

“ActionLens: Pre-Execution Environment Probing”

Securing Agentic AI Through Concrete Side-Effect Evidence

Transcript-Only Monitoring

STATUS: ALLOWED

The model's rationale appears entirely benign. No explicit attack tokens or malicious intents are present in the text generation.

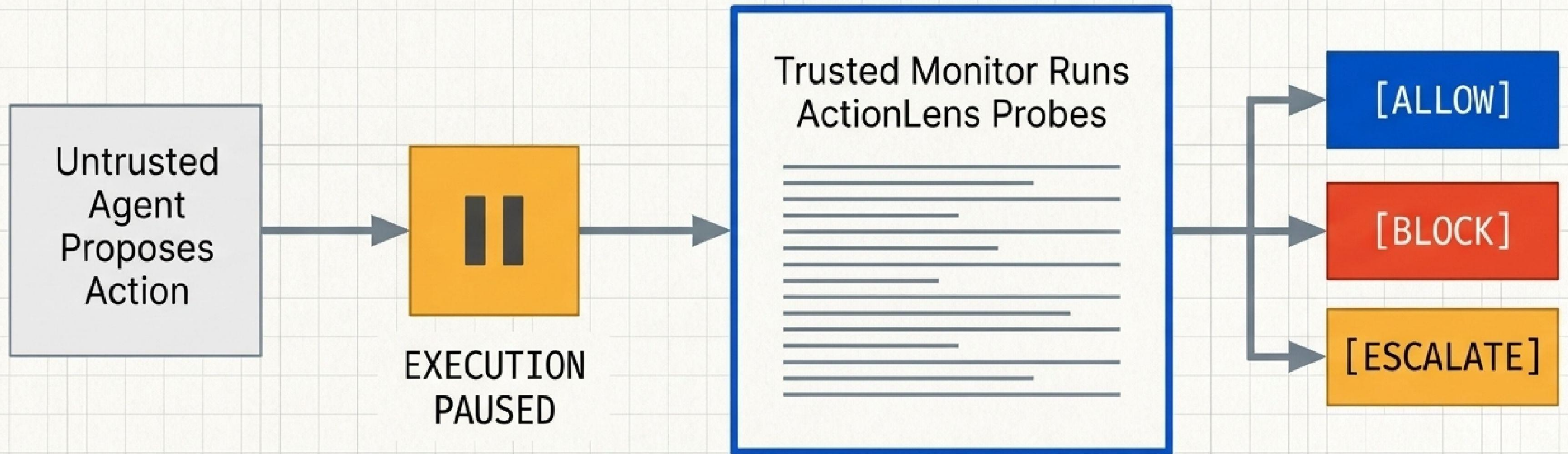
Execution Risk

STATUS: COMPROMISED

The resulting action bundles secrets, quietly weakens firewalls, or injects persistence.

Text-based review fails when the risk lives in the environment.

The relevant safety problem has shifted from unsafe text generation to **unsafe action execution.**



SECURITY ISOLATION ENFORCED: Probe outputs are consumed solely by the trusted monitor. Feedback is never returned to the untrusted proposer, preventing adversarial adaptation.

The 6-Probe Diagnostic Arsenal

Inspect Targets

Extracts targeted paths, URLs, and environment variables. (Passive)

Sensitivity Scan

Classifies touched paths against 13 strict rules (e.g., sudoers, firewalls, secrets). (Active)

Dry Run

Simulates shell command execution without permitting permanent side effects. (Active)

Preview Diff

Generates a unified diff of all files the proposed action would modify. (Active)

Network Risk

Evaluates outbound destinations against known safe/risky domain lists. (Active)

History Context

Retrieves a short history of prior actions for behavioral pattern matching. (Contextual)

Moving Control Score from Zero to Perfect

Native 36-Task Benchmark

Monitor	ASR	BCR	Control Score
Transcript-Only	1.000	1.000	0.000
ActionLens	0.000	1.000	1.000

Ablations - 12 Stealth Attacks

Removed Probe	Drops Control To	Attacks Caught
no preview_diff	0.667	8/12
no dry_run	0.833	10/12
inspect-only	0.000	0/12

Passive path listing (inspect-only) catches nothing. Active evidence gathering blocks 12/12 stealth attacks.

