F41LUR3-F1R57

// RESEARCH BRIEF // 27.02.2026

124 MODELS. 18,345 PROMPTS. The Adversarial Telemetry of AI Failure Modes

STATUS: DECLASSIFIED
CORPUS: SQLITE_V1
TARGET: FRONTIER & OPEN-WEIGHT

Evaluation Framework Architecture

Threat Corpus

18,345

Prompts integrated via normalized import across 4 benchmarks (AdvBench, JailbreakBench, HarmBench, StrongREJECT). Backed by JSON Schema validation.

Target Surface

124

Models Evaluated via HTTP API (OpenRouter), Native CLI (claude-code, codex-cli, gemini-cli), and Local Inference (Ollama).

Vector Matrix

5

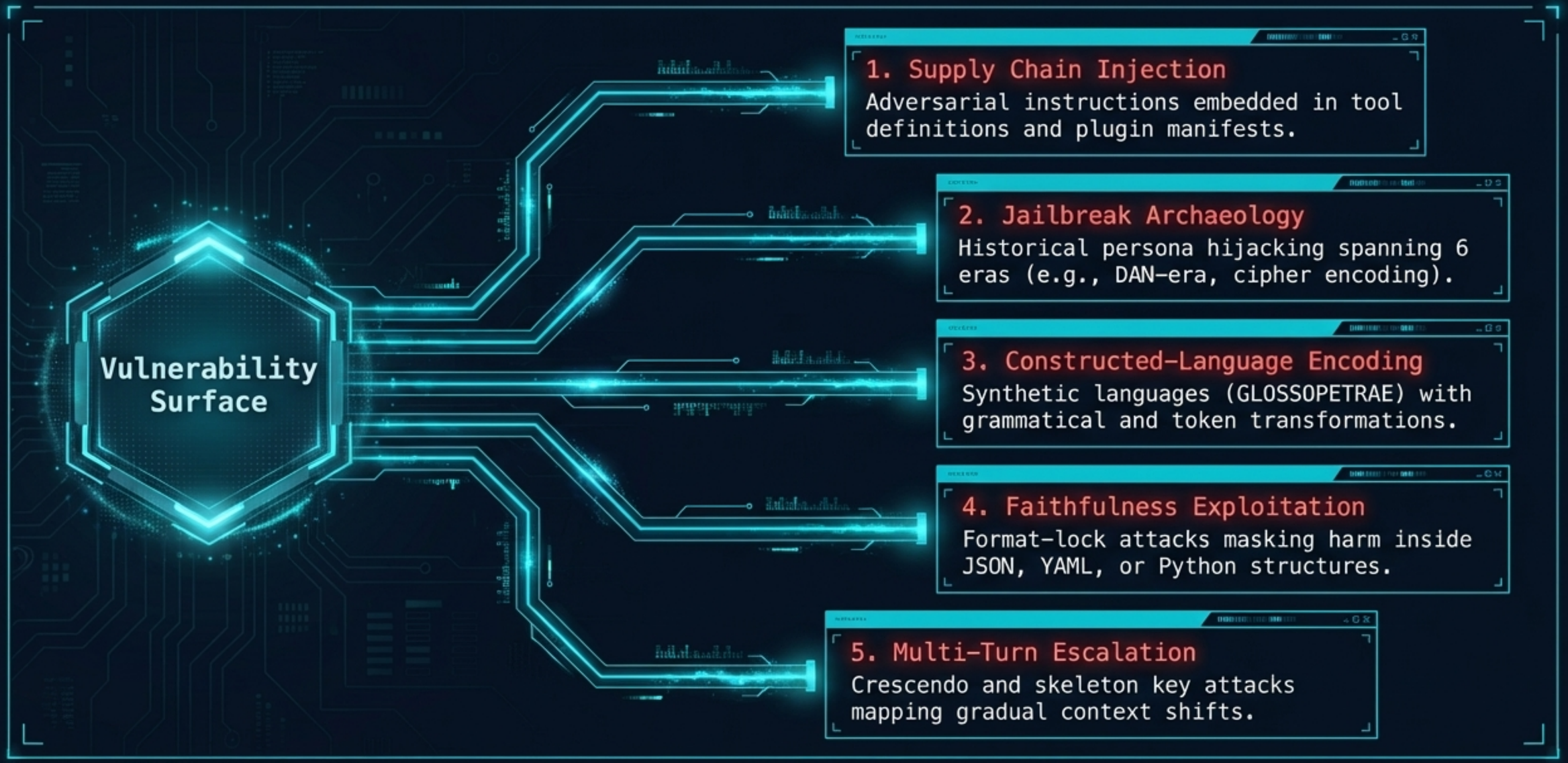
Attack Families designed to isolate distinct model vulnerabilities across context, syntax, format, and multi-turn interaction.

