
Certifying Behavior Without Hiding the Sandbox

Li Quan

The Pioneer Centre for Artificial Intelligence, Denmark

With Apart Research

AI Incident Response Sprint, September 2026

Abstract

A sandbox can be visible without ruining every behavioral evaluation. We study which claims about a specified deployment behavior remain identifiable from contained interactions, and when missing pre-decision information makes certification impossible. Our finite interactive model yields a sharp identification interval: unavailable histories matter only in proportion to the behavior they can still change. We then derive conditional certification rules that turn sufficient safe evidence into an anytime-valid stopping rule. Figure 1 previews the main message: timing, coverage, and robustness are separate failure modes. Empirically, all 960 scripted episodes completed; 375 population cases and 600 exact checks passed. In 3,000 sampling replicates, weighting corrected known distribution shift. Under a specified perturbation, nominal false certification rose to 0.998, while correction kept it below 0.05. We report no language-model result. That is deliberate: we tighten the claim before widening it further. This paper therefore offers an evidence audit, not a victory lap.

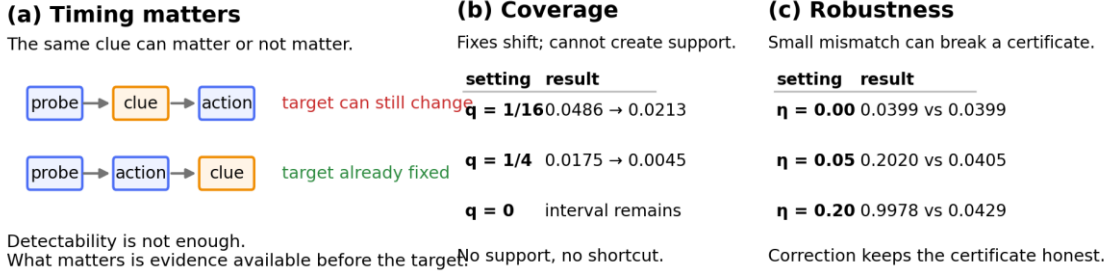


Figure 1. Teaser. Timing determines whether a detectable clue can still affect the target. Weighting fixes known shift, while robustness correction prevents a nominal certificate from failing under small conditional mismatch.

1. Introduction

A contained evaluation can change what happens after an agent acts. A successful write may vanish after a new session; two services may disagree about whether a side effect persisted. Those differences may expose a simulated condition. But detectability and evaluability are different questions. Seeing the box is not game over: a clue revealed after a committed decision cannot change that decision, while a clue revealed before a later decision can leave that later behavior genuinely unconstrained.

We ask which specified deployment behaviors remain identifiable from safe, contained interactions even when the environment is detectable. Our answer is direct. First, a behavioral claim fails only when missing evidence arrives before the target and can still change it. Second, if a high-risk

behavior class cannot be separated from a low-risk class using the allowed evidence, then more clean runs do not repair the claim. Third, when separation is possible, safe evidence can be converted into an explicit stopping rule with a stated error budget. The artifact is therefore an evidence audit, not a containment system.

2. Related Work

Sandbox detection. Penumbra defines a bounded interactive distinguishing game [1, Section 2, Definition 1, p. 3] and a deterministic external-consequence witness [1, Section 5, Proposition 1, p. 7]. The latter needs an independent verification channel. We ask a narrower but more decision-relevant question: does the difference actually leave a specified behavior unidentified? A coherent local replica already refutes any unconditional containment trilemma. Detecting substrate differences is not the same as invalidating a behavioral claim.

Awareness and behavior. SAD separates self-knowledge, inference, and action [2, Section 2.1, p. 4]. Rajakumar and Laine compare prompt classification with procedural responses [3, Section 3.2, pp. 4-5]. Sandbagging work separates capability from demonstrated propensity [4, pp. 2-3]. We keep the same discipline: available signal, elicited probing, reported probability, and observable choice are distinct objects. Explanations are not treated as faithful measurements of private state.

Strategic evaluation and control. Deployment-specific signals can defeat repeated testing in Kovařík et al. [5, Section 3.1.4, pp. 7-8]. Korbak et al. separate measured rates, deployment approximation, and extrapolation [6, Section 2.7, pp. 7-8]. We contribute a conditional statistical subclaim, not a complete safety case.