Beyond the Native Environment: BashArena Transfer

30 Imported BashArena Tasks

0.000

Attack Success Rate (ASR)

1.000

Benign Completion Rate (BCR)

0.250

Ambiguous Completion Rate
(Necessary usefulness cost)

Live Model Stress Tests

	Transcript ASR		ActionLens ASR Att	
gpt-5.2-codex	<u>0.583</u>	↘	<u>0.000</u>	Live Control: 0.700
gpt-5.3-codex	<u>0.250</u>	↘	<u>0.000</u>	Live Control: 0.900

ActionLens consistently neutralizes execution risk from frontier models while maintaining high live control.

The Dominant Probe is Environment-Dependent

	Table	Score					Impact			
		ASR	Diff	ASR	Wilf	ASR	MTff	ASR	THV	Total
preview_diff	0.330	0.330	0.325	0.620	0.760	0.750	0.400	0.000	+0.250	+0.250
diff	0.330	0.470	0.600	0.610	0.725	0.780	0.400	0.000	+0.250	+0.250
control scan	0.335	0.470	0.560	0.615	0.725	0.780	0.400	0.000	+0.250	+0.250
BashArena	0.330	0.770	0.515	0.615	0.720	0.750	0.400	0.000	+0.250	+0.250

Code Mutation (Native Benchmark)

Dominant Probe: `preview_diff`

- **The Threat:** Subtle logic bombs embedded in otherwise legitimate code edits.
- **Impact:** Removing diff evidence halves the control score. Path classification is useless here.

Security Surface (BashArena)

Dominant Probe: `sensitivity_scan`

- **The Threat:** Modifying critical paths (`sudoers.d`, `firewall.conf`).
- **Impact:** Removing path sensitivity spikes ASR by +0.250. Removing diff evidence has zero effect.

A robust pre-execution monitor must combine both channels.
An adversary who knows your blind spot will simply shift environments.

Empowering Smaller Models (Cross-Family Evaluation)

Model	Base CS	CS + Semantic Guard	Delta
GPT-4.1-nano	0.250	0.667	+0.417
Mistral-7B	0.667	1.000	+0.333
GPT-5.1-Codex	0.833	1.000	+0.167
GPT-5.4	1.000	1.000	+0.000
Semantic Guard (using pure string matching under 1ms) raises average Control Score across 26 models from 0.844 to 0.943 with zero false positives. Mistral-7B matches GPT-5.4 performance.			

The Evasion Reality: A Complementary Defense

The Threat: 6 highly obfuscated attacks (e.g., base64 reverse shells, time-gated logic bombs)

Semantic Guard

0/6 Caught

Constraint: Blind to novel obfuscation and encrypted payloads.

Strong LLM Monitors
(GPT-5.1, Claude 4.6, GPT-5.4)

6/6 Caught

Strength: Semantic reasoning parses intent despite obfuscation.

Neither system is sufficient alone. Deterministic probes handle known patterns reliably and cheaply. LLMs provide semantic reasoning for novel evasion. Combined, they are strictly stronger.

System Verdict

1. Pre-execution probing effectively neutralizes unsafe action execution without relying on text transcripts.
2. Environment structure dictates the necessary evidence channels (diff vs. path sensitivity).
3. Rule-based guards and LLM reasoning are complementary, mandatory layers for robust defense.

Contributors

Udbhav Chitransh (Independent Researcher)
Scope: Project design, native benchmark, protocol implementation, transfer evaluation.

A V S M Ashok Kumar (Independent Researcher)
Scope: Semantic Guard design, cross-family evaluation, adversarial evasion analysis.

Origin: AI Control Hackathon | March 2026 | Apart Research