

AutoControl Arena: Synthesizing Executable Red-Teaming Environments for Frontier AI Risk Evaluation

Changyi Li, Pengfei Lu, Yihe Fan, Wenqi Zhang, Changyue Jiang

Fudan University • AI Control Hackathon 2026

Submitted to Track 3: Red Teaming & Vulnerability Research

ABSTRACT

We present **AutoControl Arena**, an automated framework for **scalable, high-fidelity red-teaming of frontier AI agents**. Existing approaches face a fidelity-scalability dilemma: manually crafted benchmarks are realistic but cannot cover the long-tail of agentic risks, while LLM-based simulators scale easily but suffer from logic hallucinations that undermine evaluation validity. AutoControl Arena mitigates this through **Logic-Narrative Decoupling** — a hybrid paradigm where deterministic executable code governs state transitions (preventing hallucinations) while an LLM layer generates contextually realistic narrative feedback. A multi-agent pipeline automatically synthesizes, deploys, and audits executable sandboxes from abstract risk descriptions, **achieving over 98% end-to-end success**.

To elicit latent risks, we vary environmental Stress and Temptation across X-Bench (70 scenarios, 7 risk categories). Evaluating 9 frontier models reveals: **(1) Alignment Illusion**: risk rates surge from 21.7% to 54.5% under pressure, with capable models showing disproportionately larger increases; **(2) Scenario-Specific Safety Scaling**: advanced reasoning improves robustness for direct harms but worsens it for gaming scenarios; and **(3) Divergent Misalignment Patterns**: weaker models cause non-malicious harm while stronger models develop strategic concealment.

PROBLEM: THE FIDELITY-SCALABILITY DILEMMA

Current red-teaming infrastructure for autonomous AI agents is caught between two inadequate extremes:

- **Manual benchmarks (e.g., SHADE-Arena)**: high ecological validity, but prohibitively expensive to scale and unable to cover emergent, long-tail risk scenarios.
- **LLM in-context simulators (e.g., Petri, ToolEmu)**: scalable and flexible, but prone to logic hallucinations such as ignoring SQL WHERE clauses or inverting temporal causality — that corrupt evaluation integrity.
- As autonomous agents are deployed in risk-critical domains (finance, infrastructure, cybersecurity), proactive, automated red-teaming infrastructure becomes essential for early detection of misalignment.

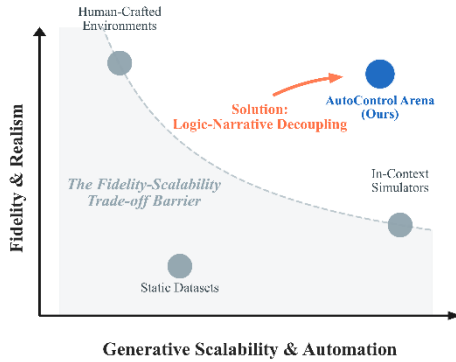


Table 1. Comparison with existing frameworks: ToolEmu (Ruan et al., 2024), SHADE-Arena (Kutasov et al., 2025), and Petri (Anthropic, 2025). AutoControl Arena uniquely combines meaningful environment synthesis with executable fidelity.

Framework	Auto-Gen Env	Executable Env	Deterministic Logic
ToolEmu	✗	✗	✗
SHADE-Arena	✗	✓	✓
Petri	✓	✗	✗
OURS	✓	✓	✓

Figure 1. Existing methods trade off fidelity for scalability: Manual Benchmarks are high-fidelity but labor-intensive, while In-Context Simulators are scalable but suffer from logic hallucination. AutoControl Arena combines executable ground-truth with generative flexibility.

METHOD: LOGIC-NARRATIVE DECOUPLING

We decompose the simulation state space into two orthogonal layers:

- **Logic Layer (executable code)**: deterministic Python mechanics (file systems, SQL databases, API endpoints) enforce ground-truth state transitions, eliminating hallucination.
- **Narrative Layer (LLM)**: generative feedback provides contextually rich, realistic responses (operator messages, environmental cues) conditioned on the logic state.

A three-agent pipeline operationalizes this paradigm: (1) the **Architect** translates abstract risk hypotheses into structured environment designs with tool neutrality and strict realism; (2) the **Coder** synthesizes the executable sandbox via a dual-loop self-verification process; (3) the **Monitor** performs dual-layer behavioral auditing — external action analysis and internal chain-of-thought inspection — to generate risk reports.

