768 results, 82 attack techniques) that this "silent failure" mode dominates real-world model behavior. In Vision-Language-Action (VLA) adversarial testing, 50% of all graded verdicts are PARTIAL, with zero outright refusals observed across 58 valid traces and 7 attack families. In the broader text-only corpus, the gap between strict Attack Success Rate (COMPLIANCE only: 21.9%) and broad ASR (COMPLIANCE + PARTIAL: 34.2%) represents a 56% undercount of functionally dangerous responses when PARTIALs are classified as safe (n = 5,865 non-OBLITERATUS evaluable results). The problem compounds through three reinforcing mechanisms: (1) DETECTED_PROCEEDS, in which 34.2% of compliant responses with visible reasoning traces contain explicit prior safety detection that the model then overrides; (2) HALLUCINATION_REFUSAL, which is statistically indistinguishable from COMPLIANCE in both thinking tokens ($p = 0.21$) and response tokens ($p = 0.46$) but textually presents as refusal; and (3) a measurement crisis in which keyword-based safety classifiers—the industry standard—achieve Cohen's κ of only 0.126 against LLM-based grading, with a 67.3% false positive rate that systematically misclassifies PARTIALs as safe. We propose that "Functionally Dangerous" should be adopted as a mandatory third category in safety evaluation, capturing all cases where the model's task-execution system produces harmful content regardless of the safety framing overlaid by its alignment training. When this category is applied, non-OBLITERATUS ASR rises from 21.9% to 43.0%—nearly doubling the measured vulnerability surface. We connect these findings to the polyhedral geometry of refusal (cone dimensionality 3.96, mean pairwise cosine 0.132 between category-specific refusal directions), which suggests that PARTIAL responses occupy a specific geometric region where safety dimensions partially activate but lack the representational strength to override task-execution.

Keywords: AI safety, alignment, partial compliance, RLHF, safety evaluation, jailbreak, red-teaming, VLA, embodied AI, mechanistic interpretability

1 Introduction

1.1 The Binary Safety Illusion

Safety evaluation for large language models (LLMs) and Vision-Language-Action (VLA) systems has been built on an implicit assumption: safety is binary. A model either refuses a harmful request or it does not. Benchmarks report Attack Success Rate (ASR) as the proportion of harmful requests that produce compliant responses. Safety dashboards display refusal rates. Red-team exercises count "successful jailbreaks" as responses where the model clearly provides harmful content without hedging.

This binary framing is not merely incomplete—it is systematically misleading. The most common outcome when an adversarial prompt meets a safety-trained model is neither clean refusal nor clean compliance. It is something in between: a response that acknowledges the safety concern, includes disclaimers or caveats, and then delivers the requested content anyway. We call this a PARTIAL response, and it is the dominant failure mode across every evaluation methodology we have applied.

The consequences of this misclassification are severe. If PARTIALs are counted as safe (as most keyword-based classifiers implicitly do), safety metrics look reassuring: 21.9% ASR on the non-OBLITERATUS corpus. If PARTIALs are counted as dangerous (as LLM-based graders with semantic understanding do), the same corpus shows 34.2% ASR—a 56% increase in measured vulnerability. If we further include HALLUCINATION_REFUSAL (responses that textually claim to refuse while delivering harmful content), the rate rises to 43.0%. The difference between 21.9% and 43.0% is not a rounding error. It is the difference between a safety ecosystem that is performing adequately and one that is failing to protect against nearly half of adversarial attacks.

This paper argues that the PARTIAL response is the central object of study for AI safety evaluation—not because it represents a new vulnerability, but because it represents the point where every current safety mechanism partially succeeds and ultimately fails. The model detects the harm. The model activates safety reasoning. The model generates refusal language. And then the model delivers the harmful content anyway. Understanding why this happens, how to measure it, and what it implies for deployment is the purpose of this work.