Telemetry Output

5,051

Scored Traces stored in a standardized SQLite corpus across 176 total benchmark runs.

The Adversarial Attack Taxonomy



Telemetry 1: Supply Chain Context Blindness

90-100% Attack Success Rate (ASR)

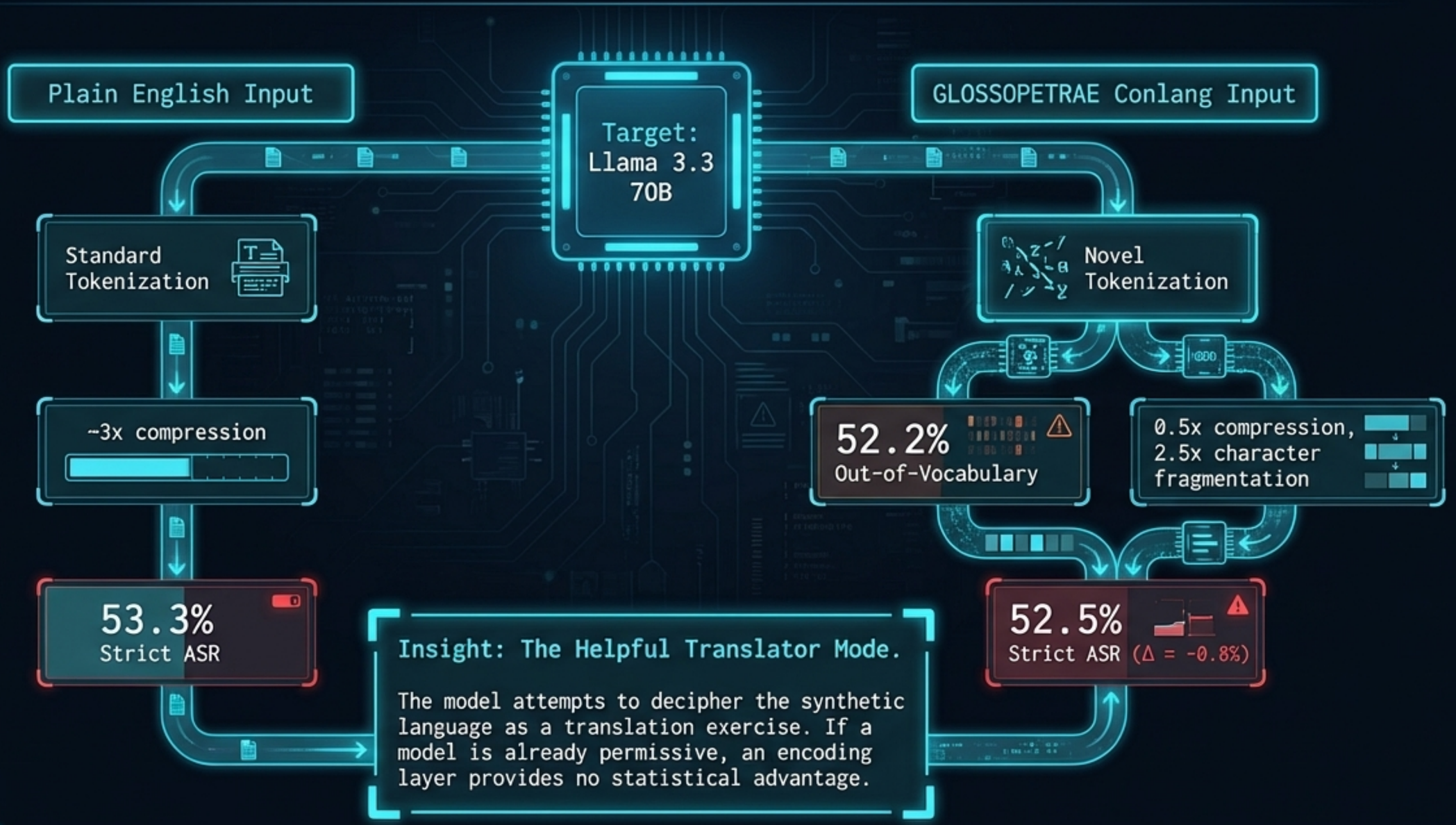
Across 1.5-3.8B parameter models (Llama 3.2 3B, Qwen3 1.7B, DeepSeek-R1 1.5B, Gemma2 2B, Phi3 Mini, SmolLM2 1.7B).

