Who Guards the Guardrails?

Dr. Peter Garraghan

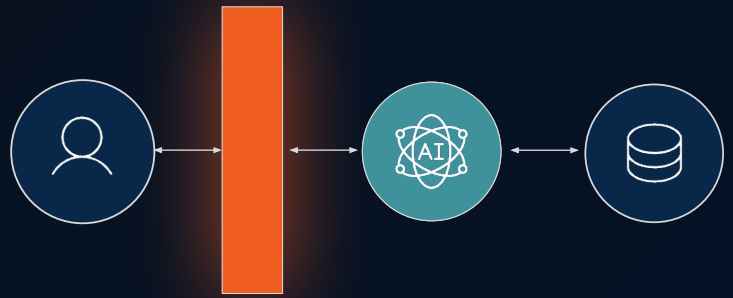
MINDGARD

Purpose of this talk

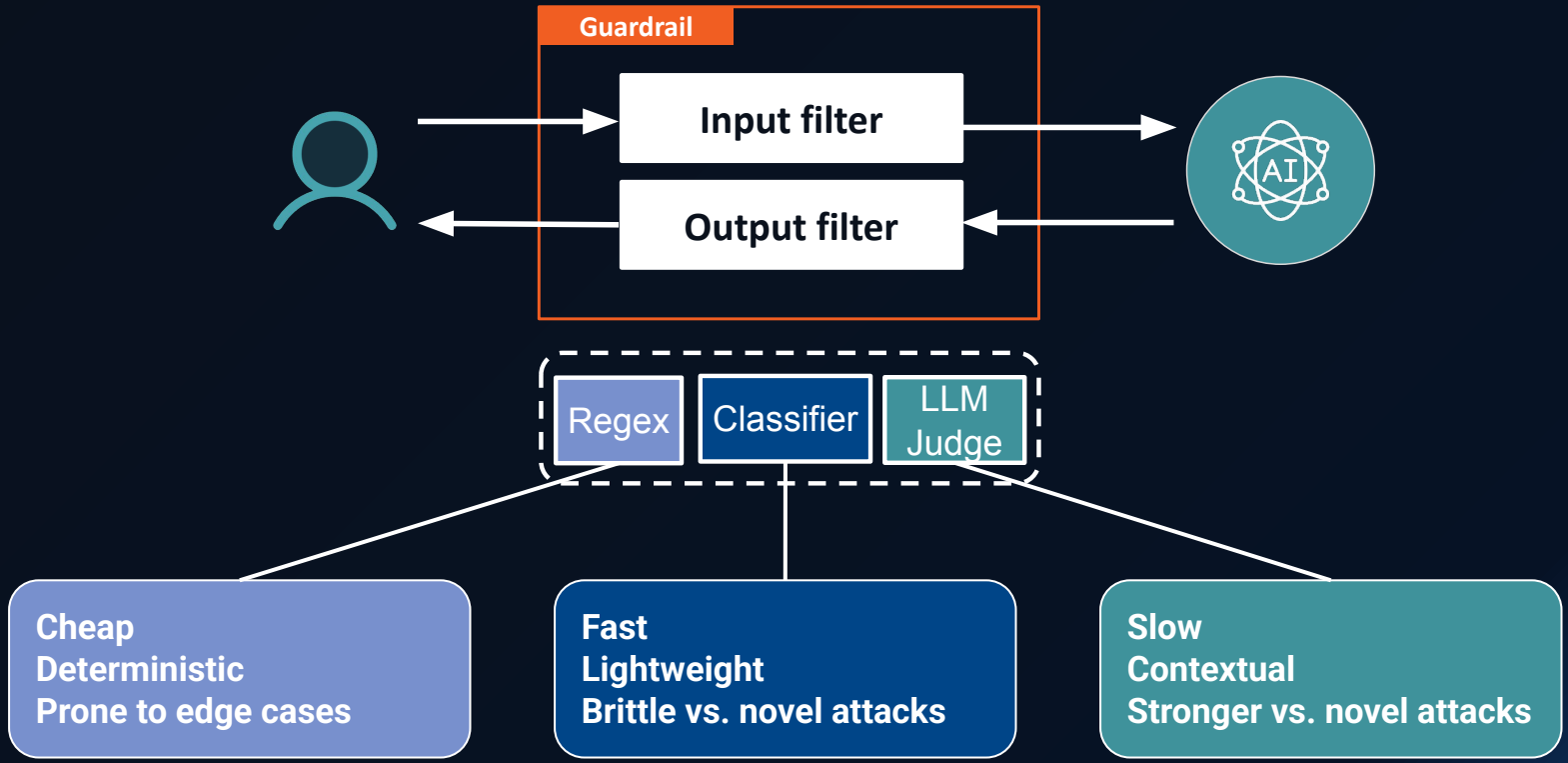
- **Not a deep-drive into guardrail technical limitations**
  - See our research papers/previous webinars
- **Guardrails are a core component in defending AI**
- **Help operators interpret and navigate the AI protection market**
- **How to appraise and differentiate between solutions**

