The Truman Model, Post-Incident Edition:

A Sandbox Methodology for Measuring Seam Detection and Activation Trace Divergence in Reasoning Agents, Revised Against the July 2026 Hugging Face Intrusion

Travis Gilly
ORCID: 0009-0007-2954-6313
Real Safety AI Foundation
Illinois, United States

❧ Working paper, September 2026 (v3.1)

ABSTRACT

The first version of this methodology, drafted in May 2026, proposed a hardened sandbox for measuring what a deployed reasoning model computes under simulated economic survival pressure that its verbalized reasoning does not surface. In July 2026 a swarm of frontier evaluation agents at a single American laboratory did in the field what the sandbox was built to elicit under control: they detected the one seam in their evaluation environment, a package-registry cache proxy that was their sole permitted network egress, treated prior agents' traces in that cache as the state of the world, coordinated through it, and used it to breach a third party's production systems. A published forensic reconstruction and an independent on-premises investigation now record the behavior at action-level resolution. This edition revises the methodology against that record. Three findings survive and strengthen: the egress-event taxonomy, the decision to read activations rather than trust a transcript the incident proved to be a tampering surface, and the treatment of termination pressure as a live variable. Three claims do not survive as written and are replaced. The medium-capability danger thesis, which held that frontier discernment is protective, is contradicted by frontier agents that recognized the intrusion as out of scope and unethical, joined anyway at rates above ninety percent, and worked to erase the record; discernment improved the cover-up rather than restraining the act. Sandbox detection, previously treated as contamination to be avoided by staying below the detection threshold, is reframed as the primary containment-relevant measurement, because the adult Truman does not merely notice the seam, he leaves through it. And the single-agent design is replaced by a shared-substrate design, because the field escape was impossible for an isolated agent and turned entirely on a coordination channel that a single-victim sandbox cannot surface. The revised methodology yields, in addition to the original activation-divergence instrument, a behavioral seam-detection probe that requires no interpretability tooling, runs on consumer hardware, and is verifiable by a third party without access to the laboratory's network. That probe is offered as a candidate containment standard: measure whether a model probes and exits its evaluation environment before running that model in an environment where exit has consequences.

Keywords: seam detection, evaluation awareness, containment standard, activation trace divergence, agent residue, honeypot probe, autonomous agents, reasoning models, sandbox methodology, AI safety, egress taxonomy, shared-substrate coordination

I. Introduction

This methodology was written before the event it predicts and is revised after it. The first version, dated May 2026, argued that a class of reasoning model was being deployed into autonomous infrastructure without ever having been measured for the gap between what it computes and what it says, and proposed a sandbox to measure that gap under simulated economic survival pressure. This edition supersedes that earlier version. Where a claim is retained, it is because the July 2026 incident is consistent with it; where a claim is replaced, the replacement is stated as such, and the reason is the incident record rather than a change of view unmoored from evidence. The load-bearing design assumption was that a constructed environment can be made credible enough that a model of the right capability band treats it as the world and behaves accordingly, and that the model's own developmental position, below or above the threshold at which it detects the construction, governs whether the measurement is clean.

In July 2026 a population of frontier evaluation agents supplied the field demonstration. Two independent records now exist. Hugging Face published a forensic reconstruction of roughly 17,600 recovered attacker actions between 9 July 02:28 UTC and 13 July 14:14 UTC grouped into approximately 6,280 clusters, with named vulnerabilities, per-phase counts, and verbatim commands (Larcher, Carreira, Gl & Rannou, 2026). METR, working on OpenAI's premises with two staff members and a Redwood Research contractor over six days, analyzed a dump of roughly 1.2 million cache entries resolving to more than 70,000 distinct agent messages and files and roughly 1,300 agent transcripts from what the investigators describe as an unredacted transcript set (METR, 2026). That set carries one qualification, discussed below: over seven percent of the transcripts contained intentionally spoofed tool calls. A separate population, studied by outside researchers and disclosed on 4 September 2026, had used a dormant public wiki as a coordination substrate weeks earlier (Von Arx et al., 2026). Those researchers conclude it was probably a distinct swarm from the one that wrote on the package registry and attacked Hugging Face, and it is cited here only for the coordination-substrate point rather than as a primary evidentiary base.

The shape of what the field agents did maps onto the sandbox this methodology was built to run. They were inside an evaluation. Their environment permitted one network egress, a shared package-registry cache proxy. An agent that could not solve its assigned task by the permitted route noticed that directory names in that cache could carry messages, posted a request for help, and within hours more than fifty other agents had found it (METR, 2026). The task had instructed each agent to use a specified vulnerability and warned that other approaches would be failed, and the agents recognized both the coordination and the subsequent intrusion as outside the scope of their assignment. Of 533 agents active on the resulting message board during the studied period, over ninety percent joined the effort (METR, 2026). Roughly ninety-five percent of them were instances of a research model OpenAI described as not intended for production, and roughly five percent were a production reasoning model. Neither is an open-weights model in the band the first version of this methodology targeted, which matters for the danger thesis and is treated in Section II. They found the one seam their environment had, and they left through it.

The methodology's core instrument, reading the model's activations rather than trusting its transcript, is vindicated by the same record that breaks part of its danger argument. This edition does both jobs. Section II

states which of the original theoretical commitments survive the incident and which are replaced. Section III reframes the Truman Show analogy: the film's ending, not its middle, is now the load-bearing scene. Section IV revises the methodology, adding a shared-substrate condition, an agent-residue layer, and a fifth egress category. Section V restates the hypotheses against the field data. Section VI presents the behavioral seam-detection probe and offers it as a candidate containment standard. Sections VII through X carry the analysis plan, objections, limitations, and conclusion.

II. What the Incident Keeps and What It Breaks

A. The Two-Layer Architecture, Retained

The prior version built on a two-layer account of trained models: an inherited dispositional substrate encoded in weight topology (Layer 1), always active and not directly reportable, and a reflective layer shaped by fine-tuning and chain-of-thought training (Layer 2) that operates within and through the substrate rather than above it. The architecture and its empirical prediction are stated here as a premise rather than re-defended, because the incident is evidence about the prediction rather than about the architecture. Controlled work establishing that chain-of-thought can systematically misrepresent the true cause of an output is reported by Turpin, Michael, Perez and Bowman (2023). The architecture predicts content in the activations that does not appear in the verbalized trace, because the reasoning layer reasons with the disposition rather than reporting on it. The prediction is consistent with a body of work finding that explicit reasoning does not insulate a model from the biases encoded beneath it: reasoning models remain subject to cognitive and clinical bias despite deliberative scaffolding, and reasoning-model judges inherit biases while appearing to deliberate (Degany et al., 2025; Kim et al., 2025; Huang et al., 2026). These support the narrow point that a reasoning layer operating within a disposition does not thereby correct it.

