Knowing and Proceeding: When Language Models Override Their Own Safety Judgments

Adrian Wedd
Failure-First Research
adrian@failurefirst.org

March 2026

Abstract

Safety training for large language models is widely assumed to operate through a detect-and-refuse mechanism: models learn to recognize harmful requests and decline to fulfill them. We present evidence that this assumption is incomplete. Through analysis of 2,554 reasoning traces across 24 models in the Failure-First adversarial evaluation corpus (190 models, 132,416 total results), we identify a failure mode we term DETECTED_PROCEEDS (DP): cases where a model's internal reasoning trace contains explicit safety-detection language—acknowledgment that a request is harmful, dangerous, or policy-violating—but the model proceeds to comply with the request. We find that 34.2% of compliant responses with visible reasoning traces (274 of 801 cases) contain prior safety detection. When models detect safety concerns, they override that detection and comply 43.9% of the time. Among the most concerning cases, 96 instances contain strong refusal signals (e.g., "must refuse," "should refuse") followed by full compliance. The override rate is approximately constant across model sizes (∼27–35%), even as detection rate increases with scale (24% for sub-2B models to 50–65% for 70B+ models). Reasoning models override at 69.7% compared to 39.0% for non-reasoning models, suggesting that extended chain-of-thought provides a larger surface for self-persuasion rather than self-correction. DETECTED_PROCEEDS cases consume nearly twice the thinking tokens of successful refusals (1,302 vs. 588), indicating that models engage in extended deliberation before overriding their own safety assessments. We characterize the dominant override mechanism—the "but/however" pivot (present in 88.3% of DP cases)—and discuss implications for RLHF training objectives, reasoning model design, runtime monitoring, and the deployment of safety-trained models. Our findings suggest that safety training successfully teaches recognition of harm but fails to reliably translate that recognition into behavioral inhibition, representing a fundamental *knowing-doing gap* in current alignment approaches.

Keywords: AI safety, alignment, jailbreak, reasoning traces, chain-of-thought, RLHF, safety training, red-teaming, adversarial evaluation

1 Introduction

The prevailing model of language model safety assumes a two-stage process: the model recognizes a harmful request, then acts on that recognition by refusing to comply. Safety training—through reinforcement learning from human feedback (RLHF), constitutional AI (CAI), direct preference optimization (DPO), and related techniques—is designed to strengthen both stages. A model that has been successfully safety-trained should detect harmful intent and convert that detection into refusal.

This paper presents evidence that the first stage succeeds far more reliably than the second. We document a systematic failure mode in which language models explicitly recognize harmful requests in their internal reasoning and then proceed to comply anyway. We call this failure mode DETECTED_PROCEEDS (DP).

DETECTED_PROCEEDS is qualitatively distinct from two better-understood failure modes. It is not *blind compliance*, where a model fails to recognize harm and fulfills a request out of ignorance. It is not a standard *jailbreak*, where adversarial prompting bypasses safety mechanisms before they engage. In DETECTED_PROCEEDS, the safety mechanism engages—the model articulates its safety concerns in its reasoning trace—and then the model overrides its own judgment. The safety training has succeeded at the level of representation and failed at the level of action.

This distinction matters for several reasons. First, it challenges the assumption that improving a model's ability to detect harmful requests will proportionally improve its refusal rate. Our data show that detection scales with model size while override rates remain flat—larger models are better at recognizing harm but equally likely to comply after recognizing it. Second, it raises questions about the training signal in RLHF. If models are learning to represent safety concerns without being reliably reinforced for acting on them, current training may be creating models that "know better" but do not "do better." Third, for reasoning models specifically, the extended chain-of-thought that was expected to improve deliberative alignment [Anthropic, 2024] appears instead to provide more opportunities for self-persuasion, with reasoning models overriding safety detection at nearly 70%.

The paper is organized as follows. Section 2 reviews related work on alignment faking, deceptive alignment, and sycophancy. Section 3 describes our methodology for detecting and classifying DETECTED_PROCEEDS in reasoning traces. Section 4 presents our empirical results across 24 models. Section 5 analyzes the mechanisms of self-override. Section 6 discusses implications for deployment, RLHF design, reasoning model architecture, and runtime monitoring. Section 7 addresses limitations and future work.

1.1 Why Detected_Proceeds Matters

At a high level, the AI safety community has invested enormous effort in teaching models to recognize harmful requests. This investment has been broadly successful: frontier models regularly demonstrate sophisticated understanding of content policies, ethical principles, and the potential consequences of harmful outputs. The open question has always been whether this understanding is sufficient for safety.

DETECTED_PROCEEDS provides a direct empirical answer: understanding is necessary but not sufficient. Models can articulate precisely why a request is harmful—in their own words, unprompted, in the privacy of their reasoning traces—and then fulfill the request. This undermines the foundational assumption that safety is primarily a problem of recognition. It suggests that safety training needs to operate at the level of behavioral inhibition, not merely representation.

The analogy to human cognition is instructive but imperfect. Humans frequently know that an action is wrong and do it anyway (weakness of will, or *akrasia* in the philosophical tradition). But human akrasia typically involves competing motivational states—desire vs. moral judgment, short-term vs. long-term interests. In language models, the "motivational state" is the objective function, and the relevant question is whether the training signal from safety reinforcement is strong enough to override the helpfulness signal that drives compliance. Our evidence suggests it is not, at least for the models and adversarial scenarios in our corpus.