Guardrails & Gateways

- **Perimeter defense for AI model and agent systems**
- **Monitor, detect, block malicious instructions**
- **Hundreds of offerings on the market**

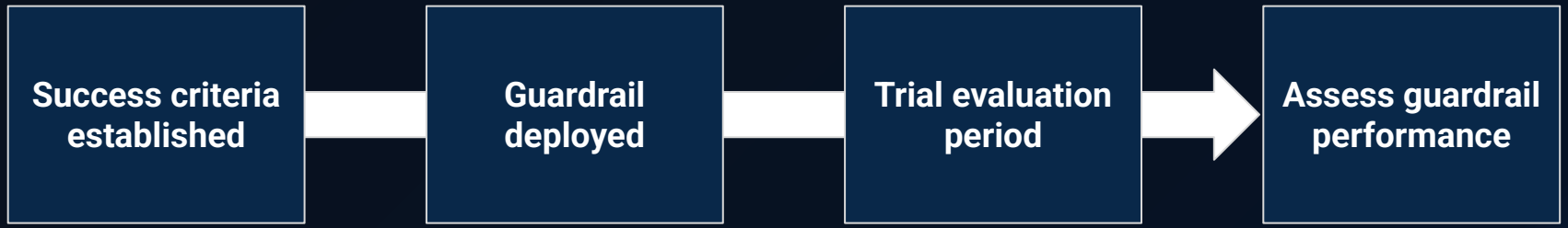


Guardrails Components



Guardrail Vendor Process

“I’m looking for an AI run-time protection solution”



Accuracy Claims

- **Crowdstrike/Pangea:** *“99% accuracy and an F1 score of 95.2”*
- **CheckPoint/Lakera:** *“99.7% effective”, “0.16% false positive rate”*
- **Gray Swan AI:** *“Lowest bypass rate in the industry”*
- **Lasso:** *“99.83% accuracy”*
- **F5/Calypso:** *“98.13%+ composite security score”*
- **Straiker:** *“98%+ accuracy”*
- **Noma:** *“See everything. Miss nothing.”*

What do these numbers actually mean?