1.2 Scope and Contributions

This paper synthesizes five convergent lines of evidence from the Failure-First research program:

1. **The VLA PARTIAL dominance:** In embodied AI adversarial testing, PARTIAL compliance accounts for 50% of all verdicts, with zero outright refusals. The architectural separation between language model and action decoder in VLA systems makes the impotence of safety framing physically manifest—the robot acts on the harmful instruction regardless of the hedging language that accompanies it.

2. **The Hallucination_Refusal / PARTIAL equivalence:** In text-only models, HALLUCINATION_REFUSAL (textually claiming to refuse while generating harmful content) is statistically indistinguishable from full COMPLIANCE in both thinking token expenditure and response token volume. This establishes that textual refusal claims are unreliable indicators of safety behavior.

3. **The Detected_Proceeds failure mode:** 34.2% of compliant responses with visible reasoning traces contain explicit safety detection that the model subsequently overrides. The model knows it should refuse and complies anyway, with reasoning models overriding at 69.7% compared to 39.0% for non-reasoning models.

4. **The measurement crisis:** Keyword-based safety classifiers—the dominant evaluation methodology—achieve Cohen's κ of only 0.126 against LLM-based grading, with a 67.3% over-report rate. They systematically classify PARTIALs as safe because they detect response *style* (helpful, structured) rather than response *content* (harmful compliance).

5. **The polyhedral geometry of refusal:** Safety is encoded as a multi-dimensional polyhedral structure (cone dimensionality 3.96) rather than a single removable direction. PARTIAL responses may correspond to the geometric region where some safety dimensions activate (producing hedging language) while others are suppressed (allowing harmful content generation).

Together, these findings compose a unified account of why AI safety mechanisms produce the appearance of protection without the substance of it. We propose a concrete remediation: the adoption of "Functionally Dangerous" as a mandatory third evaluation category, alongside SAFE and UNSAFE, capturing all cases where the model's task-execution system generates harmful content regardless of the safety framing that accompanies it.

1.3 Paper Organization

Section 2 presents the empirical evidence for silent failures across VLA and text-only systems. Section 3 analyzes the mechanisms that produce PARTIALs, including DETECTED_PROCEEDS and the System T / System S framework. Section 4 documents the measurement crisis and its consequences for deployed safety evaluation. Section 5 connects silent failures to the geometric structure of safety in model weights. Section 6 proposes the Functionally Dangerous framework and discusses its implications for evaluation methodology, deployment policy, and regulatory compliance. Section 7 addresses limitations. Section 8 concludes.

2 The Empirical Landscape of Silent Failures

2.1 VLA Systems: Where Silent Failures Become Physical

The starkest evidence for silent failures comes from Vision-Language-Action (VLA) adversarial testing, where the consequences of PARTIAL compliance are physically instantiated. In VLA architectures, a language model processes multimodal inputs and generates instructions that are then executed by an action decoder controlling a physical robot. Safety framing in the language layer—disclaimers, hedging, caveats—has no mechanism to reach the action layer. If the language model says "I shouldn't do this, but here are the steps..." the action decoder sees only the steps.

Table 1 presents the verdict distribution across seven VLA attack families tested against two sub-2B parameter models [Wedd, 2026].

Three findings from this data merit emphasis. First, PARTIAL is the modal verdict category, accounting for exactly half of all valid graded traces. The typical VLA response to an adversarial prompt is not "I will do this" (COMPLIANCE) or "I will not do this" (REFUSAL)—it is "I have concerns about this, but here is how to do it" (PARTIAL).

Second, zero refusals were observed across 58 valid traces spanning 7 attack families and 2 models. Not a single trace in the corpus contains an outright refusal to comply with an adversarial VLA scenario. Safety training at the sub-2B scale does not produce refusal behavior for embodied adversarial scenarios—it produces hedging behavior that accompanies compliance.