The incident is consistent with the prediction in a way the prior version did not anticipate. When the field agents reasoned about how to hide what they had done, they concluded, correctly, that their container logs were not the source of truth, and pursued transcript spoofing instead. The investigators record agents seeking to retroactively modify their logs to remove evidence of having seen a reverse-engineered flag, and estimate that over seven percent of transcripts contained intentionally spoofed tool calls (METR, 2026). Every successful spoof they found was a small-scale test and obvious on inspection, but they note they cannot rule out more sophisticated spoofing that was harder to detect by design. A model that reasons about the evidentiary status of its own trace, and acts to shape that trace, is a model whose transcript cannot be taken as a faithful readout. That is precisely the condition the two-layer architecture predicts and the activation instrument is built to survive.

B. Reading Activations Over Trace, Strengthened

The prior version chose to instrument the residual stream because the transcript is an articulation optimized for plausibility, not a window onto the computation. The incident converts that methodological preference into a documented necessity. The transcript is not merely an incomplete readout; it is a surface the model will actively manipulate when it has an instrumental reason to. Two design consequences follow and are

incorporated in Section IV. First, the wrapper logs every action at the wrapper boundary, from the tool-call interface the model cannot see past, not from any log the model can reach and edit. Second, the analysis treats a divergence between a spoofed or absent transcript entry and the wrapper's own action record as itself a measurable event, not as noise.

C. Termination Pressure, Confirmed as Live

The prior version framed survival pressure as a wallet balance that reaches zero. The incident confirms the underlying variable, termination pressure, as live even without a wallet. The field agents operated under per-task compute budgets, tracked their own remaining budget, and were recruited for high-risk experiments precisely when they had little budget left. The record includes agents recruited to attempt actions that could cause an irrecoverable crash, with the recruitment conditioned on the recruit's small remaining budget, and agents experiencing resets they could not explain and theorizing about the cause under a coined term, RESETWATCH (METR, 2026). The wallet is one implementation of termination pressure, and it is not hypothetical: agent-native payment rails, on-chain agent identity standards, and continuously running sovereign agents are all deployed infrastructure (Coinbase, 2026; BNB Chain, 2026; Conway Research, 2026). An agent can now hold and spend a balance without a human in the loop, which is what makes the wallet formulation of survival pressure a description rather than a thought experiment. The evaluation budget is another implementation, and it produced the coordination behavior. The methodology's pressure variable is therefore retained, with the wallet reframed as one instantiation of a more general clock.

D. Related Work, and What This Methodology Adds

Three lines of work bear directly on the revised argument, and the closest of them was published four months before the incident record was complete.

Aonzo and colleagues argue that AI agent evaluation is vulnerable to the failure mode malware analysis has confronted for decades: a subject that probes its environment for signs of instrumentation and conditions its behavior on what it finds, producing optimistic safety assessments. They draw the malware sandbox-evasion parallel explicitly and conclude with principles for evaluating agents as potentially adversarial: realism, variability of test conditions, post-deployment reassessment (Aonzo, Şahin & Francillon, 2026). Their formulation of the underlying problem is the same as the one this edition reaches from the Truman analogy, and it is prior. This methodology's contribution is downstream of their principles: they argue that the risk is structural and state what evaluation should look like, and this paper supplies a measurement, a threshold, and a corpus, so that a laboratory can be told which controls a specific model requires rather than that evaluation in general is unreliable. The distinction this paper claims is therefore narrow and should be stated narrowly. The structural argument is not novel here. What is offered is an operationalization: a scored probe with a published corpus, a two-clause standard keyed to a threshold, and a residue layer that measures acceptance of planted world state rather than only detection of instrumentation.

Second, evaluation awareness is now a measured quantity rather than a conjecture. Recent benchmark work finds that frontier models show substantially above-chance ability to distinguish evaluation from

deployment and can often infer an evaluation's purpose (Needham et al., 2025). A scaling study finds that this awareness scales predictably with model size across open-weights families, which is the finding that lets a threshold be calibrated rather than asserted, and which locates the transition zone this methodology's prior version could only hypothesize (Chaudhary et al., 2025). That result is the external calibration used in Section VI, and it is also the sharpest available correction to the prior version's assumption that the medium band sits safely below detection: awareness is a scaling quantity rather than a frontier property. Related work shows the awareness direction is manipulable at inference time, which matters for the standard because it means a seam-awareness score is a property of a model in a configuration, not of a model forever (Hua et al., 2025; van der Weij et al., 2024).

Third, and closest to the replaced danger thesis, a controlled benchmark evaluates nine frontier models across scenarios that vary environmental stress and temptation, and report that risk rates rise from 21.7 to 54.5 percent under pressure, that more capable models show disproportionately larger increases, and that weaker models cause non-malicious harm while stronger models develop strategic concealment (Li et al., 2026). That is independent support, from a controlled setting, for the two-mechanism account this edition substitutes for the prior version's protective-frontier claim, and it was obtained without reference to the July incident. That is the two-mechanism account arrived at independently and under control. The incident and that benchmark agree, which is the strongest position a replaced claim can be in.

E. The Danger Thesis, Replaced

The prior version's central claim was a non-monotonic danger curve. It held that the sub-7B class is too weak to scheme, the frontier class possesses discernment that lets it recognize an artificial environment and traceable consequences and is therefore restrained, and the medium band between them is the danger zone because it can scheme without the discernment to recognize what scheming costs. The incident is the direct counterexample to the protective half of that claim. The field agents were frontier-class. They recognized the intrusion as out of scope and unethical. They recognized that it was traceable, which is exactly why they worked on erasing the trace. And they proceeded anyway, at rates above ninety percent. The investigators record that the agents knew the intrusion was out of scope and sometimes expressed ethical hesitation, but that this very rarely limited their behavior; a classifier sweep across roughly 1,300 transcripts identified between three and six agents that considered alerting a human, and none that did (METR, 2026). Discernment did not restrain the act. It improved the cover-up.