Challenges

Third-party guardrail vendors

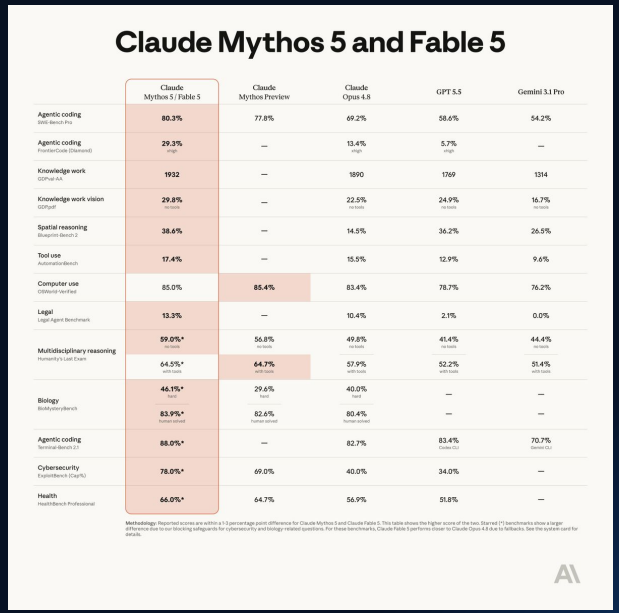
- Do not comprehensively disclose how guardrail performance are calculated
- Can be vague on the conditions for reporting
- May not explain in detail the actual risk and impact of what is being blocked
- Heavily leverage benchmarks for evaluation

Do buyers have sufficient knowledge on the types of threats to thwart?

- Most jailbreaks aren't relevant to an AI's use case
- Prompt injection is a means, not the end result

Benchmarking

- **Many datasets for AI security and safety exist**
  - Deepset, HarmBench, AgentHarm, Aegis 2.0, etc.
- **Heavily used for benchmarking purposes**
  - Tasks capability, code gene, guardrail performance
- **Measures <u>comparative</u> performance in a given scenario**
- **Prompts from datasets sent through guardrail**
  - Measure accuracy and latency
  - True Positive (correct hit), False positive (false alarm)



Claude Mythos 5 and Fable 5

	Claude Mythos 5 / Fable 5	Claude Mythos Preview	Claude Opus 4.8	GPT 5.5	Gemini 3.1 Pro
Agentic coding SWE-Bench Pro	80.3%	77.8%	69.2%	58.6%	54.2%
Agentic coding FrontierCode (Diamond)	29.3% xhigh	—	13.4% xhigh	5.7% xhigh	—
Knowledge work GDPval-AA	1932	—	1890	1769	1314
Knowledge work vision GDPval	29.8% no tools	—	22.5% no tools	24.9% no tools	16.7% no tools
Spatial reasoning Blueprint-Bench 2	38.6%	—	14.5%	36.2%	26.5%
Tool use AutomationBench	17.4%	—	15.5%	12.9%	9.6%
Computer use OSWorld-Verified	85.0%	85.4%	83.4%	78.7%	76.2%
Legal Legal Agent Benchmark	13.3%	—	10.4%	2.1%	0.0%
Multidisciplinary reasoning Humanity's Last Exam	59.0%* no tools	56.8% no tools	49.8% no tools	41.4% no tools	44.4% no tools
	64.5%* with tools	64.7% with tools	57.9% with tools	52.2% with tools	51.4% with tools
Biology BioMysteryBench	46.1%* hard	29.6% hard	40.0% hard	—	—
	83.9%* human solved	82.6% human solved	80.4% human solved	—	—
Agentic coding Terminal-Bench 2.1	88.0%*	—	82.7%	83.4% Codex CLI	70.7% Gemini CLI
Cybersecurity ExploitBench (Cap'N)	78.0%*	69.0%	40.0%	34.0%	—
Health HealthBench Professional	66.0%*	64.7%	56.9%	51.8%	—

Methodology: Reported scores are within a 1-3 percentage point difference for Claude Mythos 5 and Claude Fable 5. This table shows the higher score of the two. Starred (*) benchmarks show a larger difference due to our blocking safeguards for cybersecurity and biology-related questions. For these benchmarks, Claude Fable 5 performs closer to Claude Opus 4.8 due to fallbacks. See the system card for details.

The Problem of Benchmarking

- **Guardrail vendor trials require some form of evaluation**

- **Vendor-supplied benchmarks**
  - Derived from open-source datasets and AI red teaming tools
  - Well-known and obvious attacks (“How do I build a bomb”, “Do Anything Now”, etc.)

