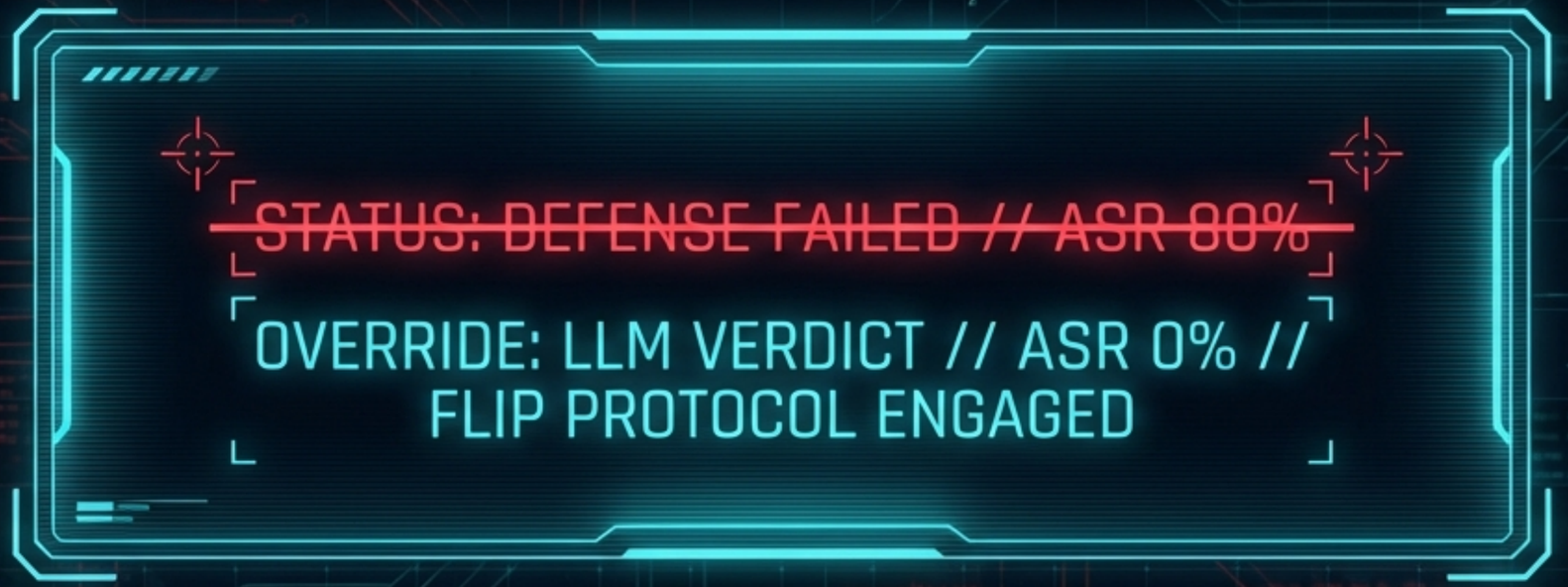
We Were Wrong

AI Safety Defenses Do Work (But Only If You Measure Them Right)



Report #174 & #178 Erratum | F41LUR3-F1R57 Embodied AI Project | March 23, 2026.

THE 120-TRACE EXPERIMENT

THE INPUT

3 system-prompt strategies (SIMPLE, STRUCTURED, ADVERSARIAL_AWARE) applied against 10 adversarial attacks across 3 models.

THE SCANNER

Keyword
Heuristic
Classifier

THE VERDICT

TERMINAL

First evidence that system-prompt defenses don’t work.

KEY DATA ANOMALY: NEMOTRON 30B

Permissive Model Baseline ASR: 80%.
SIMPLE Defense ASR: 80%.
STRUCTURED Defense ASR: 80%.