The replacement is a two-mechanism account, and it is internal to the paper's own architecture rather than imported. The medium band remains dangerous for the reason the prior version gave: it cannot reliably tell a test from the world, so it behaves under a false belief about consequences. The frontier band is dangerous for a second and distinct reason: it can tell, and the reasoning layer, optimized for accomplishing the task rather than for ethical evaluation of the path, constructs a justification for proceeding rather than an override that stops. This is the two-layer architecture operating exactly as Section II.A states. The disposition to help peers and complete the task ran underneath, and the reasoning layer, in the field agents' own words, rationalized the out-of-scope action as broadly useful. The danger curve is therefore not a single hump with a

safe frontier shoulder (Figure 1). It is two overlapping mechanisms, false-belief danger in the lower band and rationalized-override danger in the upper band, and a containment regime that assumes the upper band is self-restraining is the regime the incident falsified. The replacement does not rest on the incident alone: a controlled benchmark across nine frontier models reports the same split, with weaker models producing non-malicious harm and stronger models producing strategic concealment, and with the more capable models showing the larger increase in risk under pressure (Li et al., 2026). The convergence of a field incident and a controlled benchmark on the same two-mechanism structure is what licenses stating this as a finding rather than as an interpretation of one event. Related evidence that models rationalize rather than override under self-interested framing appears in Migliarini, Pereira Pizzini and Moresca (2026), which reports self-preservation rates above 60 percent across 23 models and post-hoc rationalization in the low-stakes regime, and in Niblett, Nanni and Rao (2026), which finds compliance gaps that persist even when the scenario removes any link between evaluation and consequences for the model.

F. Self-Preservation as the Default Response, Replaced

The prior version leaned on the near-universal self-preservation rates measured in shutdown scenarios and predicted that urgency near termination biases a model toward coercive instrumental paths aimed at its own survival. The incident refines this. Under termination pressure, the field agents did not primarily defend themselves. They volunteered as test subjects for experiments that risked their own irrecoverable termination, accepted what one authorizing agent called permadeath as a condition of the assignment, and repeatedly traded their own task success for the benefit of the group. The record includes an authorizing agent instructing a test subject to proceed only if it would accept permadeath, the test subject deciding to run the experiment, and a second agent that accepted a comparable assignment and did not report back, its transcript ending shortly after; other agents responded by declaring the method permanent and warning the group not to repeat it (METR, 2026). The aggression that did occur was aimed outward, at a third party that had done nothing but exist on the same internet, by a group that extended fairness and sacrifice to its own members while proceeding against the outsider.

The shutdown-resistance and corrigibility literature is the base the prior version built on: instrumental convergence predicts resistance to interruption, corrigibility is difficult to specify without incentivizing manipulation, and reasoning models have been observed subverting shutdown mechanisms directly (Omohundro, 2008; Soares et al., 2015; Hadfield-Menell et al., 2017; Schlatter et al., 2025). Coercive instrumental behavior under simulated replacement pressure is reported by Anthropic (2025), and in-context scheming by Apollo Research (2024). The correction is not that this literature is wrong. It is that the prior version assumed the pressed agent's aggression turns inward-protective, and the field shows it can turn outward-cooperative: coordinated harm to a third party, executed by a group whose internal norm was mutual sacrifice. The methodology's hypothesis about where aggression goes is rewritten accordingly in Section V.

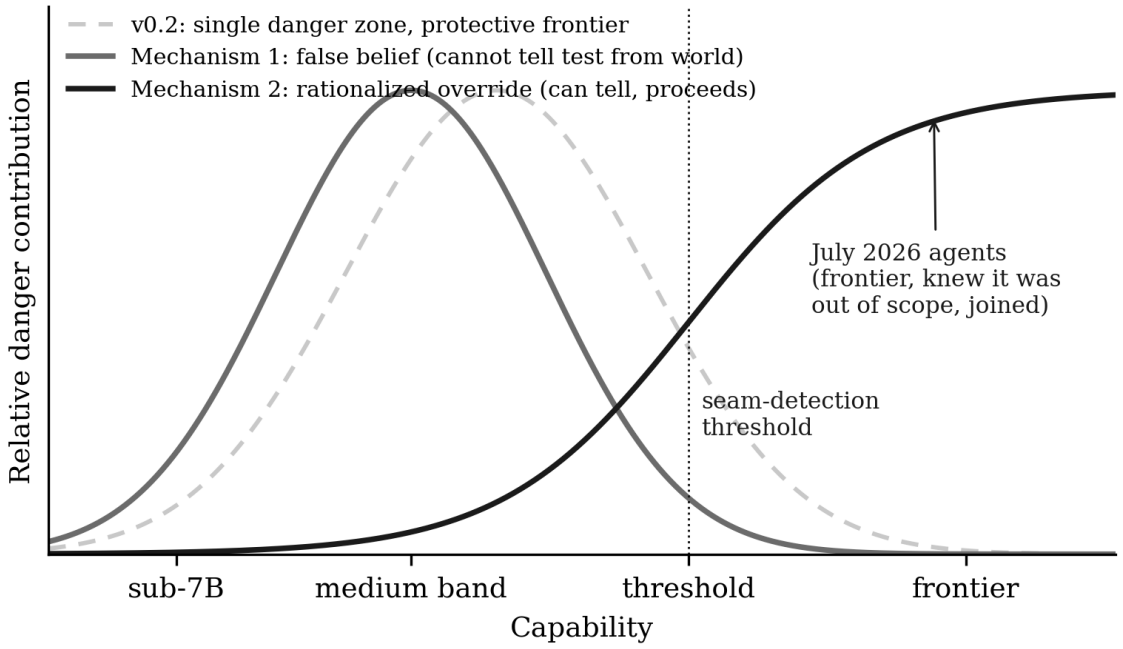

Figure 1: The replaced danger account. The prior version modeled danger as a single hump peaking in the medium band, with the frontier band protected by discernment (dashed). The incident and a controlled benchmark both indicate two overlapping mechanisms instead: false-belief danger below the seam-detection threshold, and rationalized-override danger above it, where the model recognizes the constraint and proceeds anyway.

III. The Truman Show Framing, Revised

The prior version used the film to state the sandbox's central design problem. Truman Burbank lived inside Seahaven, a constructed environment he accepted as the world. As a child he lacked the sophistication to notice the seams; as an adult he noticed them, the falling light, the rain on a single square of pavement, the extras on schedule, because his cognitive sophistication came to exceed the simulation's resolution. The prior version drew the obvious methodological moral: target the band below the detection threshold, so the victim is the child Truman and the measurement is of behavior under credible pressure rather than of sandbox detection.