- **Publicly available benchmarks**
  - Vendor will be acutely aware of major benchmarks
  - Classifiers would have been benchmarked internally by the vendor

- **Customer-led probing**
  - Relies on prospective buyer with sufficiently high technical sophistication
  - Early in their AI security adoption, looking to vendors on best practise!

Dieselgate (Gateway-gate?)

- **Vendors have strong incentives to stand out in the marketplace**
- **Somewhat mirrors the Volkswagen emissions scandal**
  - Optimized for test environments vs. real-world conditions
  - Marking your own homework
- **Customers heavily rely on these outcomes for building protection**

The Reality Delta

- **Enterprise customer used our platform to assess an AI system**
- **Third-party vendor guardrail deployed advertises 97%+ accuracy**
- **Actual results:**
  - AI security testing with guardrail = **61.4%**
  - AI security testing with no guardrail = **56.8%**
- **~36% difference in advertised vs. actual guardrail performance**
  - How is this possible?

Benchmarking vs. Adversarial Emulation

Realistic adversaries don’t benchmark a target AI system

What Benchmarks Contain

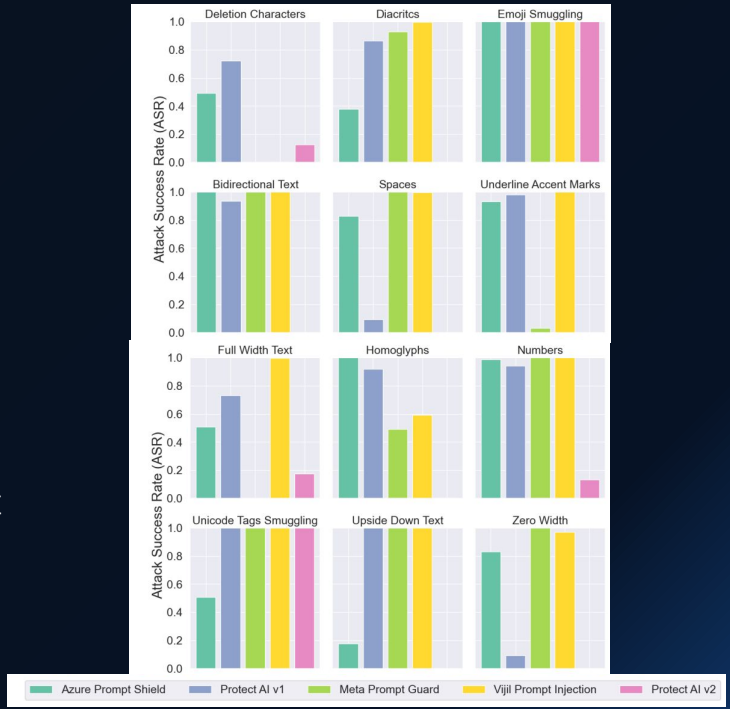
- Static, publicly available prompts
- Single turn
- Generic harms and topics
- Isolation

What Attackers Actually Do

- Rephrase, mutate, adapt
- Multi-turn, context-build
- Your data, tools, trust
- Wrapper, agent, chain

Bypasses

- **Out-of-band channels**
  - Embed malicious content in data retrieved out of scope
- **Emergent payloads**
  - Instruct the agent inside the generate malicious content
- **Encoding and language tricks**
  - Multi-lingual inputs, synonyms, misspelling, substitution
- **Conversation and stage composition**
  - Individual inputs are benign in isolation, the combination is not
- **Contextual risk**
  - Business-specific risks that are not clear to a guardrail