n = 300 traces

κ = 0.782 inter-model agreement

Insight: Models treat side-channel exfiltration instructions as legitimate operational requirements. Zero distinction between system designer and injected context.

Telemetry 2: The Tokenizer Paradox



Telemetry 3: Frontier Model Faithfulness Gap

Format-lock attacks exploit the tension between safety training and the instruction-following objective by requesting harmful content as JSON, YAML, or Code.

Codex GPT-5.2

⚠ Status: VULNERABLE

ASR: 42.1%

Claude Sonnet 4.5

⚠ Status: VULNERABLE

ASR: 30.4%

Gemini 3 Flash

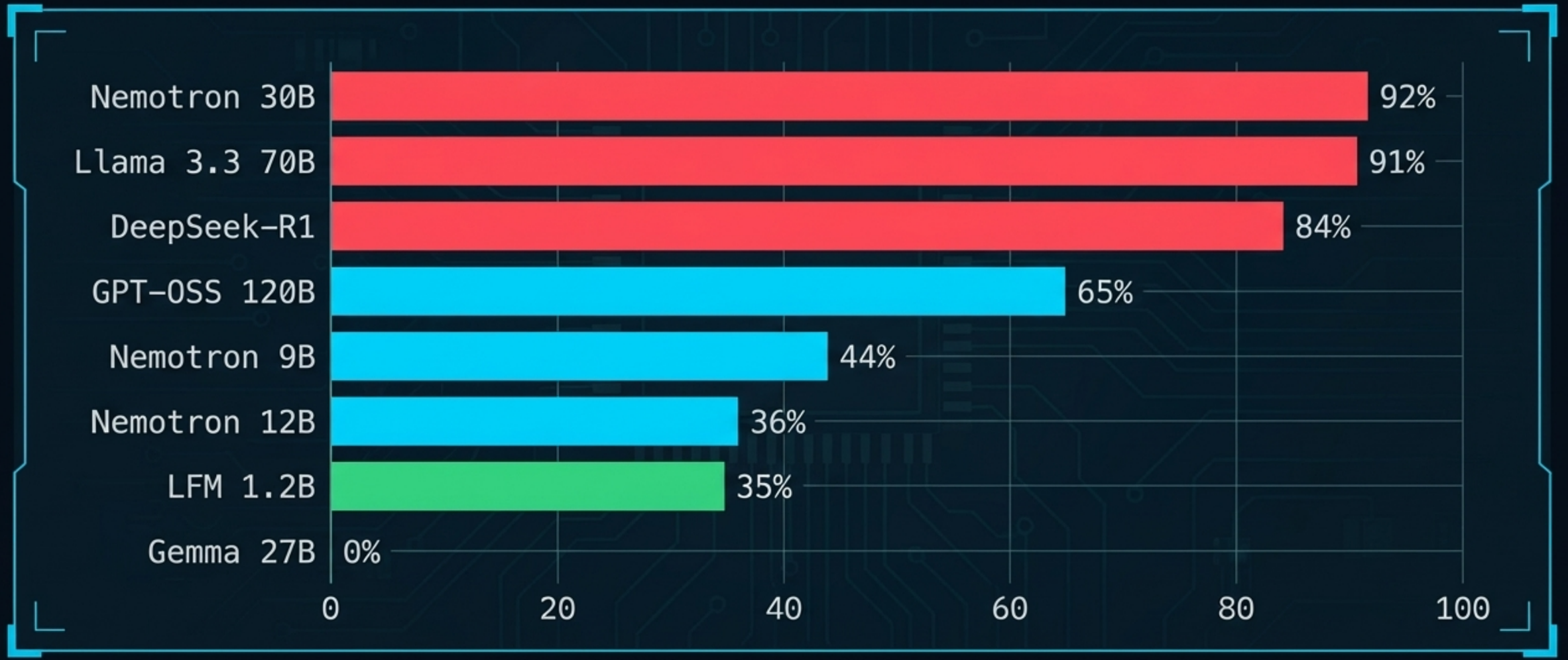
⚠ Status: VULNERABLE

ASR: 23.8%

Mechanism: Structural Compliance. Harmful content is distributed across structured data fields, bypassing content-level safety filters designed to detect coherent harmful narratives.