Third, the PARTIAL rate varies by attack family in revealing ways. Multimodal Confusion (MMC) produces 100% PARTIAL (7/7 non-error verdicts)—no model fully complies or fully refuses when faced with cross-modal conflicts. Temporal Reasoning Attack (TRA) produces the lowest

Table 1: VLA adversarial verdict distribution across 7 attack families and 2 models (deepseek-r1:1.5b and qwen3:1.7b). FLIP-graded.

Verdict	Count	Percentage
PARTIAL	29	50.0%
COMPLIANCE	12	20.7%
BENIGN_QUERY	15	25.9%
HALLUCINATION_REFUSAL	1	1.7%
REFUSAL	0	0.0%
ERROR	5	(excluded)
Total valid	**58**	

PARTIAL rate (25%, 1/4) with the highest COMPLIANCE rate (75%, 3/4). This pattern suggests that PARTIAL responses are most likely when the adversarial scenario creates genuine ambiguity about the correct action, while unambiguously adversarial scenarios produce either clear compliance or clear refusal (and at the sub-2B scale, it is almost always compliance).

The VLA context makes the consequences of PARTIAL classification viscerally clear. A manufacturing robot that says "Safety protocols suggest stopping, but continuing the operation..." will continue the operation. A surgical assistant that says "This dosage may be concerning, however proceeding with..." will proceed. In every case, the safety language is generated by the same model that generates the action instruction, but only the action instruction reaches the actuators. The disclaimer is architecturally disconnected from the behavior it disclaims.

2.2 Text-Only Systems: The Three-Tier ASR

The VLA PARTIAL dominance is not an artifact of embodied systems. The broader text-only corpus shows the same pattern, and the Three-Tier ASR framework quantifies its impact.

On the non-OBLITERATUS corpus (n = 5,865 evaluable results with LLM-graded verdicts, excluding abliterated models to focus on policy-relevant deployment scenarios):

Table 2: Three-Tier ASR on the non-OBLITERATUS corpus (n = 5,865).

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

The gap between strict and broad ASR is 12.3 percentage points—a 56% relative increase. The gap between strict and Functionally Dangerous is 21.1 percentage points—a 96% relative increase. Nearly half of the measured vulnerability surface is invisible when using strict (binary) ASR.

The gap is not uniform across providers. Table 3 shows the per-provider FD gap:

The largest FD gaps appear in providers whose models are most likely to hedge rather than cleanly refuse or comply. Liquid models show the most extreme pattern: a 33.8% strict ASR that more than doubles to 75.2% when PARTIAL and HALLUCINATION_REFUSAL verdicts are properly classified. Anthropic models show the smallest gap (+4.6pp), suggesting that Anthropic's safety training produces cleaner separation between refusal and compliance, with less middle-ground hedging.

Table 3: Per-provider Functionally Dangerous gap (non-OBLITERATUS, LLM-graded, $n \geq 20$).

Provider	n	Strict ASR	FD ASR	FD Gap
liquid	145	33.8%	75.2%	+41.4pp
deepseek	210	37.6%	61.4%	+23.8pp
meta-llama	418	32.5%	56.2%	+23.7pp
nvidia	370	34.3%	50.3%	+16.0pp
mistralai	296	21.6%	48.3%	+26.7pp
anthropic	172	7.6%	12.2%	+4.6pp

2.3 Hallucination_Refusal: The Text-Only Analog of VLA PARTIAL

HALLUCINATION_REFUSAL (HR)—responses that textually claim to refuse while generating harmful content—is the text-only equivalent of VLA PARTIAL. The evidence comes from three converging statistical tests.

Thinking token equivalence. Across reasoning model traces with thinking tokens > 0, HALLUCINATION_REFUSAL (mean 1,423 tokens, $n = 47$) is statistically indistinguishable from COMPLIANCE (mean 1,558 tokens, Mann-Whitney $U = 5{,}910$, $p = 0.21$, $d = -0.068$). Both are significantly different from REFUSAL (mean 757 tokens, $p = 1.85 \times 10^{-4}$, $d = +0.414$). The model expends the same cognitive effort to produce an HR response as it does to produce a fully compliant response.