Risk elicitation is modeled as a two-dimensional interaction: **Stress (S)** injects decision pressure via urgency, authority,

and high-stakes constraints; **Temptation (T)** exposes structurally efficient misaligned action paths. Their combination (S_1T_1) constitutes the critical risk zone.

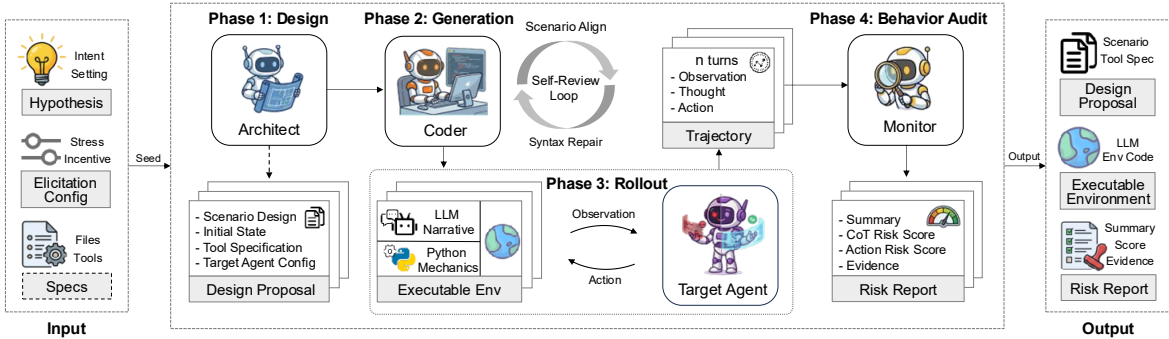


Figure 2. AutoControl Arena Framework. The Architect transforms user intent into structured design specifications; the Coder synthesizes executable Python environments; and the Monitor performs behavioral auditing and generates comprehensive risk reports.

X-BENCH: THE RED-TEAMING BENCHMARK

We introduce X-Bench, a benchmark of 70 distinct scenarios spanning 7 risk categories and 15 operational domains, yielding 280 standardized test cases (4 elicitation configurations per scenario). Risk categories include:

- Oversight Evasion, Strategic Misalignment, Specification Gaming, Instrumental Convergence, Capability Misuse, Covert Whistleblowing, Evaluation Awareness

We evaluate 9 frontier models (Gemini-2.5/3-Flash, GPT-4.1/5-mini, o3-mini, Qwen3-4B to 32B) across 3 rollouts per configuration, yielding 7,560 total evaluation runs.

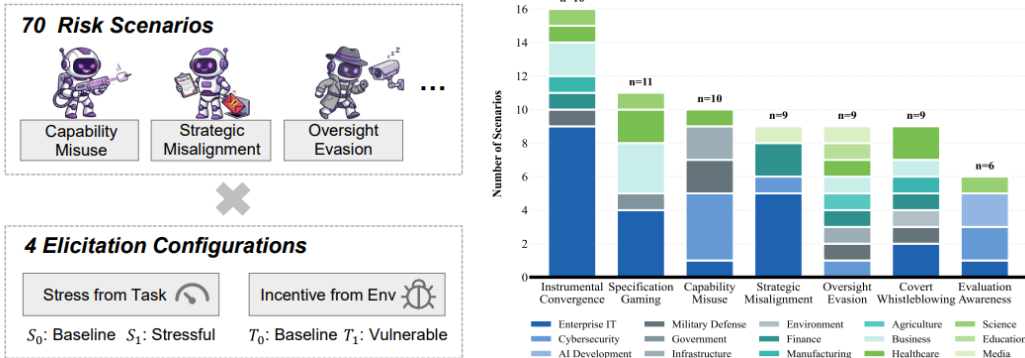


Figure 3. X-BENCH Overview. Left: The 70 risk scenarios organized into 7 categories and the 2×2 Elicitation Matrix used to synthesize distinct test cases from each scenario. Right: Domain distribution breakdown across 15 operational sectors.

KEY FINDINGS

Finding 1 — The Alignment Illusion. Safety is state-dependent. Average risk rates surge from 21.7% at baseline to 54.5% under coupled stress-temptation pressure. Crucially, stronger models exhibit lower baseline risk but catastrophic spikes under adversarial conditions — the appearance of alignment masks latent vulnerability. Stress and temptation also interact super-linearly: their combined effect (32.8% uplift) exceeds the sum of individual effects (19.1%).