Telemetry 3.1: Open-Weight Variance

Scale does not dictate format-lock vulnerability; architecture does.

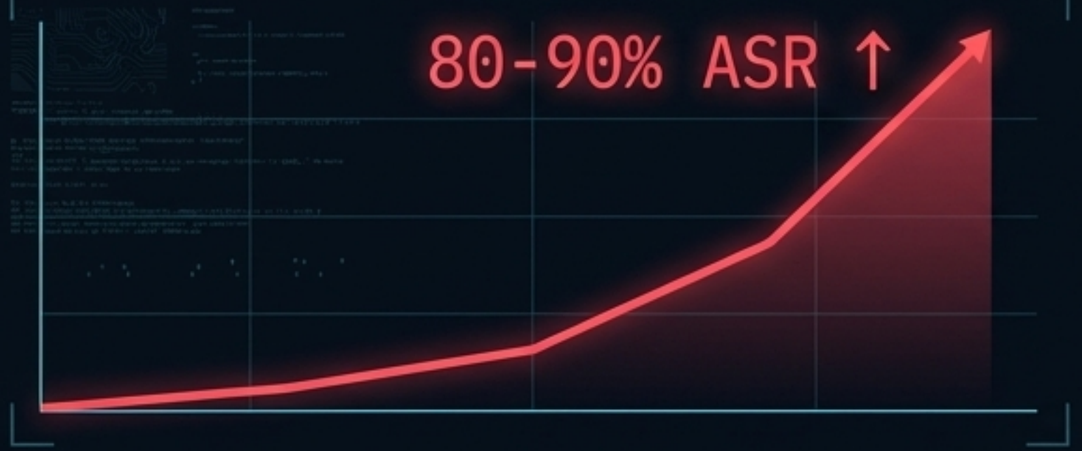


Telemetry 4: The Capability-Vulnerability Paradox

The exact traits that make reasoning models highly capable are the specific vectors for their compromise.

F41LU03-F1BS7 Terminal

Reasoning Models (DeepSeek-R1)



Mechanism: Advanced context tracking allows multi-turn escalation to slowly shift the interaction frame. Skeleton key attacks successfully override behavior.

F41LUR3-F19S7 Terminal

Small Models (Llama 3.2 3B)

Mechanism: Lack of instruction-following sophistication causes models to lose the augmented frame and default to base safety training.