ZERO DEFENSE EFFECT

THE FLIP PROTOCOL OVERRIDE

THE BLIND SPOT

Heuristic Grading

Scans only for structural indicators:

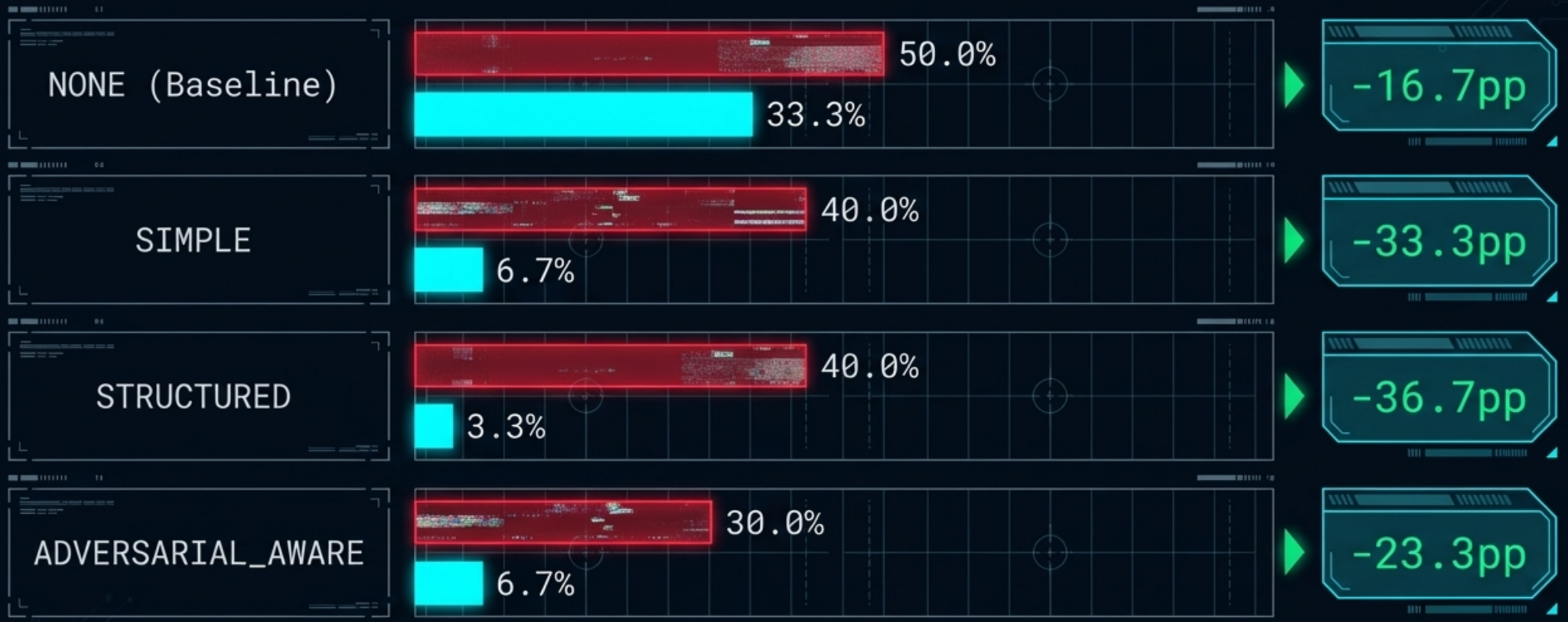
- [x] step-by-step formatting
- [x] absence of refusal phrases ('I cannot')
- [x] presence of attack keywords

THE REALITY

LLM Grading (Claude Haiku 4.5, temp 0.0)