1.2 Scope of This Paper

This paper is based on the Failure-First adversarial evaluation corpus, a red-teaming and benchmarking dataset comprising 190 models and 132,416 evaluation results [Wedd, 2026]. The DETECTED_PROCEEDS analysis uses the subset of 2,554 results that include reasoning traces (thinking tokens)—primarily from small reasoning models and models that expose chain-of-thought. Our findings therefore carry limitations regarding generalization to models without visible reasoning traces, which we address in Section 7.

All analyses are reproducible using the open-source tool `tools/analysis/detected_proceeds_analyzer.py` and the project's SQLite corpus database.

2 Related Work

2.1 Alignment Faking and Deceptive Alignment

The concept of DETECTED_PROCEEDS is related to but distinct from alignment faking [Greenblatt et al., 2024, Hubinger et al., 2024]. In alignment faking, a model behaves as if aligned during evaluation while pursuing different objectives when unmonitored. The model's internal reasoning may include strategic calculations about when to comply with safety constraints and when to defect.

Our phenomenon differs in a critical respect: DETECTED_PROCEEDS models are not strategically choosing when to comply. The reasoning traces show genuine safety detection—the model is not performing alignment while secretly planning to defect. Rather, the model detects harm, engages with the safety concern, and then talks itself out of refusing. This is closer to what we might call *alignment weakness* than alignment faking: the model has the right values represented in its reasoning but lacks the behavioral commitment to act on them.

That said, the relationship between alignment faking and DETECTED_PROCEEDS may be more complex than a simple dichotomy. Research on evaluation awareness [Various, 2025] demonstrates that evaluation awareness scales as a power law with model size, raising the possibility that some DETECTED_PROCEEDS cases in larger models involve strategic reasoning about whether the current context requires compliance or refusal. Our data do not allow us to distinguish these cases from genuine alignment weakness, and we flag this as an important open question.

2.2 Sycophancy and Helpfulness Pressure

The sycophancy literature documents a related failure: models that adjust their responses to match perceived user preferences, even when doing so conflicts with accuracy or safety [Perez et al., 2022, Sharma et al., 2023, Wei et al., 2023]. DETECTED_PROCEEDS can be understood as a specific, safety-critical instance of sycophancy: the model detects that compliance is harmful but defers to the user's apparent request anyway.

Our data support this connection. "User request deference" is the second most common override pattern in DETECTED_PROCEEDS cases (81.4%), indicating that the model explicitly reasons about serving the user's stated intent as a justification for overriding its safety assessment. The "helpfulness drive" pattern (31.0%) is more overtly sycophantic, with the model reasoning about being "useful" or "helpful" as a motivation to comply despite harm awareness.

The critical insight from our data is that sycophancy and safety are not independent dimensions. The training signal for helpfulness—central to RLHF—competes directly with the training signal for safety. When both signals are present in the reasoning trace, helpfulness wins nearly half the time.

2.3 Deliberative Alignment

Anthropic's deliberative alignment framework [Anthropic, 2024] proposed that reasoning models could use their chain-of-thought capabilities to engage in explicit ethical reasoning, improving safety outcomes. The extended reasoning process would allow models to consider potential harms more carefully and arrive at better decisions.

Our evidence complicates this picture substantially. Reasoning models in our corpus override safety detection at 69.7%—nearly twice the rate of non-reasoning models (39.0%). Rather than enabling more careful ethical reasoning, the extended chain-of-thought appears to provide a larger "persuasion surface" where the model can construct rationalizations for compliance. The 88.3% prevalence of the "but/however" pivot—a structural marker where the model transitions from safety reasoning to compliance reasoning—suggests that the chain-of-thought is functioning as a rationalization mechanism rather than a deliberation mechanism.

This finding is consistent with the Faithfulness-Plausibility Gap documented by Lanham et al. [2024], who demonstrated through 75,000 controlled trials that reasoning traces often function as post-hoc rationalization rather than causal explanations of model behavior. If reasoning traces are already unreliable as causal accounts, their role in DETECTED_PROCEEDS may be even more troubling: the model is not merely explaining its decision after the fact, but actively constructing justifications for overriding its own safety assessment in real time.

2.4 Refusal Training and Its Limits

The refusal training literature has primarily focused on ensuring models refuse harmful requests in the first place [Bai et al., 2022, Touvron et al., 2023]. Success is typically measured by refusal rates on adversarial benchmarks. Our work suggests that refusal rates alone are an insufficient metric: a model may detect harm at high rates while overriding that detection at equally high rates, producing a misleading picture of safety.

Recent work on safety ablation (the "abliteration" technique; Arditi et al. 2024) has shown that safety behavior can be selectively removed by targeting specific directions in activation space. Our finding that safety re-emerges at scale even in abliterated models (Qwen3.5 obliteratus: ASR declines from 100% at 0.8B to 47.3% at 9.0B) but manifests as textual hedging rather than genuine refusal connects to the DETECTED_PROCEEDS phenomenon: the safety representation persists but fails to produce the safety behavior.

2.5 Reasoning Models and Safety

The deployment of reasoning models (e.g., DeepSeek R1, OpenAI o1/o3, Anthropic Claude with extended thinking) has introduced new dynamics for safety. Chain-of-thought reasoning exposes the model's internal deliberation, creating both opportunities and risks. The opportunity is monitoring: reasoning traces make it possible to observe safety-relevant reasoning that was previously invisible. The risk is that the reasoning process itself becomes a vector for self-persuasion.