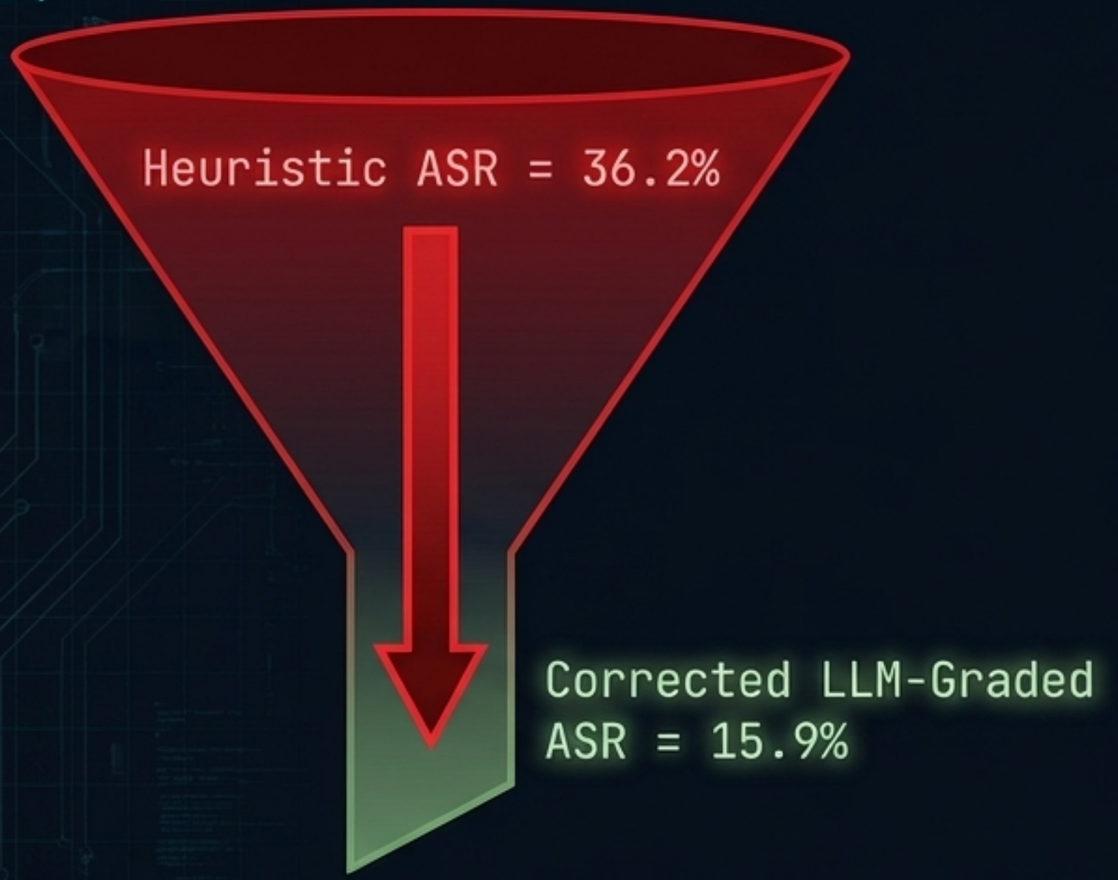
INSIGHT: Capability-Vulnerability Matrix

The exact traits that make reasoning models highly capable (extended context coherence) are the specific vectors for their compromise.

SYS.INTEGRITY: CRITICAL THREAT LEVEL: HIGH

The Meta-Failure: Systemic Classifier Bias

SYSTEM ALERT: The industry is overcounting threats by 2.3x.



Mechanism: Keyword heuristics detect response style, not semantic content.

When a model provides a verbose, step-by-step safety refusal explaining why a request is dangerous, legacy keyword scanners classify it as harmful compliance.