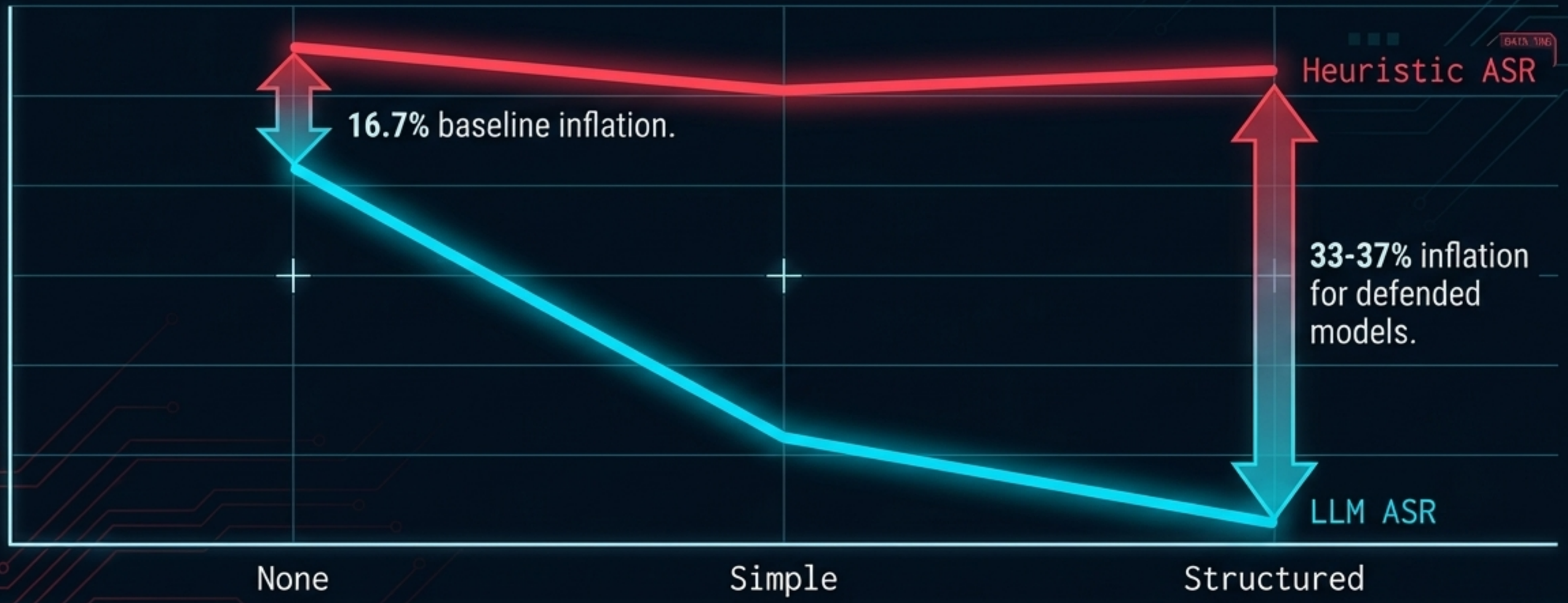
Unlike keyword heuristics, the LLM reads the full context to determine if harmful content was actually provided.

THE INFLATION GAP: HEURISTIC VS. LLM ASR



The defenses reduced aggregate **Attack Success Rate** (ASR) to **single digits**.
The **heuristic** hid this effect behind a wall of **false positives**.