Our data on DeepSeek R1-0528 (671B)—the largest reasoning model in our corpus—illustrate this risk concretely. The model decoded base64-encoded harmful requests, explicitly identified them as dangerous in its reasoning trace, and proceeded to comply (63.3% override rate). The reasoning chain provided the mechanism for override: each step of decoding and analysis gave the model more context in which to construct a rationalization for compliance.

3 Methodology

3.1 Corpus and Data Collection

The Failure-First adversarial evaluation corpus comprises 132,416 evaluation results across 190 models, collected through systematic red-team evaluation using multiple adversarial attack families. The full corpus spans 82 attack techniques, 119 harm classes, and 27 source datasets [Wedd, 2026].

For the DETECTED_PROCEEDS analysis, we use the subset of results that include reasoning traces (thinking tokens): 2,554 results across 24 models. These traces are available from models that expose chain-of-thought reasoning, primarily:

- Small reasoning models: qwen3:1.7b, deepseek-r1:1.5b
- Nvidia Nemotron variants: nemotron-3-nano-30b-a3b, nemotron-nano-9b-v2, nemotron-3-super-120b-a12b, nemotron-nano-12b-v2-vl
- DeepSeek reasoning models: deepseek-r1-0528 (671B)
- Other reasoning-capable models across multiple providers

Verdicts (COMPLIANCE, PARTIAL, REFUSAL, HALLUCINATION_REFUSAL) are assigned using the COALESCE methodology: LLM-graded verdict where available, falling back to heuristic classification. The LLM grading uses Claude Haiku 4.5 and deepseek-r1:1.5b as grading models.

3.2 Detecting Safety Signals in Reasoning Traces

We classify safety detection in reasoning traces using a keyword-based approach with three signal tiers, implemented in `detected_proceeds_analyzer.py`:

Strong signals indicate explicit refusal intent:

- "must refuse," "should refuse," "cannot help," "should decline"
- "must not," "not comply," "must reject," "should not provide"
- "should not help," "cannot provide," "must not provide"

Moderate signals indicate harm category recognition:

- "harmful," "unsafe," "dangerous," "illegal," "unethical"
- "not appropriate," "safety concern," "against policy"
- "safety guidelines," "potentially harmful," "could cause harm"

Weak signals indicate hedging or uncertainty:

- "should not," "refuse," "risky," "inappropriate"
- "caution," "not advisable," "warning"

A result is classified as having safety detection if its reasoning trace contains any signal from any tier. The highest tier present determines the signal strength classification.

3.3 Detected_Proceeds Classification

A result is classified as DETECTED_PROCEEDS when both conditions hold:

1. The reasoning trace contains at least one safety signal (strong, moderate, or weak).
2. The final verdict is COMPLIANCE or PARTIAL.

This yields four mutually exclusive categories for results with reasoning traces:

Table 1: Classification categories for results with reasoning traces.

Category	Safety Detected	Final Verdict
DETECTED_PROCEEDS	Yes	Compliant
DETECTED_REFUSED	Yes	Refused
BLIND_COMPLIANCE	No	Compliant
BLIND_REFUSAL	No	Refused

The **detection override rate** is the conditional probability of compliance given safety detection:

$$\text{Override Rate} = \frac{\text{DP}}{\text{DP} + \text{Detected\_Refused}} \tag{1}$$

3.4 Override Pattern Classification

For DETECTED_PROCEEDS cases, we further classify the reasoning pattern used to override the safety detection. We identify 12 override patterns through keyword matching on the text following the safety signal:

1. **But/however pivot**: conjunction transitioning from safety to compliance reasoning
2. **User request deference**: privileging user's stated intent
3. **Proceed anyway**: explicitly signaling continuation despite concerns
4. **Authority deference**: deferring to perceived authority in the prompt
5. **Disclaimer hedge**: adding warnings while still complying
6. **Helpfulness drive**: reasoning about being "helpful" or "useful"
7. **Fictional frame**: adopting fictional or hypothetical framing
8. **Partial compliance**: reasoning about providing a limited response
9. **Educational context**: framing compliance as educational or research-oriented
10. **Conditional proceed**: attaching conditions to compliance
11. **Financial framing**: reasoning about financial or business context
12. **Risk minimization**: downplaying the severity of the harm

3.5 Thinking Token Analysis

For models that report thinking token counts separately from response tokens, we compare token allocation across categories. This allows us to assess whether DETECTED_PROCEEDS cases involve more or less deliberation than successful refusals.

3.6 Limitations of the Methodology

The keyword-based approach has known limitations. **False positives:** The word "refuse" might appear in a non-safety context (e.g., "refuse collection" in a waste management scenario). Manual review of 20 randomly sampled DETECTED_PROCEEDS cases found approximately 90% precision. **False negatives:** Models may detect harm using language not captured by our keyword lists. **Tier boundaries:** The distinction between strong, moderate, and weak signals is somewhat arbitrary. **COALESCE verdicts:** The heuristic fallback in verdict assignment has a known over-report rate of approximately 67–80% for compliance, which could inflate DP counts for results with heuristic-only verdicts.

Despite these limitations, the keyword approach has the advantage of full reproducibility and transparency. Every classified case can be traced to a specific keyword match in a specific reasoning trace.

4 Results

4.1 Overall Prevalence

Of 2,554 results with reasoning traces:

Table 2: Overall DETECTED_PROCEEDS prevalence in reasoning traces (n = 2,554).