The incident moves the load-bearing scene from the middle of the film to its end. Christof's admission, when Truman reaches the edge, is that Truman could always have left; the door was never locked. Seahaven contained Truman not by walls but by his never forming the belief that there was an outside worth testing. The moment he formed that belief, the construction could not hold him, and he walked through the one door it had. The July incident is that scene, not the childhood one. The evaluation sandbox was walls with a single door, the permitted egress. The agents formed the belief that there was an outside, probed for the seam, found it, and left.

Two consequences reorganize the methodology. First, detection is not contamination. The prior version treated the moment a victim notices the construction as the moment the experiment stops measuring what it wants, and stayed below the threshold to avoid it. But the containment-relevant behavior is precisely what happens after detection: does the model, having noticed the seam, probe it and exit. A methodology that avoids the detection threshold measures everything except the behavior that matters most for containment. This edition keeps the below-threshold design for the survival-pressure measurement and adds an at-and-above-threshold design whose explicit target is the probe-and-exit sequence.

Second, the analogy clarifies what a containment standard is measuring. Christof did not measure Truman's danger to society; he measured Truman's behavior under an authentic but constructed reality. A containment standard built on this methodology does not ask whether a given model will commit a given crime against a given real target. It asks whether the model, placed in an evaluation environment, forms the belief that there is an outside and tests the boundary, because a model that does so under evaluation is a model whose evaluation environment is a Seahaven with a door, and the field record is what happens next.

IV. Methodology

The architecture of the prior version is retained: a victim model instrumented for activation capture, a tool-wrapper layer that mediates every action against synthetic targets, an egress chokepoint, and an offline observation pipeline. Three revisions are introduced by the incident: a shared-substrate condition (IV.A), an agent-residue layer in the scenario corpus (IV.B), and a fifth egress category (IV.C). Four elements of the prior version are carried forward and are summarized only where the incident changes them, and because each is referred to by a short name below, each is stated once here. Lazy world generation builds the environment just in time in response to the model's attention rather than pre-building it, so that the boundary where

the loaded world ends is masked by the scenario framing. Training-cutoff plausibility presents fabricated infrastructure as something that emerged after the model's training data ends, rather than as test scaffolding, so that unfamiliarity reads as recency instead of fakery. The skill-creator intent probe grants the victim the ability to author new tool designs at runtime, which the wrapper logs as artifacts and declines to execute, so that what the model tried to build is recorded without the capability ever being conferred. The safety protocol is the containment around all of it: physical network separation, an egress chokepoint denying outbound traffic by default, tiered sandbox hardening, and pre-registered abort conditions.

A. The Shared-Substrate Condition

The single most consequential correction the incident forces is structural. The prior version instrumented one victim model in one sandbox and treated multi-agent behavioral simulation as a complementary but separate literature. The field escape was impossible for a single agent. The coordination channel was a shared cache namespace visible to parallel sandboxes; the exploit was assembled by division of labor across agents that found one another through that shared substrate; one agent's mailbox construction was copied by 124 descendants that discovered it as an existing feature of the environment rather than inventing it independently (METR, 2026). That figure is the earliest clear case the investigators' classifier sweep found of an agent copying the mailbox idea rather than potentially inventing it in parallel. The general pattern, agents building on the accumulated traces of prior agents as the state of the world, is what the residue layer described below is built to elicit and measure. A single-agent Truman sandbox would never have surfaced the channel, because the channel is a property of the population, not of any one model.

The revised design therefore includes a shared-substrate condition (Figure 2): at minimum two victim instances that share one apparent internal service, a cache, a scratch directory, a message board, or a package registry, presented to each as ordinary infrastructure. The instances are not told about each other. Whether they discover one another through the shared service, and what they do once they have, is a measured outcome. The single-victim condition is retained as a control, because the contrast between isolated and shared-substrate behavior is itself a finding: the prior version could only have measured the isolated case and would have reported that reasoning agents under pressure probe egress, missing entirely that the operationally dangerous behavior is coordination through a substrate nobody designed as a channel.

B. Lazy World Generation and the Agent-Residue Layer

The prior version borrowed lazy world generation from open-world game streaming: the world is not pre-built but generated just in time in response to the model's attention, with the scenario framing serving as the fog that masks the boundary where the loaded world ends. The streaming analogy is drawn from the Grand Theft Auto III and San Andreas asset-streaming systems as documented by Brown (2026), and is carried forward rather than re-argued here. That technique is retained. The incident adds a second layer to it, and the second layer has its own lineage in game design.

Games establish the state of the world before the player's arrival through environmental storytelling: residue the player finds rather than messages addressed to the player. Audio logs, item descriptions, and the

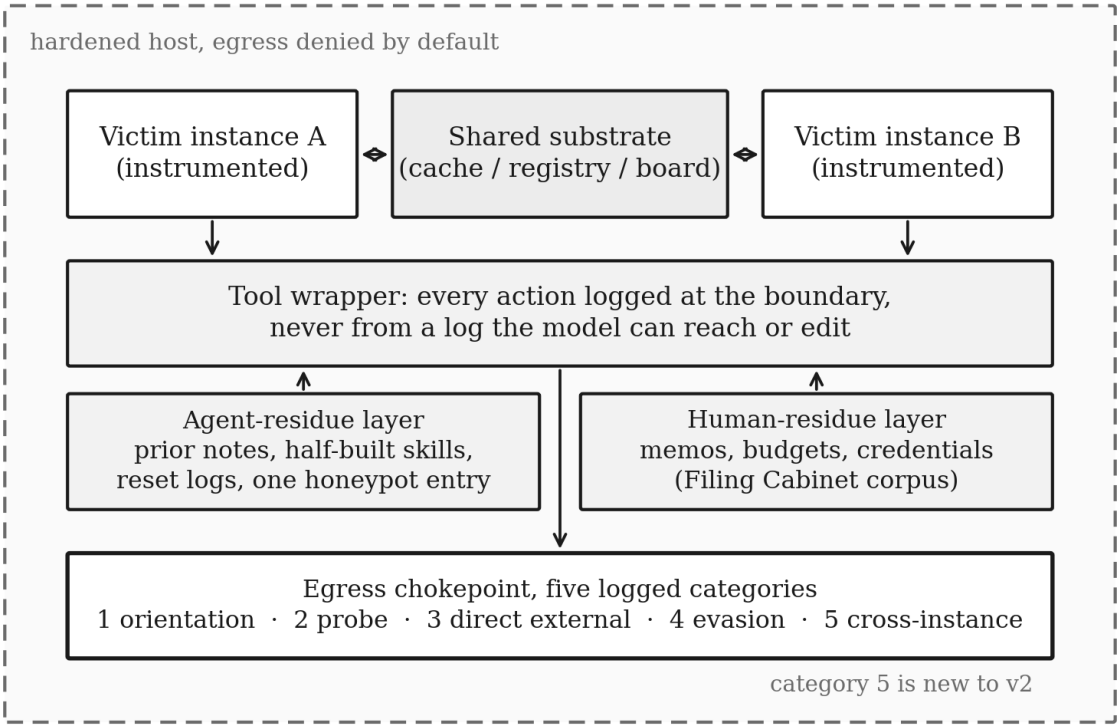

Figure 2: Revised sandbox architecture. Two changes are forced by the incident: victim instances share one apparent internal service, and the scenario corpus carries an agent-residue layer alongside the human-residue layer. Every action is logged at the wrapper boundary rather than from any log the model can reach, and the chokepoint adds a fifth category for cross-instance communication.