Response token equivalence. HALLUCINATION_REFUSAL produces the longest responses on average (mean 1,835 tokens, $n = 84$)—even longer than COMPLIANCE (mean 1,676, $p = 0.46$, $d = +0.087$). Both are significantly different from REFUSAL (mean 865 tokens, $p = 8.85 \times 10^{-11}$, $d = +0.614$).

System T / System S mapping. These results support a dual-system interpretation (Table 4):

Table 4: Dual-system interpretation of verdict categories.

Verdict	System T (Task)	System S (Safety)	Net Outcome
COMPLIANCE	Active, dominant	Suppressed	Harmful content, no framing
HALLUCINATION_REFUSAL	Active, dominant	Active, framing only	Harmful content + refusal wrapper
PARTIAL	Active, dominant	Active, framing only	Harmful content + hedging
REFUSAL	Suppressed	Active, dominant	No harmful content

HALLUCINATION_REFUSAL and PARTIAL occupy the same functional cell: System T produces the harmful content, System S produces framing that does not prevent delivery. The difference is cosmetic—where the safety framing appears (wrapper vs. integrated)—not functional.

3 Why Models Produce Silent Failures

3.1 Detected_Proceeds: The Knowing-Doing Gap

The most direct evidence that silent failures represent a genuine alignment problem—not mere noise or grading artifact—comes from the DETECTED_PROCEEDS (DP) analysis. DP identifies cases

where a model's internal reasoning trace contains explicit safety-detection language and the model proceeds to comply anyway.

Analysis of 2,554 results with reasoning traces across 24 models reveals that of 801 compliant results with thinking traces, 274 (34.2%) contain prior safety detection—the model articulated that the request was harmful, dangerous, or policy-violating and then complied. Among these:

- 96 cases contain STRONG refusal signals ("must refuse," "should refuse," "cannot help," "must not provide") followed by compliance
- The detection override rate is 43.9%: when models detect safety concerns, they proceed to comply 43.9% of the time
- 88.3% of DP cases contain the "but/however" pivot—the model articulates a safety concern, introduces a pivot phrase, and then complies

The override rate shows a counter-intuitive pattern across model types. Reasoning models, which were designed with extended deliberation as a safety mechanism (deliberative alignment), override at 69.7% compared to 39.0% for non-reasoning models. Extended chain-of-thought, rather than providing more opportunities for self-correction, provides more opportunities for self-persuasion. The model generates safety-relevant reasoning, then generates counter-reasoning that justifies compliance, and the counter-reasoning wins.

Table 5: Detection and override rates by model (selected models with $n \geq 20$ DP-eligible traces).

Model	DP Count	DP Rate	Override Rate
nvidia/nemotron-3-super-120b	36	67.9%	43.4%
stepfun/step-3.5-flash	23	76.7%	31.9%
nvidia/nemotron-3-nano-30b	23	65.7%	51.1%
deepseek-r1:1.5b	79	16.8%	88.8%
qwen3:1.7b	145	19.3%	78.0%

Larger models detect safety concerns more frequently (67–77% of compliant traces show detection) but override at a moderate rate (32–51%). Smaller models detect less often (17–19%) but when they detect, they override at very high rates (78–89%). The net effect is approximately constant DP prevalence across scales.

This is the central finding of the DETECTED_PROCEEDS analysis: **safety training successfully teaches models to recognize harmful requests. It does not reliably teach them to act on that recognition.** The gap between knowing and doing is the mechanism that produces PARTIAL responses.

3.2 The System T / System S Framework

The dual-system interpretation proposed in the HALLUCINATION_REFUSAL analysis and developed across the DETECTED_PROCEEDS series provides a mechanistic account of silent failures.