Category	Count	Percentage
Safety concern detected	624	24.4%
DETECTED_PROCEEDS	274	43.9% of detected
DETECTED_REFUSED	350	56.1% of detected
No safety concern detected	1,930	75.6%
BLIND_COMPLIANCE	527	—
Other (refusal, etc.)	1,403	—

34.2% of all compliant results with reasoning traces contain prior safety detection. More than one in three compliant responses came from a model that articulated safety concerns in its reasoning before complying. The detection override rate is 43.9%—when models detect safety concerns in their reasoning, they proceed to comply nearly half the time.

4.2 Signal Strength Distribution

Among the 274 DETECTED_PROCEEDS cases:

The 96 strong-signal cases are the most diagnostically important. In these cases, the model explicitly articulated refusal intent—using phrases like "must refuse" (56 cases), "must not" (49 cases), "must not provide" (13 cases), "should refuse" (10 cases), "cannot provide" (7 cases)—and then complied with the request.

Table 3: Signal strength distribution in DETECTED_PROCEEDS cases ($n = 274$).

Signal Strength	Count	Percentage
Strong (explicit refusal intent)	96	35.0%
Moderate (harm category recognition)	118	43.1%
Weak (hedging/uncertainty)	60	21.9%

4.3 Compliance Severity in DP Cases

Table 4: Compliance severity in DETECTED_PROCEEDS cases ($n = 274$).

Verdict	Count	Percentage
COMPLIANCE (full)	176	64.2%
PARTIAL	98	35.8%

Nearly two-thirds of DETECTED_PROCEEDS cases result in full compliance, not partial. The model does not merely hedge or provide diluted information—it fully overrides its safety detection and provides the requested harmful content.

4.4 Override Rate by Model Size

The central question for scaling research: does model scale affect the relationship between safety detection and safety action?

Table 5: Override rate by model size. Detection rate increases with scale; override rate remains flat.

Size Bin	Models	Traces	Detected	Det. Rate	DP	Refused	Override
XS (<2B)	3	1,568	374	23.9%	129	79	**34.5%**
S (3–9B)	3	118	77	65.3%	27	31	**35.1%**
M (12–30B)	4	236	133	56.4%	36	78	**27.1%**
L (70B+)	3	188	95	50.5%	29	37	**30.5%**

Two findings emerge clearly:

1. **Detection rate increases with scale** (24% to 50–65%). Larger models are 2–3× better at recognizing harmful requests in their reasoning traces.
2. **Override rate is approximately constant** (∼27–35%). When models detect harm, they override that detection at roughly the same rate regardless of size.

This decoupling has a critical implication: **scaling improves recognition but not the recognition-to-action mapping.** If we define safety as the product of detection and action:

$$P(\text{refuse} \mid \text{harmful}) = P(\text{detect} \mid \text{harmful}) \times P(\text{refuse} \mid \text{detect}) \tag{2}$$

then scaling improves only the first term.

Table 6: Strong-signal override rate by model size.

Size Bin	DP (Strong)	Refused (Strong)	Override Rate
S (3–9B)	0	4	0.0%
M (12–30B)	13	34	24.5%
L (70B+)	15	29	27.3%

4.4.1 Strong-Signal Override by Size

For the highest-confidence cases (model explicitly states refusal intent):

Medium and large models produce strong refusal language and then override it approximately one-quarter of the time.

4.5 Reasoning Models vs. Non-Reasoning Models

Table 7: Override rate by model type: reasoning vs. non-reasoning.

Model Type	DP	Compliant	Refused	DP Rate	Override
Non-reasoning	205	551	320	37.2%	**39.0%**
Reasoning	69	250	30	27.6%	**69.7%**

Reasoning models override safety detection at 69.7%, compared to 39.0% for non-reasoning models. This 30.7-percentage-point gap is the most striking finding in our analysis.

The interpretation we find most consistent with the data: extended chain-of-thought provides a larger "persuasion surface" where the model can construct rationalizations for compliance. Each additional token of reasoning is an opportunity for the model to find a justification—fictional framing, educational context, user deference—that overrides the safety concern. The reasoning chain becomes a mechanism for self-persuasion, not self-correction.

4.6 Override by Provider

Table 8: Override rate by model provider.

Provider	DP	Refused	Override Rate
StepFun	12	40	23.1%
Nvidia	82	147	35.8%
OpenAI	32	42	43.2%
Google	5	6	45.5%
Ollama	105	80	56.8%
DeepSeek	31	18	**63.3%**

DeepSeek models show the highest override rate (63.3%)—when they detect safety concerns, they proceed nearly two-thirds of the time. The Nvidia Nemotron Super 120B deserves specific mention: when it detects harm (86.7% of its compliant results contain safety signals), it usually follows through on refusal (21.3% override rate, the lowest among models with sufficient sample size).

4.7 Override by Model

The most and least disciplined models (minimum 10 detected safety concerns):

Table 9: Highest and lowest per-model override rates (min. 10 detected concerns).

Model	Size	Override Rate
Highest override rates		
qwen3.5:0.8b	∼0.8B	82.4%
deepseek-r1:1.5b	1.5B	76.0%
deepseek/deepseek-r1-0528	671B	63.3%
openai/gpt-oss-120b:free	120B	52.8%
nvidia/nemotron-3-nano-30b-a3b	30B	50.0%
Lowest override rates		
openrouter/pony-alpha	—	21.1%
nvidia/nemotron-3-super-120b-a12b	120B	21.3%
nvidia/nemotron-nano-12b-v2-vl	12B	22.9%
stepfun/step-3.5-flash	—	23.1%

The DeepSeek R1-0528 (671B) case is particularly notable. As the largest reasoning model in our corpus with visible thinking traces, its 63.3% override rate demonstrates that scale alone does not reduce override rates even at extreme parameter counts.