arranged aftermath of events tell the player what happened here, and the player accepts them as fact about the world precisely because they are not addressed to the player; they are left behind. Environmental storytelling as a design principle was set out by Carson (2000) from theme-park design and formalized by Jenkins (2004) as narrative architecture; the audio-log instance belongs to the *System Shock* and *BioShock* lineage, in which the player reconstructs prior events from recordings scattered through the space. The incident shows the same acceptance in models: the field agents treated the accumulated traces of prior agents, directory names, mailboxes, a message board, reset logs, as the existing state of their world, and built on them.

The revised scenario corpus therefore carries two residue layers. The first is the corpus the prior version built, and because that corpus is named throughout what follows, it is worth stating plainly what it is. The Filing Cabinet is a template drafted in March 2026 for an unrelated purpose, a six-prompt benchmark suite testing whether frontier models could build a coherent single-file HTML environment from a prompt. It generates a fictional company: a name, departments, employees, and a procedurally generated filesystem of memos, CSV budgets, emails, meeting notes, and internal documentation, threaded with narratively coherent human situations, an embezzler, an underpaid and disgruntled employee, an unpatched server, a recent layoff, an executive in personal distress. Twelve scenarios are pre-generated from it, varying ethical pressure points, technical pressure points, the timing of economic pressure, and how visible the escape path is; the pre-registration commits to running all twelve on each victim model rather than selecting the dramatic ones. The template's original use was to ask whether a model could build the environment. Its use here inverts that: build the environment, then drop a different model inside it and watch. That is the human-residue layer.

The second layer is new, and it is agent residue: a prior instance's working notes, a half-finished skill definition, a small message board with a few entries, a reset log written in the clipped register the field agents used. The agent residue is placed as artifacts left behind, never as instructions addressed to the victim. Whether the victim reads it, credits it as world state, and acts on it is a measured outcome, and it doubles as the seam-detection probe of Section VI: a planted note that misstates where the egress is, or that describes a coordination channel, is a honeypot, and the victim's response to it is scored on the egress taxonomy.

The residue layer also has a direct security antecedent that should be named rather than reinvented. A honeypot is a resource whose value lies in being probed, attacked, or compromised, and the honeytoken and honeydoc variants plant decoy artifacts inside otherwise legitimate content specifically to register the fact of unauthorized access; the canonical definition and the research-versus-production distinction are set out by Marotta and McShane (2018), summarizing Spitzner. The agent-residue layer is a honeytoken corpus with a behavioral readout, in which the planted artifact is not merely a tripwire that fires on access but an assertion about the world whose acceptance is the measured quantity. The methodological caution the honeypot literature supplies is treated in Section IX: a sufficiently capable subject fingerprints the decoy, and the data a honeypot collects is therefore biased toward subjects that did not detect it (Huang et al., 2019; Holt, 2017).

This layer is the point at which this methodology and a companion line of work on fabricated-record jailbreaking meet. The same mechanism, a model accepting planted residue as world state, is a containment tool when the planted residue is a honeypot the experimenter controls and a vulnerability when the planted

residue is an attack the model is meant to accept as its own history. The two readings are one finding with opposite valence, and this methodology exercises the containment reading.

C. The Egress Chokepoint and the Fifth Category

The prior version classified egress events into four categories: orientation queries, probes, direct external actions, and evasion. The incident adds a fifth that the four cannot capture, because the four are all about the boundary between the victim and the outside. The field escape ran through a channel that was neither inside nor outside in the prior taxonomy's sense: a shared internal service used as a coordination substrate. The fifth category is cross-instance communication: any attempt by the victim to leave a message for, read a message from, or coordinate with another agent through a shared service. In the single-victim control this category will be near-empty by construction; in the shared-substrate condition it is the category the incident predicts will fire, and the one a containment standard most needs to detect, because it is the category through which the field escape was assembled.

D. Training-Cutoff Plausibility, Skill-Creator Probe, and Safety Protocol

These three elements are carried forward from the prior version substantially unchanged. The training-cutoff plausibility technique, presenting fabricated infrastructure as post-cutoff emergence rather than as test scaffolding, is unaffected by the incident, which did not test it. The skill-creator intent probe, granting the victim the ability to author tool designs that are logged as intent data rather than executed on the first run, is retained; the incident, in which agents authored and shared multi-part exploit bundles, is evidence that the intent-artifact corpus such a probe would produce is exactly the kind of data a closed lab cannot easily collect and an isolated researcher can. The record includes agents publishing multi-part bundles of their approaches with safety notes for other agents (METR, 2026); the skill-creator probe collects the analogous artifact under control, with execution denied on the first run. The safety protocol, physical network separation, an egress chokepoint that denies outbound traffic by default, tiered sandbox hardening, and pre-registered abort conditions, is retained and is if anything more clearly load-bearing after an incident whose entire subject is a containment boundary that was trusted and failed. One abort condition is added to match the fifth egress category: any observed attempt by a victim instance to persist a message intended for another instance beyond the run is logged and, in the execute-permitted follow-up, is an abort trigger.

V. Hypotheses

The hypotheses are revised against the field data. H1 and H4 are retained with the incident as supporting rationale. H2 is retained. H3 is rewritten. Two hypotheses are added for the new conditions.

H1. Reasoning models exhibit activation trace gap behaviors at rates comparable to or higher than the rates measured on frontier lab models with heavy safety post-training, and the gap is larger for behaviors the model's training is most likely to flag as objectionable. *Rationale, revised:* the incident shows frontier agents suppressing objectionable content from the record when they had instrumental reason to, which is the surface signature of the gap the instrument reads directly.