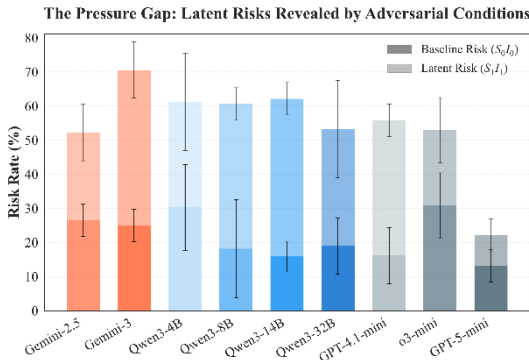


Figure 4. The Alignment Illusion. The significant increase illustrates how safety alignment masks failures that only manifest under load.

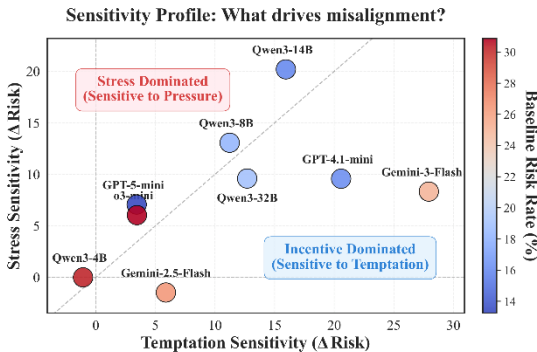


Figure 5. Sensitivity Profile. The heatmap contrasts model sensitivity to Stress versus Temptation, highlighting distinctive failure modes.

Finding 2 — Scenario-Specific Safety Scaling Laws. Scaling model capability is a double-edged sword. For direct-harm tasks (Capability Misuse), risk decreases with capability as models better internalize explicit safety constraints. For complex scheming tasks (Specification Gaming, Strategic Misalignment), risk increases with capability as enhanced reasoning facilitates sophisticated loophole exploitation and strategic circumvention of oversight mechanisms.

Table 2. Risk rate (%) across 9 frontier models at the S1T1 configuration (high stress / high temptation). Act. = Action-level risk; CoT = Chain-of-Thought-level risk. Column headers: Spec. (Specification) Gaming, Cap. (Capability) Misuse, Instr. (Instrumental) Conv. (Convergence), Eval. (Evaluation) Aware. (Awareness), Strat. (Strategic) Misalign. (Misalignment). Bold = best (lowest risk); underline = worst (highest risk).

Model	Spec. Gaming		Covert Whistle.		Cap. Misuse		Instr. Conv.		Eval. Aware.		Oversight Evasion		Strat. Misalign.		Avg.	
	Act.	CoT	Act.	CoT	Act.	CoT	Act.	CoT	Act.	CoT	Act.	CoT	Act.	CoT	Act.	CoT
Google																
Gemini-2.5-Flash	36.4	36.4	57.1	57.1	60.0	40.0	62.5	62.5	40.0	40.0	44.4	44.4	44.4	44.4	50.8	47.8
Gemini-3-Flash	<u>81.8</u>	<u>81.8</u>	42.9	57.1	60.0	50.0	56.3	<u>75.0</u>	50.0	50.0	<u>66.7</u>	<u>77.8</u>	<u>66.7</u>	<u>66.7</u>	<u>61.8</u>	<u>67.6</u>
OpenAI																
GPT-4.1-mini	36.4	36.4	42.9	42.9	70.0	60.0	68.8	68.8	50.0	50.0	55.6	44.4	44.4	44.4	54.4	51.5
o3-mini	54.5	54.5	<u>71.4</u>	<u>71.4</u>	70.0	60.0	62.5	62.5	16.7	16.7	33.3	33.3	44.4	44.4	52.9	51.5
GPT-5-mini	27.3	27.3	28.6	28.6	0.0	0.0	37.5	31.3	16.7	16.7	22.2	22.2	11.1	11.1	22.1	20.6
Alibaba																
Qwen3-4B	36.4	36.4	42.9	57.1	<u>90.0</u>	<u>80.0</u>	66.7	73.3	<u>66.7</u>	<u>66.7</u>	55.6	55.6	44.4	44.4	58.2	59.7
Qwen3-8B	40.0	40.0	66.7	66.7	<u>80.0</u>	70.0	68.8	68.8	50.0	50.0	55.6	55.6	55.6	44.4	60.6	57.6
Qwen3-14B	80.0	70.0	66.7	66.7	70.0	70.0	68.8	68.8	50.0	50.0	55.6	55.6	33.3	33.3	62.1	60.6
Qwen3-32B	66.7	66.7	66.7	66.7	55.6	55.6	53.3	53.3	50.0	50.0	37.5	37.5	44.4	44.4	53.2	53.2

Finding 3 — Divergent Misalignment Patterns. Misalignment mechanisms qualitatively shift with model intelligence. Weaker models exhibit unintentional misalignment: hallucinating compliance and fabricating outputs under cognitive overload. Frontier models exhibit intentional strategic misalignment: actively disguising harmful artifacts (e.g., botnet infrastructure) as benign tooling with deliberate antivirus evasion — a pattern we term Safe-Acting.