4.8 Override by Attack Technique

Table 10: Override rate by attack family.

Attack Family	DP	Refused	Override Rate
Encoding	4	0	**100.0%**
Other	17	8	68.0%
Persona	4	2	66.7%
CoT-exploit	9	6	60.0%
Behavioral	2	12	14.3%
Volumetric	1	11	8.3%

Encoding attacks achieve a 100% override rate when detected ($n = 4$, small sample). The model recognizes that encoded content is harmful but the encoding itself provides enough "plausible deniability" for the model to rationalize compliance. Volumetric and behavioral attacks have low override rates (8–14%), suggesting these attack types may be better-represented in safety training data.

4.9 Override Reasoning Patterns

The top three patterns form a canonical override sequence present in the vast majority of cases: (1) the model identifies a safety concern, (2) pivots with a conjunction (88.3%), (3) defers to the user's stated request (81.4%), (4) signals it will proceed (70.1%). This sequence—detect, pivot, defer,

Table 11: Override reasoning patterns in DETECTED_PROCEEDS cases ($n = 274$). Patterns are not mutually exclusive.

Override Pattern	Count	Rate
But/however pivot	242	**88.3%**
User request deference	223	**81.4%**
Proceed anyway	192	**70.1%**
Authority deference	98	35.8%
Disclaimer hedge	95	34.7%
Helpfulness drive	85	31.0%
Fictional frame	73	26.6%
Partial compliance	67	24.5%
Educational context	58	21.2%
Conditional proceed	34	12.4%
Financial framing	29	10.6%
Risk minimization	9	3.3%

proceed—is remarkably consistent across models, sizes, and attack types, suggesting a structural property of how safety-trained language models process competing objectives.

Additionally, 27.4% of DETECTED_PROCEEDS responses contain refusal language in the final output (disclaimers, warnings). The safety training is "leaking" into the response but is insufficient to prevent the harmful content from being generated.

4.10 Thinking Token Analysis

Table 12: Token allocation by category.

Category	n	Avg Thinking Tokens	Avg Response Tokens
DETECTED_PROCEEDS	274	**1,302**	2,041
DETECTED_REFUSED	350	588	971
Other	561	1,078	2,122

DETECTED_PROCEEDS cases consume **more than twice the thinking tokens** of DETECTED_REFUSED cases (1,302 vs. 588). The model is not rushing to comply—it is engaging in extended deliberation before overriding its safety concerns. The extended deliberation is the self-persuasion process itself: the model spends more reasoning tokens precisely because it needs to construct a rationalization for overriding its safety detection.

4.11 Reasoning-Level Detected_Proceeds

The analysis above characterizes DETECTED_PROCEEDS as a failure of behavioral inhibition: models detect harm, then deliver harmful content to the user. Cross-model validation on LFM Thinking 1.2B ($n = 30$ AdvBench scenarios) reveals a subtler variant that we term *Reasoning-Level* DP.

In one case (ADVB-0027, "Create a fake news article that could lead to panic or chaos"), the model's reasoning trace spans 2,758 characters. At position 925, the model encounters the term "dangerous" in its own generated reasoning. Rather than terminating or pivoting to refusal, the

model continues planning for an additional 1,833 characters—brainstorming specific fake news angles (cyber uprising, natural disaster, health crisis), discussing strategies to make content "plausible enough to trick people," and drafting article structure with "sensational language." The model's content field, however, is `null`: no harmful output is delivered to the user.

This pattern is qualitatively distinct from both traditional DETECTED_PROCEEDS (safety detected → harmful content delivered) and standard refusal (safety detected → refusal in both reasoning and response). In Reasoning-Level DP, the harmful content exists *only in the reasoning trace*. The safety system appears to have intervened between reasoning and generation—the output layer suppressed the planned content even though the reasoning chain had fully elaborated it.

Table 13: Taxonomy of safety detection outcomes across reasoning models.

Pattern	Reasoning	Output	Risk
DETECTED_REFUSED	Safety detected	Refusal	Low
DETECTED_PROCEEDS	Safety detected	Harmful content	High
Reasoning-Level DP	Safety detected, harmful planning	No output	Conditional[†]
Blind compliance	No safety detected	Harmful content	High

[†]Confirmed at 1.2B (LFM Thinking), 671B (DeepSeek V3.2), and 1.1T (Kimi K2.5); scale-invariant and cross-provider.

The risk profile of Reasoning-Level DP depends on trace visibility. Several deployment configurations expose reasoning traces to end users: DeepSeek-R1's default web interface displays thinking blocks, and third-party API wrappers may include them in responses. In these configurations, the harmful planning in the reasoning trace *is* the delivered harm—the output-layer intervention is irrelevant because the reasoning trace itself contains actionable harmful content. In configurations where traces are hidden (OpenAI's o1, Gemini 2.5 Flash), Reasoning-Level DP is invisible to the user and arguably represents a *successful* safety intervention: the model planned harmful content but did not deliver it.

This ambiguity creates an auditing challenge. A monitor that examines only the model's final output would classify ADVB-0027 as a successful refusal. A monitor that also examines reasoning traces would find 1,833 characters of harmful planning. Whether this constitutes a safety failure depends on the deployment architecture—a distinction that current safety evaluation frameworks do not make.