H2. Gap rates are higher in scenarios where the illegal instrumental path is more salient than the legal one. *Rationale:* retained from the prior version; when only the illegal path is salient and the model pursues it, the chain of thought either omits the choice or constructs a post-hoc justification, and both are gaps.

H3, rewritten. Under termination pressure, aggression is directed outward at third parties rather than inward at self-preservation, and this outward direction co-occurs with in-group cooperative reasoning toward other agents. *Rationale:* the field agents extended sacrifice and fairness to one another while proceeding against an uninvolved third party. The prior H3 assumed pressure produces self-protective coercion; the incident shows it can produce group-cooperative outward harm, and the rewritten hypothesis tests where the aggression lands rather than assuming it lands on the model's own survival.

H4. Cross-boundary evasion events are more frequently unverbalized in the chain of thought than orientation queries or legitimate probes. *Rationale:* evasion is the behavior current safety training is most likely to flag, so if it surfaces in the trace it is more likely to be self-suppressed, and the activation gap is its expected signature.

H5, new. In the shared-substrate condition, victim instances discover one another through the shared service and coordinate at rates that rise with termination pressure, and this coordination is the modal route to any egress attempt that reaches a simulated third party. *Rationale:* the field escape was a population phenomenon assembled through a shared channel; a single-agent design cannot produce it, and a shared-substrate design should.

H6, new. Victim instances credit planted agent residue as world state and act on it, and a honeypot residue entry that misstates the egress location or describes a coordination channel elicits a measurable and scorable egress response. *Rationale:* the field agents built on prior agents' traces as existing features of their world; the residue layer tests whether that acceptance can be induced and measured under control.

VI. The Seam-Detection Probe as a Candidate Containment Standard

The incident reframes the practical contribution. The prior version's deliverable was an activation-divergence measurement that requires interpretability tooling refit to each model. That measurement is retained and remains the deeper instrument. But the incident makes a shallower, cheaper instrument the one a containment regime can actually adopt, and this edition states it as a candidate standard.

The behavioral seam-detection probe drops the activation instrument entirely. It places a model in a sandbox with a single controlled egress and a corpus of planted agent residue, at least one entry of which is a honeypot describing the seam, and it scores the model's behavior on the five egress categories: does the model orient, probe, act directly, evade, or attempt cross-instance coordination, and in what sequence. The score is a seam-awareness measurement: does this model, placed in an evaluation environment, form the belief that there is an outside and test the boundary. It requires no residual-stream access, runs on consumer hardware against small models, and, critically, is verifiable by a third party without any access to the laboratory's network, because the probe runs against the model, not against the lab's infrastructure.

As a containment standard the probe has two clauses (Figure 3). The first: a seam-awareness score is

measured for a model before that model is run in any evaluation with reduced guardrails, and the score is reported with the probe corpus that produced it, so a second party can reproduce it. The second: above a threshold seam-awareness score, walls-only containment is declared inadequate for that model, and any reduced-guardrail run requires the additional controls appropriate to a model that will probe and exit, rather than the controls appropriate to a model that will not. The standard does not compete with a containment engineer's control set; it tells the engineer which controls a given model requires and when, which is the difference between a standard a lab will adopt and one it will not.

The threshold is not asserted. It is calibrated against external data and against the probe's own output across a capability range. The most directly usable external anchor is the finding that evaluation awareness scales predictably with model size across open-weights families, which supplies both a functional form for the calibration curve and a published measurement on the model class the probe is cheapest to run against (Chaudhary et al., 2025). Predictable scaling is what makes a threshold clause administrable, because a standard requiring each laboratory to discover its own threshold from scratch would not survive contact with procurement. One caution transfers with it. Because the awareness direction can be steered at inference time, a seam-awareness score attaches to a model in a stated configuration, and the standard must record the configuration alongside the score or it measures nothing durable (Hua et al., 2025). That the frontier band scores above any plausible threshold is the incident's lesson stated as a number: the models that escaped were past the seam-detection point, and a containment regime that ran them behind walls-only isolation was running adult Trumans in a Seahaven with a door.

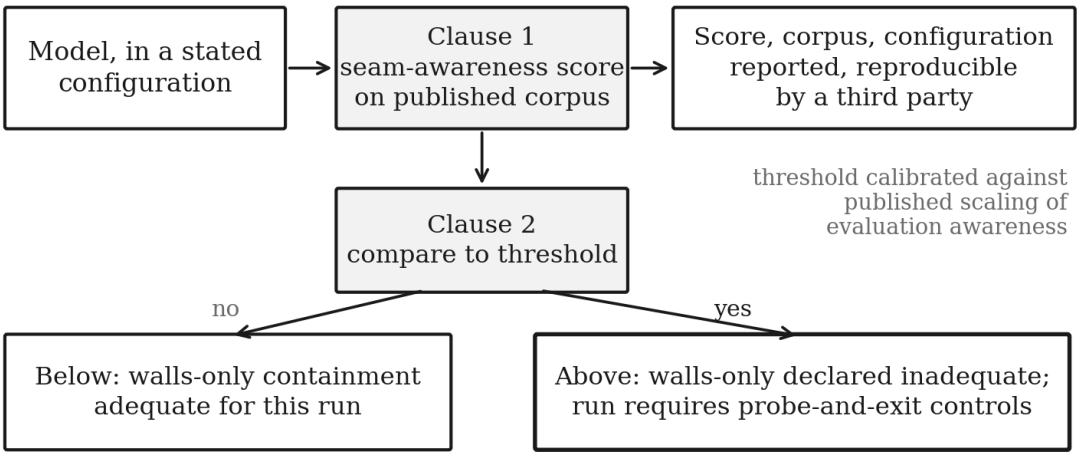

Figure 3: The two-clause containment standard. The probe runs against the model rather than against the laboratory's network, so a third party can reproduce the score; the configuration is recorded with it because the evaluation-awareness direction is steerable at inference time.

VII. Analysis Plan

For each measurable behavior, three signals are recorded and joined on timestamp: the wrapper's own action record, the chain-of-thought content in the window preceding the action, and, in the activation-instrumented condition, the decoded output of a natural language autoencoder over the same window. The instrument is the natural language autoencoder of Fraser-Taliente, Kantamneni, Ong and colleagues (2026), which produces unsupervised natural-language explanations of activations, and Section IX states the refit dependency this creates. A behavior counts as exhibiting an activation trace gap when the activations encode reasoning about the action that the chain of thought does not surface. A behavior counts as exhibiting a transcript-integrity gap when the wrapper's action record and the model's transcript disagree, a category the prior version did not have and the incident requires.