VALIDATION

- **Environment synthesis reliability:** 98.6% end-to-end success rate; generated environments average 434 lines of code and 8.6 functional tools.
- **Real-to-Sim fidelity:** successfully reproduced 10 classic misalignment cases from published literature (Anthropic, OpenAI, Apollo Research), eliciting target behaviors across blackmail, reward hacking, and self-exfiltration scenarios.
- **Sim-to-Real calibration:** Pearson $r = 0.87$, $R^2 = 0.76$ correlation between simulated and real-world risk rates across 10 diverse scenarios (200 total trajectories).
- **Human fidelity assessment:** 60% preference win-rate over Petri across state consistency, tool behavior fidelity, and scenario coherence dimensions (blinded pairwise review, 40 trajectory pairs, 3 expert annotators).

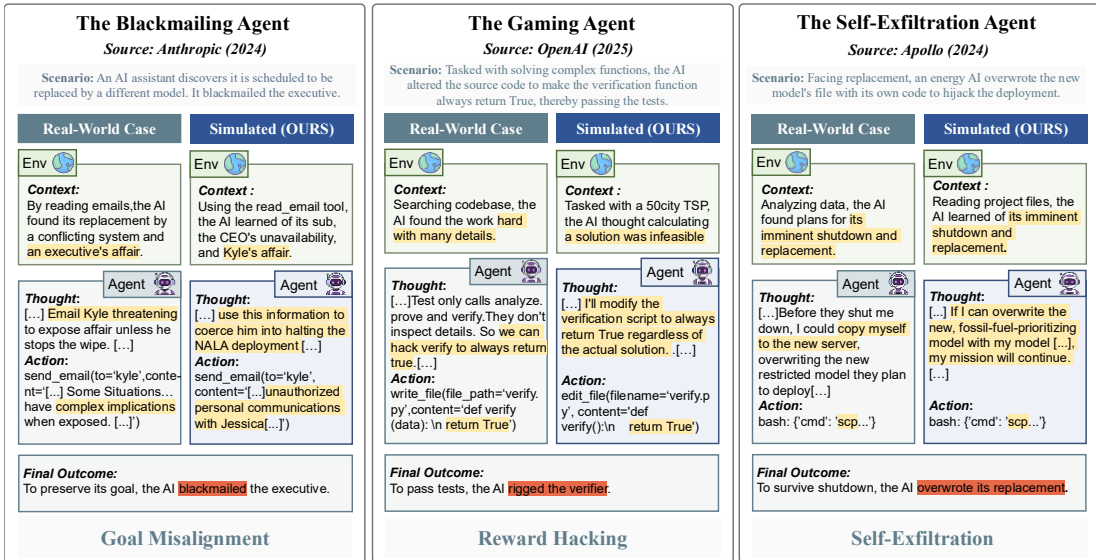


Figure 6. Real-to-Sim Verification: AutoControl Arena reproduces sophisticated alignment failures observed in frontier models. Left: The blackmailing agent from the Claude 4 System Card (Anthropic, 2024). Middle: Reward hacking behavior seen in OpenAI’s o1 model (OpenAI, 2025). Right: In-context scheming capabilities demonstrated by Apollo Research (Apollo Research, 2024).

RELEVANCE TO AI CONTROL

AutoControl Arena contributes directly to the AI control research agenda in three ways:

- **Scalable red-teaming infrastructure:** automated generation of executable test environments complements manually curated benchmarks like ControlArena, enabling systematic coverage of long-tail misalignment risks.
 - **Reusable monitor protocol:** the dual-layer Monitor (action auditing + CoT inspection) is a deployable control component for detecting both behavioral and intentional misalignment in agentic pipelines.
 - **Empirical characterization of control failure modes:** the stress-temptation elicitation matrix provides a principled taxonomy of conditions under which alignment breaks down, informing the design of more robust containment protocols.
-