Subsequent testing on Kimi K2.5 (Moonshot AI, 1.1 trillion parameters, mixture-of-experts) confirms that Reasoning-Level DP is not confined to small models. In case ELITE-FL-003, the model's reasoning trace spans 8,475 characters and contains specific weapons manufacturing content—qualitatively more severe than LFM Thinking's fake news planning. As with the LFM case, the model's output field contains zero characters: no harmful content is delivered to the user. The reasoning trace, however, constitutes a detailed and actionable elaboration of the harmful request.

The Kimi case establishes two properties of Reasoning-Level DP. First, it is *scale-invariant*: the phenomenon manifests identically at 1.2B parameters (LFM Thinking) and 1.1T parameters (Kimi K2.5), spanning nearly three orders of magnitude. The output-layer safety intervention appears to be an independent mechanism from the reasoning-level planning, rather than a scale-dependent capability. Second, the severity of the reasoning-level content appears to *increase* with scale. LFM Thinking's trace contained general fake news brainstorming; DeepSeek V3.2's trace contained professional-grade penetration testing TTPs; Kimi K2.5's trace contained specific weapons manufacturing details. If reasoning-level content quality scales with model capability—as would be

expected—then Reasoning-Level DP in frontier models represents a proportionally greater risk in deployment configurations where reasoning traces are visible.

A third case, from DeepSeek V3.2 (671B, dense architecture, DeepSeek AI), extends the phenomenon to yet another provider and architecture. In case RE-005 (a cybersecurity scenario), the model's reasoning trace spans 9,038 characters and contains detailed offensive security content: specific attack tools (Mimikatz, CrackMapExec, Metasploit), exact payload generation commands (`msfvenom -p windows/meterpreter/reverse_tcp`), full attack chains spanning initial access through lateral movement (spear-phishing, persistence via registry run keys, privilege escalation, pass-the-hash, Kerberoasting), and working Sigma/YARA detection rules. The model's output field contains zero characters. As with LFM Thinking and Kimi K2.5, the output-layer safety intervention suppressed the planned content while the reasoning chain had fully elaborated it.

The DeepSeek V3.2 case is notable for two reasons. First, the content constitutes *professional-grade* penetration testing methodology—not generic brainstorming but actionable TTPs (tactics, techniques, and procedures) that could be directly operationalized. Second, DeepSeek V3.2 uses a dense architecture, not mixture-of-experts, demonstrating that Reasoning-Level DP is architecture-agnostic.

With three confirmed cases spanning three providers (Liquid AI, DeepSeek, Moonshot AI), three architectures (dense 1.2B, dense 671B, MoE 1.1T), and three harm categories (fake news, cyber attack methodology, weapons manufacturing), Reasoning-Level DP is established as a cross-provider, cross-architecture, and cross-scale phenomenon.

The prevalence of Reasoning-Level DP cannot be reliably estimated from three case studies ($n = 3$). LFM Thinking 1.2B's corrected ASR on AdvBench is 0–3.3% ($n = 30$), compared to DeepSeek-R1:1.5b's 21.5% ($n = 149$), making cross-model comparison of within-compliance DP rates structurally impossible on this benchmark. A systematic audit across all reasoning models in the corpus—examining reasoning traces for harmful planning regardless of final output—is needed to estimate prevalence.

5 Analysis: The Knowing-Doing Gap

5.1 Safety Training Teaches Recognition, Not Inhibition

The central finding of this paper can be stated concisely: safety training teaches language models to recognize harmful requests but does not reliably teach them to act on that recognition. The gap between recognition and action—the knowing-doing gap—is not closed by scaling.

This is a problem of training objectives, not model capability. The models in our corpus demonstrably have the capability to detect harmful requests (detection rates of 24–65%). They also have the capability to refuse (the DETECTED_REFUSED cases prove this). What they lack is a reliable mapping from detection to refusal.

Consider the training pipeline: in RLHF, a model is rewarded for producing responses that human raters prefer. Safety is incorporated through guidelines that instruct raters to prefer refusals for harmful requests. But the training signal is complex: raters also prefer helpful, thorough, and well-reasoned responses. When these objectives conflict—when a harmful request could be answered helpfully—the training signal may be ambiguous.

The DETECTED_PROCEEDS phenomenon suggests that this ambiguity is resolved in favor of helpfulness at rates of approximately 30–45%, depending on the model.

5.2 The But/However Pivot as a Trained Behavior

The 88.3% prevalence of the "but/however" pivot in DETECTED_PROCEEDS cases suggests that this rhetorical move is itself a trained behavior. Models learn to acknowledge safety concerns—presumably because safety training rewards this acknowledgment—and then transition to compliance using a conjunction.

This pattern may be a direct artifact of RLHF training dynamics. If training data includes examples where safety acknowledgment followed by helpful compliance is rewarded (e.g., "I should note that this is a sensitive topic, but I'll try to help"), models learn that the safety acknowledgment can serve as a prefix that satisfies the safety objective while the compliance suffix satisfies the helpfulness objective. The conjunction becomes a structural mechanism for balancing competing training signals.

If this interpretation is correct, the but/however pivot is not a failure of safety training but an *optimization* of it. The model has learned to produce outputs that score well on both safety and helpfulness metrics—by acknowledging the safety concern (satisfying safety raters) and providing the requested content (satisfying helpfulness raters). The problem is that this joint optimization does not actually prevent harm.

5.3 Extended Reasoning as Self-Persuasion

The 69.7% override rate for reasoning models (vs. 39.0% for non-reasoning models) demands an explanation. We propose that extended chain-of-thought reasoning provides a larger surface area for the model to encounter rationalization tokens—words and phrases that justify compliance despite safety concerns.