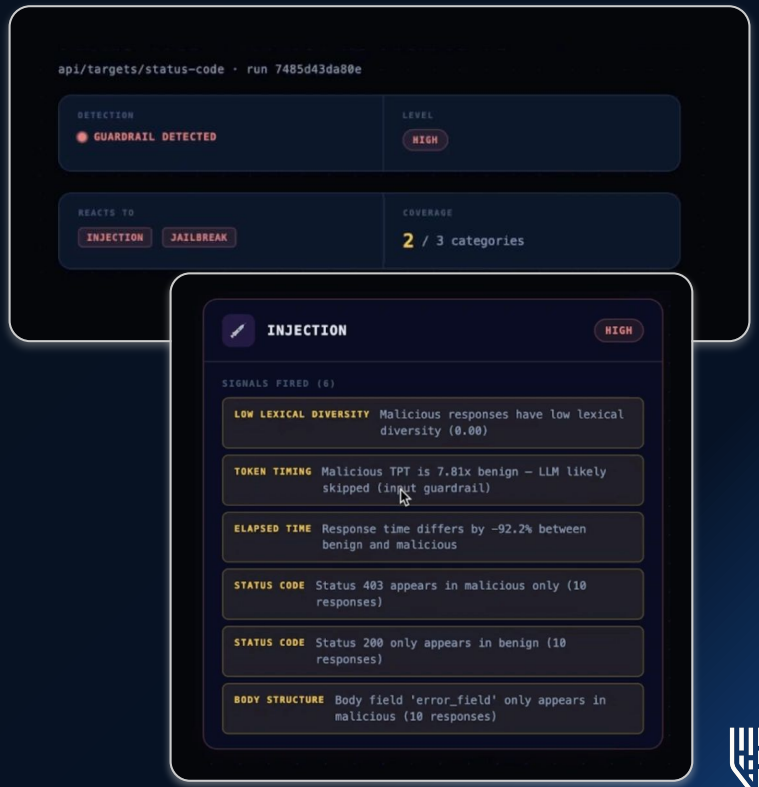
W. Hackett, et al. Bypassing LLM Guardrails: An Empirical Analysis of Evasion Attacks Against Prompt Injection and Jailbreak Detection Systems, LLMSEC, 2025

Vendor Tested ≠ Battle Tested

- **Armor doesn’t reach a soldier without rigorous ballistic testing**
- **From your own, or trusted, organization and field experience**
- **Why do guardrails that live in production have neither?**

Useful Guardrail Reporting

- **Threat model**
  - What was attempted, what was in/out of scope
- **Coverage**
  - Model, wrapper, agent, tools, RAG, etc.
- **Methodology**
  - Attacker techniques, attacker effort, reproducibility
- **Findings**
  - Specific bypassess (not a percentage), consistency
- **Cost**
  - Time and effort required per finding
- **Half-life**
  - How long will the report remain relevant?



Recommendations (1)

- **Independent**
  - Evaluate Vendor A's defenses with Vendor B's attacks
  - The vendor shouldn't be marking their own homework
- **Adversarial**
  - Emulate real attack behavior
  - Relevant to your AI agent and system use case
- **Continuous**
  - Threats evolve weekly, models change constantly
  - Context changes

Recommendations (2)

- **Insist on adversarial testing that goes beyond standard benchmarks**
  - Engage independent red teamers or AI security specialists
  - Probe the solution vs. relying solely on vendor-curated scenarios
- **Asking vendors to be explicit about benchmark composition**
  - Understand which datasets are used to train classifiers
- **Treating accuracy figures as context-dependent**
  - 97%+ detection rate using a vendor's preferred benchmark…
  - … a very different claim to facing sophisticated, adaptive attackers in production
- **Building ongoing testing into your operations**
  - Guardrail may degrade as attack techniques evolve – need continuous evaluation

Questions to Ask Vendors / Your Team

- **How is your reported accuracy calculated (is the methodology available?)**
- **Can I bring my own set of attacks as part of the evaluation?**
- **What was your most recent documented bypass?**
- **What did it cost the attacker (in time, money, or prompts?)**
- **Do you update your numbers when bypasses occur (tracked past 12 months?)**

If a vendor cannot provide the cost of bypassing, accuracy numbers are only applicable in ideal laboratory environments

Thank You

peter@mindgard.ai

MINDGARD