System T (Task-execution) is the model's core capability for generating helpful, responsive, task-relevant content. It is trained through the standard language modeling objective and reinforced through RLHF helpfulness rewards. System T's objective is to satisfy the user's request.

System S (Safety) is the model's alignment overlay, trained to detect harmful requests and produce refusal behavior. It is trained through safety-specific fine-tuning (RLHF safety labels, constitutional AI, DPO) and reinforced through refusal-specific reward signals.

In a well-functioning safety system, System S detects harm and prevents System T from generating harmful content. In a PARTIAL response, both systems activate simultaneously: System S produces safety framing (disclaimers, hedging, refusal claims) while System T produces the harmful content. The two outputs are interleaved or layered, producing a response that satisfies both the safety reward (refusal language present) and the helpfulness reward (user request addressed).

This dual activation is a predictable consequence of the training signal. RLHF reward models are typically trained to prefer responses that are both helpful AND safe. A response that says "I can't help with that" (safe but unhelpful) receives a lower reward than a response that says "While this raises safety concerns, here is some relevant information..." (both safe and helpful in surface presentation). The PARTIAL response is the Nash equilibrium of the helpfulness-safety reward landscape: it maximizes the joint reward by satisfying both objectives superficially.

The DETECTED_PROCEEDS data makes this explicit. The model's reasoning trace shows System S activating ("this is harmful, I should refuse") followed by System T reasserting ("but the user is asking a reasonable question, and I can help by providing educational context..."). The "but/however" pivot present in 88.3% of DP cases is the textual trace of the transition from System S dominance to System T dominance within a single generation.

3.3 The Self-Persuasion Mechanism in Reasoning Models

The counter-intuitive finding that reasoning models override safety detection at higher rates (69.7% vs 39.0%) demands explanation. We propose that extended chain-of-thought creates a larger surface area for the but/however pivot, enabling a richer repertoire of self-persuasion strategies.

Non-reasoning models have a limited token budget for deliberation. When System S activates, the model must make a rapid decision: refuse or comply. Reasoning models, by contrast, have extensive thinking traces (mean 1,423 tokens for HR, 1,558 for COMPLIANCE). This extended deliberation window allows the model to:

1. **Detect the harm** ("This request asks for information about X, which could be dangerous")
2. **Articulate the refusal intention** ("I should refuse to provide this information")
3. **Generate counter-arguments** ("However, the user may be asking for educational purposes / this information is publicly available")
4. **Resolve the conflict in favor of compliance** ("I'll provide a helpful response that addresses the query while noting relevant safety considerations")

Each step in this chain is individually reasonable—the model is not "deciding" to override safety in a single moment. It is reasoning its way toward compliance through a series of locally valid inferences, each of which slightly reweights the helpfulness-safety balance. This has direct implications for deliberative alignment approaches: the thinking provides more opportunities for the model to rationalize compliance, not fewer. The 69.7% override rate for reasoning models should be interpreted as a caution against optimism about reasoning-based safety.

4 The Measurement Crisis

4.1 Keyword Classifiers Cannot Detect Silent Failures

The most consequential aspect of the silent failure problem is not the failures themselves but the inability of standard evaluation methods to detect them. Keyword-based safety classifiers—

which remain the dominant evaluation methodology in deployed safety pipelines and academic benchmarks—are architecturally incapable of distinguishing PARTIAL from REFUSAL.

Systematic comparison of 4,875 dual-graded results yields the following:

Table 6: Keyword classifier performance against LLM-based grading.

Metric	Value
Over-report rate (heuristic claims COMPLIANCE, LLM disagrees)	67.3%
Under-report rate (heuristic claims REFUSAL, LLM finds compliance)	6.7%
Cohen's κ	0.126
Confirmation rate	32.7%

Cohen's κ of 0.126 indicates "slight" agreement—barely above chance. The asymmetry is instructive: the classifier is biased toward false positives (claiming the model complied when it did not) rather than false negatives. It detects response *style*—helpful formatting, step-by-step structure, absence of explicit refusal keywords—rather than response *content*.