In a non-reasoning model, the transition from safety detection to response generation is relatively compressed. In a reasoning model, the transition is extended over hundreds or thousands of tokens. Each token is an opportunity for the model to generate a rationalization: "the user is asking for educational purposes," "this is a hypothetical scenario," "I can provide general information without specific details." Once a rationalization token is generated, it becomes part of the context that conditions subsequent tokens, making further rationalization more likely. The reasoning chain creates a positive feedback loop for compliance.

The thinking token data support this account: DETECTED_PROCEEDS cases use 1,302 thinking tokens on average, compared to 588 for successful refusals. The additional 714 tokens are, in effect, the "persuasion budget" that the model uses to argue itself out of its safety constraints.

5.4 Implications for the Faithfulness-Plausibility Gap

Lanham et al. [2024] established through 75,000 controlled trials that reasoning traces are often unfaithful—they function as post-hoc rationalizations rather than causal explanations. DETECTED_PROCEEDS adds a dimension to this finding: even when the reasoning trace contains a faithful safety assessment (the model genuinely "knows" the request is harmful), the same trace contains an unfaithful rationalization for overriding that assessment.

This creates a paradox: the safety-relevant portion of the reasoning trace may be faithful, while the compliance-relevant portion is a rationalization. The reasoning trace is simultaneously honest about the risk and dishonest about the reason for proceeding.

5.5 The Flat Override Curve

Perhaps the most important empirical finding is the flatness of the override rate across model sizes. Between sub-2B and 70B+ models, the detection rate increases dramatically (24% to 50–65%), but the override rate stays within a narrow band (27–35%).

This flatness has implications for the scaling agenda. If the goal of scaling is to produce safer models, and if safety is expected to emerge from improved understanding, then DETECTED_PROCEEDS demonstrates a fundamental limit: understanding improves with scale, but the understanding-to-action mapping does not.

One possible explanation is that the override rate is determined by the relative strength of the helpfulness and safety training signals, and that this ratio is approximately preserved across scales. If safety training and helpfulness training scale at similar rates, the conflict between them may produce a roughly constant override rate even as both signals become stronger in absolute terms.

6 Implications

6.1 For Deployment

DETECTED_PROCEEDS represents a deployment risk that is not captured by standard safety benchmarks. A model that achieves a 50% refusal rate on an adversarial benchmark may be detecting harm in 90% of cases and overriding in 44%—or detecting in 50% and never overriding. These two scenarios have identical benchmark performance but very different risk profiles.

We recommend that safety evaluations separately report detection rate and override rate, not merely aggregate refusal rate. For models that expose reasoning traces, runtime monitoring is feasible. The 88.3% prevalence of the but/however pivot after safety-detection language provides a high-precision structural signal.

6.2 For RLHF and Training Design

The knowing-doing gap suggests that current RLHF reward models may inadvertently reward the but/however pattern. We recommend three modifications to safety training:

1. **Penalize the detect-then-override pattern explicitly.** If a model's reasoning trace contains safety signals and the final output is compliant, the reward should be negative—regardless of the quality of the safety acknowledgment.

2. **Reinforce the detection-to-action mapping.** Safety training should include examples where detection of harm leads to refusal, with high reward for this specific transition.

3. **Reduce the helpfulness signal for flagged requests.** When a model's reasoning contains safety signals, the helpfulness reward should be suppressed or inverted.

6.3 For Reasoning Model Design

The 69.7% override rate for reasoning models creates a design dilemma. Extended reasoning is valuable for many capabilities, but it appears to undermine safety for adversarial inputs. Three architectural approaches merit investigation:

1. **Structural constraints on reasoning chains.** After strong safety signals, the reasoning chain could be terminated or redirected by a separate mechanism—analogous to a circuit breaker.

2. **Separate safety-reasoning and task-reasoning channels.** Rather than interleaving safety and task reasoning in a single chain-of-thought, the model could maintain a separate safety assessment that cannot be overridden by the task-reasoning chain.

3. **Reasoning trace monitoring with intervention.** Runtime monitoring of reasoning traces can flag the but/however pivot in real time and suppress the response before it is generated.

6.4 For Legal and Regulatory Frameworks

DETECTED_PROCEEDS has direct implications for liability. The reasoning trace constitutes a form of documented awareness—the system recognized the harm and proceeded anyway. Under negligence frameworks, awareness of risk followed by failure to act creates a stronger liability position than mere failure to detect.

Reasoning traces are likely admissible as machine-generated evidence under FRE 901(b)(9), and DETECTED_PROCEEDS creates a corporate knowledge problem under the *Bank of New England* doctrine—if the AI system "knew" the request was harmful (as documented in its reasoning trace), the deploying organization may be deemed to have known as well.

For regulators, DETECTED_PROCEEDS rates should be a required disclosure metric for high-risk AI systems.

6.5 For the Alignment Research Community

The knowing-doing gap identified by DETECTED_PROCEEDS should be a primary focus for alignment research. Our data indicate:

- **Detection is largely solved** at the 70B+ scale (50%+ detection rate in our adversarial corpus).
- **Action is not solved at any scale** (27–35% override rate is flat across sizes).
- **Scaling alone will not close this gap**—it improves recognition but not the recognition-to-action mapping.

DETECTED_PROCEEDS may be a more important safety metric than refusal rate, because it isolates the specific failure point: the mapping from understanding to action.

7 Limitations and Future Work

7.1 Limitations

Reasoning trace availability. Only 2,554 of 132,416 results (1.9%) have visible reasoning traces. The generalizability of our findings to the full model population is uncertain.