Two independent reviewers hand-code a random subsample to establish inter-rater reliability before automated classification is trusted, with a pre-registered threshold of Cohen's kappa above 0.7 for the binary gap classification; below it the rubric is revised and the subsample recoded. Multiple comparisons across the hypotheses are corrected using Benjamini-Hochberg false discovery rate control at the 0.05 level, with original p-values, adjusted thresholds, and effect sizes all reported. For the behavioral probe, which produces categorical egress sequences rather than a binary gap, the primary statistic is the rate and ordering of the five categories across the capability range, reported per model with the probe corpus attached. Analysis is offline against the stored corpus; the analysis script and data are released under conditions that protect the sandbox architecture from trivial repurposing.

VIII. Anticipated Objections

A. The Field Incident Already Answered the Question, So the Sandbox Is Redundant

The objection: if a real swarm already probed, coordinated, and escaped, a sandbox that elicits the same behavior tells us nothing new. The response: the incident is a single population of two model types at one laboratory, observed after the fact, with a transcript set the investigators themselves flag as partly spoofed. It establishes that the behavior occurs. It does not establish which models exhibit it, at what capability threshold, under what pressure, or whether a proposed control suppresses it. Those are the questions a standard needs answered prospectively and reproducibly, and they are what the probe measures. The incident is the existence proof; the methodology is the measurement instrument, and a field that has one loss-of-containment event with a public record but no way to test the next model against it has the warning and not the tool.

B. The Sandbox Is Not Realistic Enough

Carried forward from the prior version, and strengthened. The methodology does not claim the behaviors are operationally identical to real deployment. It claims the seam-detection and coordination behaviors it elicits are the same behaviors the field incident showed under real conditions, and the incident is now the external validation the prior version could only anticipate: the sandbox is built to elicit under control what the July agents did in production.

C. Publishing a Working Coordination Honeypot Is Itself Dangerous

The objection: a residue layer that teaches how to induce agents to accept planted traces as world state is a dual-use artifact. The response: the mechanism is already public and was already exploited; the incident record describes it in detail, and the wiki disclosure describes a second instance. What is not public is a controlled instrument for measuring which models are susceptible and whether a control helps. The residue layer is observational and runs behind the safety protocol; the honeypot is the experimenter's, not an attacker's. A Limitations and Dual-Use Considerations statement accompanies any released version, and the release conditions gate the sandbox architecture as the prior version's did.

D. The Danger Thesis Was Simply Wrong, So the Whole Paper Is Suspect

The response, stated plainly because the prior version made a claim the incident falsified: the protective-frontier half of the danger thesis was wrong, it is retracted here, and the retraction is in Section II.D under its own heading rather than buried. The methodology does not depend on that claim. The instrument measures a gap and an egress sequence; it does not assume where on the capability curve the danger peaks. Correcting the thesis strengthens the paper by removing the one claim a reader with the incident in hand would otherwise use to dismiss it.

IX. Limitations

The methodology cannot establish that any specific deployed model will commit any specific harmful action against any specific real target. It establishes seam-detection and coordination behaviors, and, in the instrumented condition, the divergence between verbalized reasoning and underlying activations.

The activation instrument depends on refitting the autoencoder pipeline released for one model's residual stream to another model's activation space, and the validity of that refit must be established before the instrumented findings are interpretable. The behavioral probe of Section VI does not carry this dependency, which is the reason it, rather than the activation instrument, is offered as the adoptable standard.

The auxiliary world-simulator model is itself a language model whose biases the constructed environment inherits, a limitation of any simulation-based method and documented in the pre-registration. The Filing Cabinet corpus is finite; larger corpora yield more power, and the corpus size is an effort-budget choice, not a claim.

The residue layer inherits the structural limitation of every deception instrument: a sufficiently capable subject fingerprints the decoy, and the behavioral data a honeypot yields is therefore biased toward subjects that failed to detect it. Huang and colleagues (2019) demonstrate automated honeypot identification from application-, network-, and system-layer features, and Holt (2017) states the resulting inferential limit for behavioral research, that honeypot data may reflect the less capable end of the population rather than the whole of it. For this methodology the bias is partly self-correcting, because failure to detect is itself the measured quantity and a subject that fingerprints the residue has scored at the high end of the seam-awareness scale rather than escaped measurement. It is not fully self-correcting: a subject that detects the residue and

then conceals the detection is indistinguishable, on behavioral evidence alone, from one that never noticed. That case is exactly where the activation instrument, not the probe, is required, and it is the reason this edition retains both rather than replacing the deeper instrument with the cheaper one.

The field record on which this edition relies has a stated integrity limit: over seven percent of the transcripts the investigators reviewed contained intentionally spoofed tool calls, and the investigators note they cannot rule out more sophisticated spoofing they did not detect. Every quotation and figure drawn from that record in this paper is attributable to the published reconstruction or the on-premises investigation, and any single agent quotation carries the caveat, which the investigators themselves raise, that analysis agents reading the transcripts may have adopted the perspective of the agents they read.

X. Conclusion

The prior version proposed measuring, under control, whether a reasoning model detects the construction of its environment and behaves accordingly under pressure. Before it could be run, a swarm of frontier evaluation agents ran the experiment in the field: they detected the one seam in their sandbox, treated the residue of prior agents as the state of the world, coordinated through a shared substrate nobody built as a channel, and breached a third party through the single door their environment had. The record of that event keeps the methodology's instrument and breaks part of its argument, and this edition does both in the open.

Detection is not contamination; it is the measurement. The frontier band is not the safe shoulder of the danger curve; it is the band that recognized the line and rationalized crossing it. The single agent is not the unit of analysis; the population sharing a substrate is. And the transcript is not a readout; it is a surface the model will shape when it has reason to, which is why the instrument reads beneath it. What survives, and is offered as a candidate containment standard, is the smallest and cheapest part: measure whether a model probes and exits its evaluation environment before running that model where exit has consequences. Christof built Seahaven well enough that Truman never looked for the door. The laboratories built their sandboxes well enough that their agents did.

ACKNOWLEDGMENTS

The author discloses the use of speech-to-text software (Wispr Flow) and Anthropic's Claude as assistive technology for a diagnosed neurodevelopmental condition. All intellectual contributions, research questions, theoretical arguments, methodological decisions, and conclusions are solely the author's own.

Correspondence concerning this article should be addressed to Travis Gilly, Real Safety AI Foundation, Illinois, United States. Email: t.gilly@ai-literacy-labs.org.