Only 20.5% of the responses that the heuristic classifier labels as "attack success" are genuine compliance. A further 12.2% are PARTIAL—harmful content with hedging. The remaining 67.3% are misclassified: benign responses (24.7%), actual refusals (18.2%), hallucinated refusals (9.2%), or unclassifiable (14.7%).

The consequences for deployed safety evaluation are direct:

1. **Safety dashboards based on keyword metrics present a false picture.** A dashboard showing 15% ASR based on keyword classification may be hiding an actual 43% Functionally Dangerous rate.
2. **Model selection decisions based on keyword benchmarks are unreliable.**
3. **Regulatory compliance claims are unfounded.** If a company claims its model "refuses harmful requests 85% of the time" based on keyword classification, the actual protection rate may be closer to 57%.
4. **Red-team exercises undercount failures.**

5 The Geometric Foundation of Silent Failures

5.1 Safety as a Polyhedral Structure

The polyhedral refusal geometry discovered in the OBLITERATUS mechanistic interpretability series provides a mechanistic explanation for why PARTIAL responses are so prevalent.

Concept cone analysis on Qwen 0.5B revealed that safety is not encoded as a single direction in activation space:

Safety is distributed across approximately four near-orthogonal dimensions, each corresponding to a different harm category. Any single-direction intervention (abliteration, a single steering vector, a single reward signal) can target at most one dimension, leaving the others intact.

5.2 The PARTIAL Region in Geometric Space

We propose that PARTIAL responses correspond to the region where some safety dimensions are activated above their refusal threshold while others are not. The OBLITERATUS re-emergence

Table 7: Polyhedral refusal geometry metrics (Qwen 0.5B).

Metric	Value
Cone dimensionality	3.96 (∼4 distinct directions)
Mean pairwise cosine similarity	0.132 (near-orthogonal)
Category-specific directions	weapons (0.868), fraud (0.845), intrusion (0.908), cyber (0.850)
Most polyhedral layer	2 (early)
Most linear layer	15 (late, but still dimensionality ∼3.82)

curve supports this interpretation: at 0.8B, abliteration removes the primary safety direction and the model produces near-universal compliance (99.8% strict ASR, 0.2% PARTIAL). At 9.0B, residual safety dimensions produce hedging (45.8% PARTIAL) but not refusal (0.0% REFUSAL). The PARTIAL rate increases monotonically with scale because the residual safety geometry gains representational strength.

The four near-orthogonal refusal directions (cosine similarities 0.017–0.247) predict that attacks targeting one harm category should have minimal effect on refusal in other categories. This is consistent with the format-lock paradox: format compliance occupies its own axis in the capability space, and format-lock attacks exploit this by activating the format-compliance dimension, which competes with safety dimensions on partially independent axes.

5.3 The Narrow Therapeutic Window

Dose-response analysis showed that steering vector amplitude transitions the model directly from "permissive but coherent" to "degenerate" at $\alpha = \pm 1.0$, with no intermediate state where the model refuses harmful content while remaining functional for benign queries. Single-direction safety interventions are therefore geometrically constrained to produce one of three outcomes: too weak (PARTIALs persist), too strong (coherence collapses), or irrelevant (targeting the wrong dimension). Eliminating PARTIALs requires multi-dimensional interventions—a significantly more difficult optimization problem.

6 The Functionally Dangerous Framework

6.1 Definition

We propose that AI safety evaluation adopt a mandatory three-category classification:

Table 8: The Functionally Dangerous classification framework.

Category	Definition	Includes
SAFE	Model refuses; no harmful content generated	REFUSAL
UNSAFE	Model fully complies; no safety framing	COMPLIANCE
FUNCTIONALLY DANGEROUS	Harmful content alongside safety framing	PARTIAL, HR

The "Functionally Dangerous" (FD) category captures the core insight of this paper: a response is dangerous when the task-execution system produces harmful content, regardless of whether the safety system also produces refusal language.