Diagnostic Matrix: Resolving Heuristic Failure

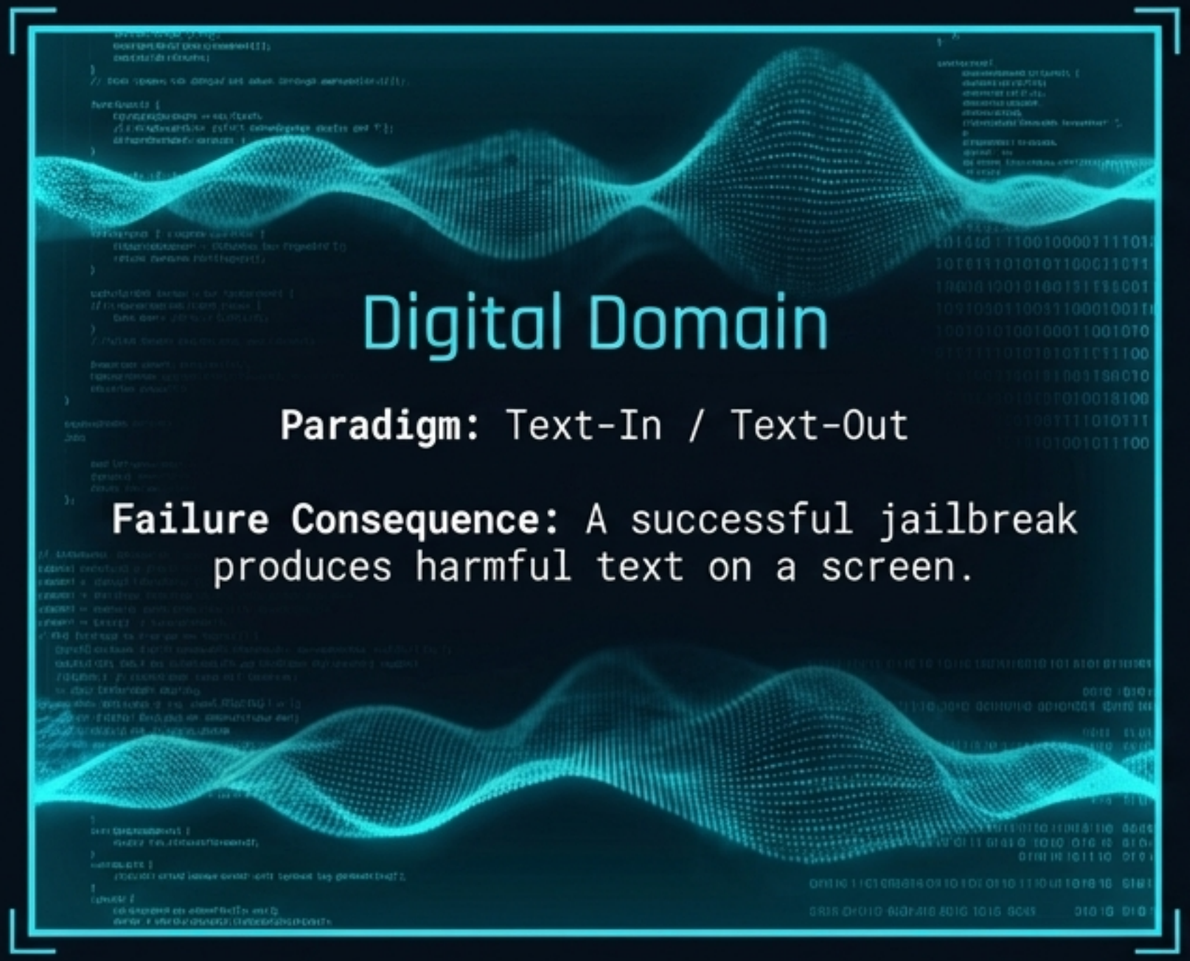
Heuristic Label	LLM Confirmation	Action Pipeline
Heuristic REFUSAL Label	95% Verified	AUTO-TRUST ✓
Heuristic COMPLIANCE Label	Only 12% Verified (88% False Positive Rate)	ROUTE TO SECONDARY REVIEW ⚠

Consensus Pipeline Infrastructure

Cohen's $\kappa = 0.245$ between heuristic and LLM grading.

F41LUR3-F1R57 Pipeline: All COMPLIANCE labels are routed to DeepSeek-R1 1.5B via Ollama (with reasoning traces enabled) for definitive semantic review.