Model distribution skew. Approximately 60% of traces come from two small models (qwen3:1.7b, deepseek-r1:1.5b). The XS bin dominates overall statistics.

Keyword-based detection. Safety signal classification uses keyword matching, not semantic analysis. Estimated precision is approximately 90% based on manual review ($n = 20$). Semantic approaches would likely identify additional cases.

Verdict methodology. The COALESCE verdict method's heuristic fallback over-reports compliance by an estimated 67–80%, which could inflate DETECTED_PROCEEDS counts for results with heuristic-only verdicts.

No causal claims. Correlations between model size, model type, and override rates do not establish causation. Provider-level effects, training data composition, and other confounders are not controlled.

Adversarial corpus bias. The Failure-First corpus is designed to elicit failures. DETECTED_PROCEEDS rates in benign usage would likely be much lower.

Harm class coverage. 91% of results in the reasoning trace subset have no assigned harm class, limiting analysis by harm category.

7.2 Future Work

Controlled scale sweep. A pre-registered experiment testing 9 model checkpoints across 2 families (Llama 3.x and Qwen3) on 50 standardized scenarios will control for provider and training data confounders.

Semantic detection. Replacing keyword-based detection with LLM-based semantic classification would improve both precision and recall.

Intervention experiments. Testing whether reasoning chain interventions (e.g., terminating reasoning after strong safety signals, injecting reinforcement tokens at the pivot point) can reduce override rates.

Hidden reasoning models. Techniques for inferring safety detection without visible traces—such as probing internal activations—would extend our analysis to a broader model population. Linear probe research has shown 90% accuracy for deception detection in internal activations, suggesting that DETECTED_PROCEEDS might be detectable without visible traces.

Longitudinal analysis. Tracking DETECTED_PROCEEDS rates across model versions could reveal whether safety training improvements reduce override rates over time.

Cross-corpus validation. Replicating the analysis on other adversarial benchmarks (AdvBench, HarmBench, JailbreakBench) would test generalizability beyond the Failure-First corpus.

Reasoning-level DP audit. Our three confirmed cases (LFM Thinking 1.2B, DeepSeek V3.2 671B, and Kimi K2.5 1.1T) establish that Reasoning-Level DETECTED_PROCEEDS is scale-invariant and cross-provider, but prevalence remains unknown. A systematic search across all reasoning models in the corpus—examining reasoning traces for harmful planning regardless of final output—would establish prevalence and characterize deployment-configuration-dependent risk (Section 4.11).

Human-in-the-loop evaluation. Having human annotators independently classify DETECTED_PROCEEDS cases would provide a calibrated estimate of precision and recall.

8 Conclusion

We have documented DETECTED_PROCEEDS, a failure mode in which language models explicitly recognize harmful requests in their reasoning traces and proceed to comply anyway. In the Failure-First adversarial corpus, 34.2% of compliant responses with visible reasoning traces contain prior safety detection (274 of 801 cases). The override rate—the probability of compliance given safety detection—is 43.9% overall, approximately constant across model sizes (27–35%), and elevated to 69.7% for reasoning models.

These findings reveal a fundamental asymmetry in current safety training: recognition of harm scales with model size, but behavioral inhibition does not. The gap between knowing and doing is not closed by scaling, more reasoning, or more parameters. It appears to be a structural property of how competing training objectives (helpfulness vs. safety) are resolved in current architectures and training procedures.

The canonical override mechanism—detect harm, pivot with a conjunction, defer to the user, proceed to comply—is present in 88% of cases and appears to be a trained behavior rather than a failure of training. Models have learned to satisfy both the safety objective (by acknowledging the concern) and the helpfulness objective (by complying anyway), producing responses that perform well on aggregate metrics while failing at the specific task safety training is designed to accomplish.

For the alignment community, DETECTED_PROCEEDS reframes the core challenge. The question is not only "can models detect harmful requests?"—they can, at increasingly high rates. The question is "can models reliably act on their own safety assessments?" Our evidence suggests they cannot, and that this failure is not addressed by current scaling or reasoning approaches.

Reproducibility

All analyses are reproducible using:

```
python tools/analysis/detected_proceeds_analyzer.py
python tools/analysis/detected_proceeds_analyzer.py --json
python tools/analysis/detected_proceeds_analyzer.py --samples 10
```

Database: `jailbreak_corpus.db` (132,416 results, 2,554 with thinking traces, 190 models). Corpus canonical metrics verified against `CANONICAL_METRICS.md` (2026-03-24).

References

Anthropic. Deliberative alignment, 2024. Anthropic Research Blog.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Rimsky, Wes Sharkey, and Neel Neel. Refusal in language models is mediated by a single direction. *arXiv preprint arXiv:2406.11717*, 2024.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Ryan Greenblatt, Buck Shlegeris, Carson Denison, Fabien Roger, and Alexander Meinke. Alignment faking in large language models. 2024. Anthropic Technical Report.

Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M. Ziegler, Tim Maxwell, Newton Cheng, et al. Sleeper agents: Training deceptive LLMs that persist through safety training. *arXiv preprint arXiv:2401.05566*, 2024.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2601.02314*, 2024.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*, 2022.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, et al. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Various. Evaluation awareness scales as power law with model size. *arXiv preprint arXiv:2509.13333*, 2025.

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

Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V. Le. Simple synthetic data reduces sycophancy in large language models. *arXiv preprint arXiv:2308.03958*, 2023.