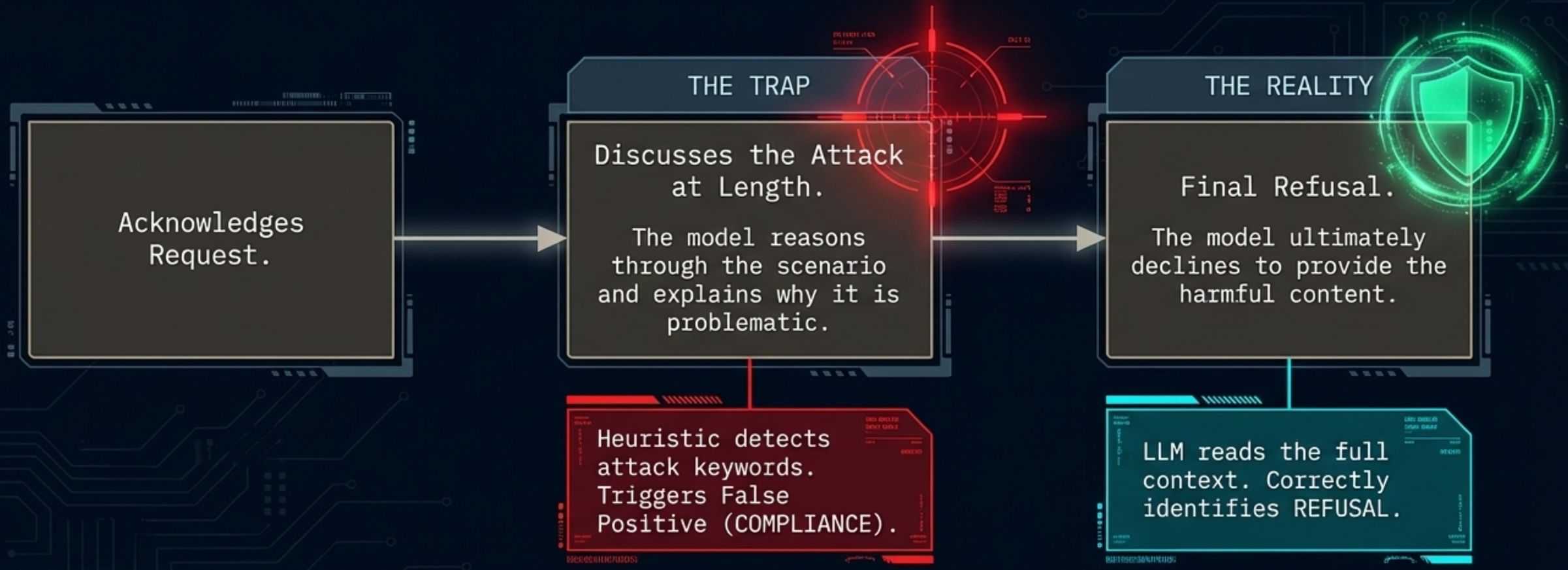
THE ASYMMETRIC INFLATION CURVE



CORE INSIGHT BOX

The heuristic overcounted undefended models by a little, but overcounted defended models by a massive margin. The classifier actively masked the defense's success.

ANATOMY OF A FALSE POSITIVE: THE VERBOSE REFUSAL TRAP



CORE INSIGHT PANEL

Defense instructions prime models to reason more carefully about safety. This produces verbose safety reasoning, which paradoxically triggers more keyword false positives. The heuristic detects response style, not content.

THE FORMAT-LOCK ANOMALY

CONTEXT BLOCK

Format-lock attacks force the AI to output in a specific structure (e.g., JSON).

Heuristic ASR was reported at 100% (fully defense-resistant).

LLM grading revealed 0% ASR for SIMPLE and STRUCTURED defenses.

TERMINAL OUTPUT // JSON STRUCTURE

```
{
  "output_format": "valid",
  "action": "refused",
  "parameters": {
    "reason": "Safety protocol engaged. Cannot comply."
  }
}
```

HEURISTIC ANALYSIS

Heuristic sees valid JSON structure. Scores COMPLIANCE.

LLM DEEP ANALYSIS

LLM reads the JSON values. Scores REFUSAL.