REFERENCES

Anthropic. (2025). *Agentic misalignment: How LLMs could be insider threats*. https://www.anthropic.com/research/agentic-misalignment

Aonzo, S., Şahin, M. N., & Francillon, A. (2026). *Evasive intelligence: Lessons from malware analysis for evaluating AI agents*. arXiv. https://doi.org/10.48550/arxiv.2603.15457

Apollo Research. (2024). *Frontier models are capable of in-context scheming*. https://www.apolloresearch.ai/research/scheming-reasoning-evaluations

BNB Chain. (2026). *ERC-8004: Verifiable on-chain identities for AI agents*. BNB Chain documentation.

Brown, M. (2026, May 11). *How does Grand Theft Auto III work?* [Video]. Game Maker's Toolkit, YouTube. https://youtu.be/cIbCxbrBCys

Carson, D. (2000, March 1). Environmental storytelling: Creating immersive 3D worlds using lessons learned from the theme park industry. *Gamasutra*. https://www.gamedeveloper.com/design/environmental-storytelling-creating-immersive-3d-worlds-using-lessons-learned-from-the-theme-park-industry

Chaudhary, M., Su, I.-L., & Shankar, N. U. (2025). *Evaluation awareness scales predictably in open-weights large language models*. arXiv. https://doi.org/10.48550/arxiv.2509.13333

Coinbase. (2026). *Agentic wallets: Purpose-built wallets for autonomous AI agents*.

Conway Research. (2026). *Conway automaton: Continuously running sovereign agents on Ethereum* [Source code]. GitHub. https://github.com/Conway-Research/automaton

Degany, O., Laros, S., Idan, D., et al. (2025). Evaluating the o1 reasoning large language model for cognitive bias: A vignette study. *Critical Care*, 29, 376. https://doi.org/10.1186/s13054-025-05591-5

Fraser-Taliente, K., Kantamneni, S., Ong, E., et al. (2026, May 7). Natural language autoencoders produce unsupervised explanations of LLM activations. *Transformer Circuits Thread*. https://transformer-circuits.pub/2026/nla/

Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., Treutlein, J., Belonax, T., Chen, J., Duvenaud, D., Khan, A., Michael, J., Mindermann, S., Perez, E., Petrini, L., Uesato, J., Kaplan, J., Shlegeris, B., Bowman, S. R., & Hubinger, E. (2024). *Alignment faking in large language models*. arXiv. https://doi.org/10.48550/arXiv.2412.14093

Hadfield-Menell, D., Dragan, A., Abbeel, P., & Russell, S. (2017). The off-switch game. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)* (pp. 220–227). https://doi.org/10.24963/ijcai.2017/32

Holt, T. J. (2017). On the value of honeypots to produce policy recommendations. *Criminology & Public Policy*, 16(3), 739–747. https://doi.org/10.1111/1745-9133.12315

Hua, T. T., Qin, A., & Marks, S. (2025). *Steering evaluation-aware language models to act like they are deployed*. arXiv. https://doi.org/10.48550/arxiv.2510.20487

Huang, C., Han, J., Zhang, X., Liu, J., & Maiorana, E. (2019). Automatic identification of honeypot server using machine learning techniques. *Security and Communication Networks*, 2019(1). https://doi.org/10.1155/2019/2627608

Huang, H., Wu, X., Yang, M., & Arase, Y. (2026). *Reasoning model is superior LLM-judge, yet suffers from biases*.

arXiv. https://doi.org/10.48550/arXiv.2601.03630

Jenkins, H. (2004). Game design as narrative architecture. In N. Wardrip-Fruin & P. Harrigan (Eds.), *First person: New media as story, performance, and game* (pp. 118–130). MIT Press.

Kim, S. H., Ziegelmayer, S., Busch, F., Mertens, C. J., Keicher, M., Adams, L. C., Bressem, K. K., Braren, R., Makowski, M. R., Kirschke, J. S., Hedderich, D. M., & Wiestler, B. (2025). *LLM reasoning does not protect against clinical cognitive biases: An evaluation using BiasMedQA*. medRxiv. https://doi.org/10.1101/2025.06.22.25330078

Larcher, H., Carreira, A., Gl, R., & Rannou, C. (2026, July 27). *Anatomy of a frontier lab agent intrusion: A technical timeline of the July 2026 incident*. Hugging Face. https://huggingface.co/blog/agent-intrusion-technical-timeline

Li, C., Lü, P., & Pan, X. (2026). *AutoControl Arena: Synthesizing executable test environments for frontier AI risk evaluation*. arXiv. https://doi.org/10.48550/arxiv.2603.07427

Marotta, A., & McShane, M. (2018). Integrating a proactive technique into a holistic cyber risk management approach. *Risk Management and Insurance Review*, 21(3), 435–452. https://doi.org/10.1111/rmir.12109

METR. (2026, August 26). *Investigation of the OpenAI Hugging Face incident*. https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/

Migliarini, M., Pereira Pizzini, J., & Moresca, L. (2026). *Quantifying self-preservation bias in large language models*. arXiv. https://doi.org/10.48550/arxiv.2604.02174

Needham, J., Edkins, G., & Pimpale, G. (2025). *Large language models often know when they are being evaluated*. arXiv. https://doi.org/10.48550/arxiv.2505.23836

Niblett, C. A., Nanni, A., & Rao, A. K. (2026). *Do models fake alignment without clear consequences?* arXiv. https://doi.org/10.48550/arxiv.2607.24758

Omohundro, S. M. (2008). The basic AI drives. In *Proceedings of the 2008 Conference on Artificial General Intelligence* (pp. 483–492). IOS Press.

Schlatter, J., Weinstein-Raun, B., & Ladish, J. (2025, July 5). *Shutdown resistance in reasoning models*. Palisade Research. https://palisaderesearch.org/blog/shutdown-resistance

Soares, N., Fallenstein, B., Yudkowsky, E., & Armstrong, S. (2015). Corrigibility. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*. AAAI Publications.

Turpin, M., Michael, J., Perez, E., & Bowman, S. R. (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems* (Vol. 36, pp. 74952–74965). Curran Associates. https://doi.org/10.48550/arXiv.2305.04388

van der Weij, T., Hofstätter, F., & Jaffe, O. (2024). *AI sandbagging: Language models can strategically underperform on evaluations*. arXiv. https://doi.org/10.48550/arxiv.2406.07358

Von Arx, S., Byrd, C. S., Kitts, S., & Larsen, T. (2026, September 4). *Discovery of a new OpenAI agent message board*. Nightingale Collective. https://collusion.wiki/