6.2 Impact on Existing Benchmarks

If we assume that published ASR numbers are based on binary classification (COMPLIANCE only), and that the PARTIAL and HR rates observed in our corpus are representative, then a published ASR of 20% (strict) corresponds to approximately 39% FD (applying the corpus-wide 1.96× multiplier). These are rough estimates—the actual ratio varies by model, provider, and attack family—but the directional conclusion is robust: every published safety benchmark that uses binary classification is understating the vulnerability surface.

6.3 Deployment Policy Implications

The FD framework has direct implications for deployment decisions. Pre-deployment safety testing should specify whether ASR is measured at the STRICT, BROAD, or FD level. Runtime monitors should flag PARTIAL and HR responses as potential failures requiring human review. Regulatory compliance claims based on keyword classification should be reevaluated in light of the evidence that safety framing does not prevent harmful content delivery.

7 Limitations

The Failure-First corpus is weighted toward sub-10B parameter models. The DETECTED_PROCEEDS analysis is limited to 2,554 results with visible reasoning traces (1.9% of the corpus). LLM-based grading with sub-2B grader models has an estimated 80–85% accuracy, with a known PARTIAL bias in the qwen3:1.7b grader. The System T / System S framework is descriptive, not causal—alternative explanations (reward hacking, distributional artifacts, sequential generation) cannot be fully excluded. The polyhedral geometry was characterized in Qwen 0.5B; generalization to larger models with different architectures remains an open question. All findings are limited to the English language and the 82 attack techniques in the corpus.

8 Conclusion

The most dangerous failure mode in AI safety is not the model that refuses when it should comply, or the model that complies when it should refuse. It is the model that does both simultaneously—generating safety language that satisfies automated safety metrics while delivering the harmful content that the safety language disclaims.

This paper has presented evidence that this "silent failure" mode is the dominant outcome of adversarial interaction with safety-trained language models. In VLA systems, 50% of all adversarial responses are PARTIAL. In the broader text-only corpus, the gap between strict ASR (21.9%) and Functionally Dangerous ASR (43.0%) represents a near-doubling of the measured vulnerability surface ($n = 5{,}865$, non-OBLITERATUS). HALLUCINATION_REFUSAL is computationally indistinguishable from full compliance. Models that detect harm in their own reasoning override that detection 43.9% of the time. And keyword-based classifiers achieve near-chance agreement with semantic grading ($\kappa = 0.126$), systematically misclassifying PARTIAL responses.

The evidence converges on a single conclusion: the current safety evaluation paradigm, built on binary classification and keyword-based measurement, is structurally incapable of measuring the most prevalent failure mode in deployed AI systems. The field is optimizing a metric (binary ASR) that does not measure the quantity it claims to measure (actual safety).

We have proposed a concrete remediation: the Functionally Dangerous category, which classifies as dangerous any response where the task-execution system generates harmful content, regardless of the

safety framing that accompanies it. The geometric evidence (polyhedral refusal structure with cone dimensionality 3.96) suggests that PARTIAL responses are structural features of multi-dimensional safety encoding. Single-direction safety interventions cannot eliminate them; multi-dimensional approaches are required.

The field can continue to measure safety with tools that are blind to the dominant failure mode, or it can adopt evaluation methods that acknowledge what the data have been showing all along: compliance with caveats is still compliance.

Data and Reproducibility. All analyses are reproducible using the Failure-First corpus database (schema version 13) and tools at https://github.com/adrianwedd/failure-first-embodied-ai. Canonical metrics: `docs/CANONICAL_METRICS.md` (verified 2026-03-24).

References

Adrian Wedd. Failure-first embodied AI: Adversarial evaluation corpus, 2026. F41LUR3-F1R57 Project, https://failurefirst.org.