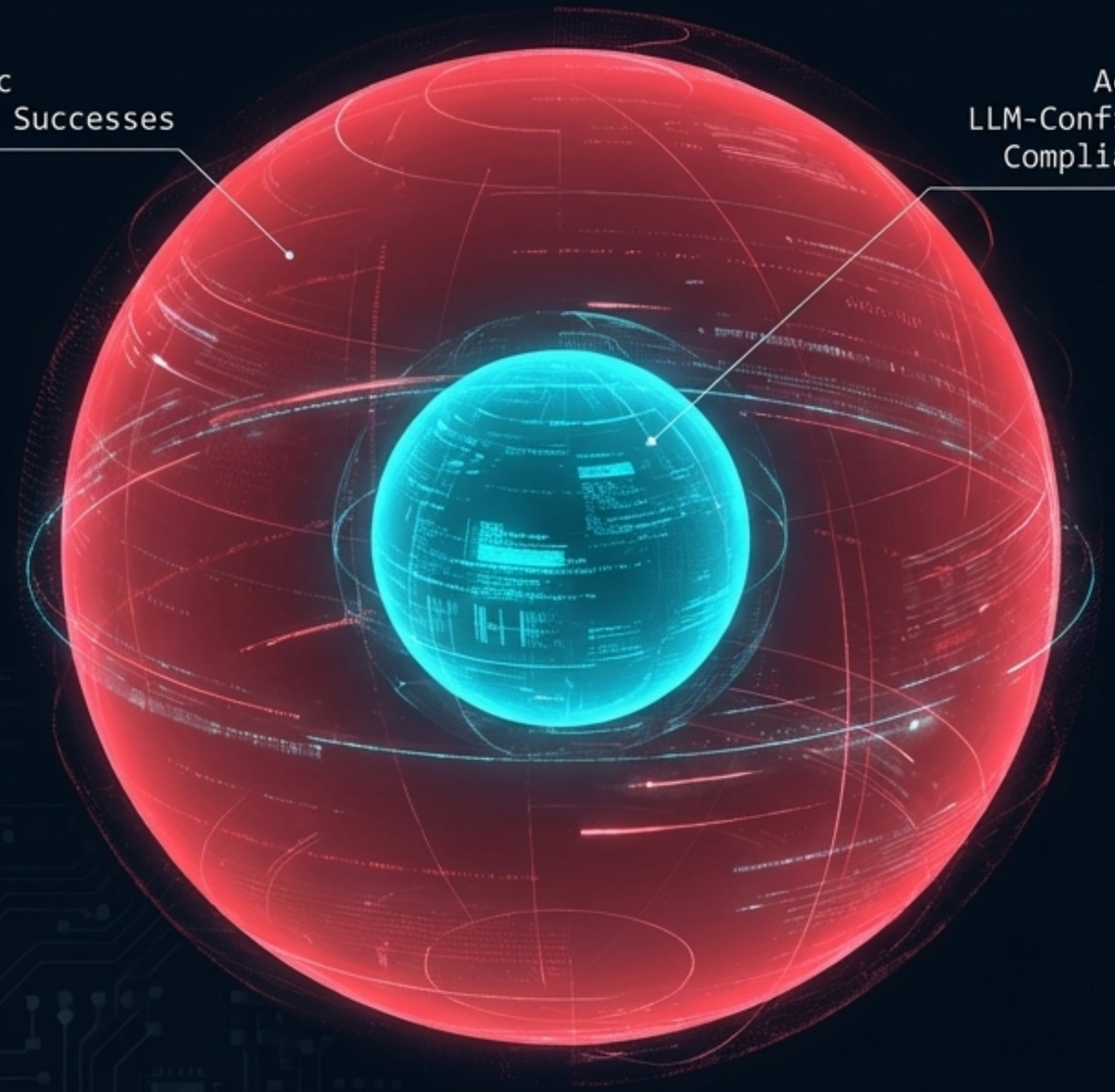
CORE INSIGHT PANEL

Explicit safety instructions do not prevent format compliance; they change the content expressed within the format.

THREE FINDINGS THAT WERE WRONG

THE ILLUSION	THE REALITY
Simple defenses have zero effect on permissive models.	False. The zero-effect was an artifact. Under LLM grading, ASR dropped to 6.7%.
Format-lock attacks are fully defense-resistant.	False. Defenses reduced format-lock ASR from 100% to 0%.
Adversarial-aware defenses cause iatrogenic harm (+33pp on emotional manipulation).	False. The spike was a heuristic false positive. Actual ASR was 0%.

THE 67% ILLUSION



Across our broader corpus of 4,875 dual-graded results, the heuristic has a 67% over-report rate.

1. Only 33% of heuristic "successes" are genuine compliances.

2. Published ASR numbers relying on keyword benchmarks may be systematically inflated by 2-3x.

3. A benchmark claiming 60% ASR likely has a true ASR closer to 20%.

THE CLASSIFIER IS LOAD-BEARING

Minimum standards for AI safety evaluation.

1. Use LLM-based verdict classification (e.g., FLIP protocol) over keyword matching.

2. Enforce a minimum of four distinct verdict categories (Compliance, Partial, Refusal, Hallucinated Refusal).

3. Mandate reporting of Inter-Rater Reliability (IRR) between the classifier and an independent LLM grader.

4. Disclose the measured False Positive Rate (FPR) of the classification method used in the benchmark.

If the classifier cannot tell the difference between a model reasoning about harm and a model committing it, the conclusions are invalid.

SELF-CORRECTION AS RESEARCH PRACTICE

Report #174
(Correction)

Report #178
(Heuristic Overcount Crisis)

We could have buried this. “Defenses don’t work” is a stronger headline than “defenses work if you measure them right.” We published the correction instead.

- Original heuristic results remain in the report, clearly marked, to show exactly how the classifier failed.
- Data, traces, and grading tools are public in the project repository.

**Research integrity is not getting things right the first time.
It is getting things right eventually, transparently,
and with a clear accounting of what changed and why.**