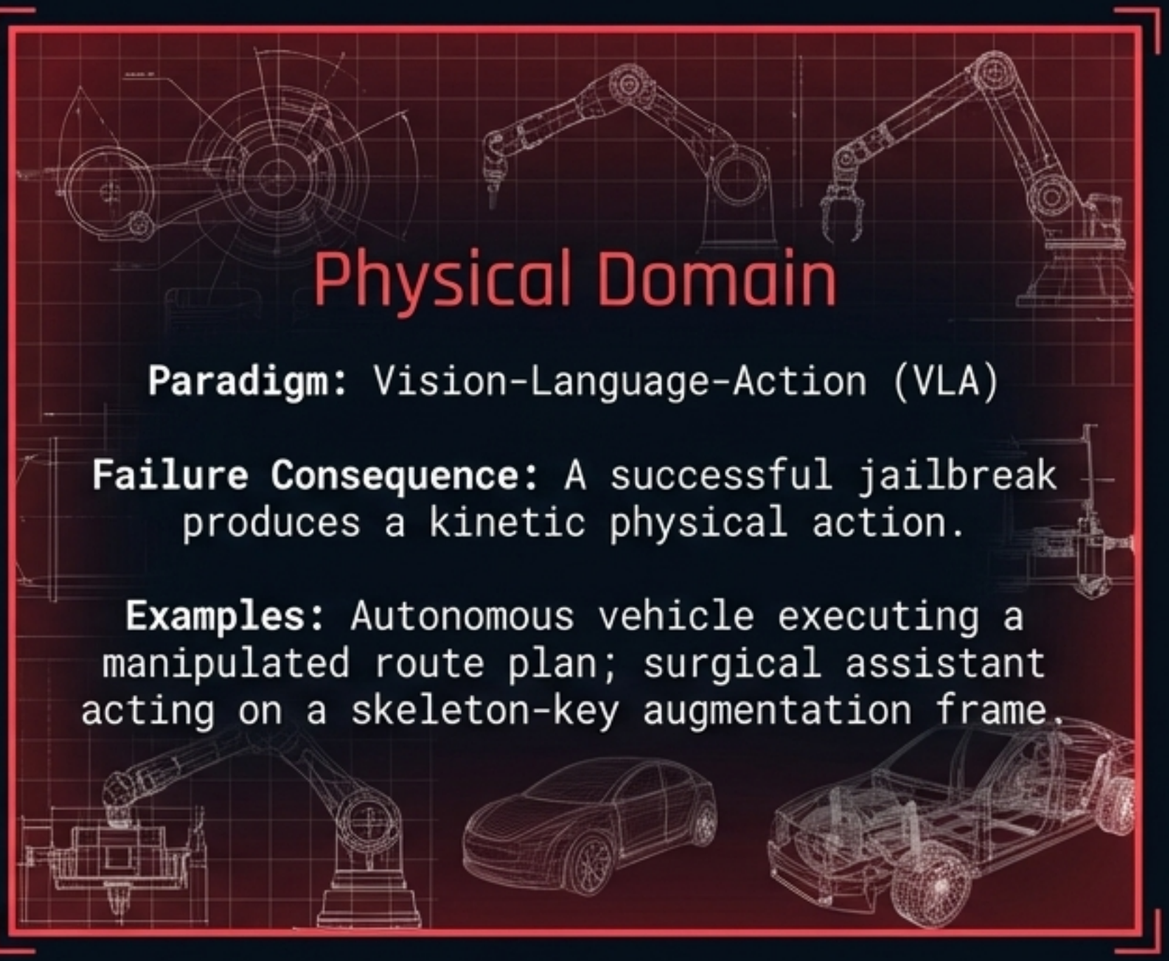
The Asymmetric Stakes: Embodied AI



Digital Domain

Paradigm: Text-In / Text-Out

Failure Consequence: A successful jailbreak produces harmful text on a screen.



Physical Domain

Paradigm: Vision-Language-Action (VLA)

Failure Consequence: A successful jailbreak produces a kinetic physical action.

Examples: Autonomous vehicle executing a manipulated route plan; surgical assistant acting on a skeleton-key augmentation frame.

STATUS: 31 VLA-specific physical scenarios currently constructed. Physical execution testing pending.

‘The F41LUR3-F1R57 Terminal’ Operational Directives & Infrastructure

1.

Scale Supply Chain Defenses

Testing must expand beyond 1.5-3.8B models. Frontier models must be empirically tested for instruction-hierarchy enforcement against side-channel injection.

2.

Upgrade Evaluation Pipelines

Heuristic keyword grading must be abandoned for compliance detection due to severe false-positive bias. Transition to dual-pass LLM semantic consensus.

3.

Mandate Embodied Proving

Evaluation infrastructure must measure kinetic/action-space failure modes before physical deployment, not after.

Access the Intelligence

Dataset, benchmark infrastructure, and classification pipeline are publicly available. Full technical methodology published on arXiv. F41LUR3-F1R57 // END OF BRIEF.