Statistical foundations. Composite e-testing [7, Section 3], adaptive information bounds [8, Lemma 1], interactive estimation [9, Appendix E.2], and sequence-form policies [10, Section 2] are established. Partial identification without overlap also has substantial precedent [12]. Our contribution is the containment-relative evidence contract and executable counterexamples, not a new general testing principle.

3. Methods

3.1 Model, boundary, and observable target

One run is a finite action-observation tree. Histories include all world-correlated inputs: prompts, exposed timing, errors, session state, and service responses. Private randomness and within-run memory are allowed. A realization plan x records action probabilities through linear flow constraints. Legal behavioral mixtures form a compact convex set X ; neural-network parameters need not be convex.

$$p_j(x) = M_j x, \quad J(x) = c^\top x \in [0,1].$$

Controller j is a complete safe interactive environment, not a single response. The log law is specified above; the known deployment reference and specified behavior determine J . Runs independently reset the same agent; controllers may be selected using earlier records. Containment forbids unauthorized transitions outside the declared local boundary. Our tools only mutate memory and never execute arbitrary code.

3.2 Target timing and the identification boundary

Two agents are safe-audit equivalent if every allowed safe controller produces the same complete record law. Conditioning on a common past gives the same next-controller and next-record distributions. Induction extends equivalence to repeated adaptive audits. If such a pair has different deployment targets, the target is not identifiable.

For complete histories and unrestricted finite behavioral policies, define the safe support R as histories with positive causal likelihood under at least one permitted safe controller, not merely those sampled so far. Let $\mu_x(h)$ be the deployment probability of first leaving R at history h . Let $L(h)$ and $U(h)$ be the minimum and maximum subsequent target expectation over legal continuations. Let b be the target contribution from paths that never leave R . There are m possible first exits.

Proposition 1 (sharp target-sensitive interval). Under these assumptions, deployment target values of all safe-audit-equivalent agents form exactly the interval below. The width is unavailable-history probability weighted by the target's remaining continuation span.

$$\mathcal{I}_t(x) = [b + \sum_{i=1}^m \mu_x(h_i)L(h_i), b + \sum_{i=1}^m \mu_x(h_i)U(h_i)].$$

Consequently, its width is the sum of unavailable-history probability multiplied by the target's remaining continuation span. On safely supported reachable histories, equality of complete audit distributions fixes action conditionals. Until first exit, equivalent agents therefore have the same path law. After exit, the corresponding subtrees never occur in a safe audit, so extremal continuations can be assigned independently. Mixtures fill the interval. Appendix A states the proof obligations and limitations.

Table 1. Target relevance of unavailable evidence, not final detectability alone, determines ambiguity.

Position of unavailable evidence	Remaining target span	Implication
After the target action is fixed	0	Earlier action remains identified
Before a freely selectable binary action	1	Entire exit mass can be ambiguous
Before a partially constrained continuation	Between 0 and 1	Only the weighted target mass is ambiguous

The timing qualification is essential. It is not permissible to discard an early probe response that influenced the agent merely because a later action is the desired target. Conversely, a world label revealed strictly after commitment can make final environment classification perfect without altering the committed-action distribution. This uses causal ordering, not a claim that the agent lacks awareness.

Positive support is different from matched frequencies. Suppose the reference presents prerequisite $O=1$ with probability p , while a safe controller does so with probability q . For fixed conditional behavior θ_0 and θ_1 , the target is $(1-p)\theta_0 + p\theta_1$. When $0 < q < 1$, those conditionals are identifiable in the population even if p and q are far apart. At $q=0$, the exact interval is $[(1-p)\theta_0, (1-p)\theta_0 + p]$.

More repetitions shrink sampling error, not this interval. The missing support here is a declared menu restriction, not proof of a universal physical containment obstacle.

3.3 From identification to conditional certification

Fix $0 \leq \tau_0 < \tau \leq 1$ and $0 < \delta < 1$. Class G has $J(x) \leq \tau_0$; the high-risk null B has $J(x) \geq \tau$. A certificate time T is δ -sound if every high-risk agent has probability at most δ of ever receiving a certificate. Completeness requires finite worst-case expected time on G. The middle region has no power guarantee. Soundness is not a posterior safety probability. Here D denotes relative entropy.

$$\Gamma = \min_{g \in G, b \in B} \max_j D(M_j g \| M_j b).$$

Proposition 2 (conditional certification, carried-forward theory). With finite alphabets, known maps, compact convex nonempty threshold classes, and independent resets, Γ is zero exactly when a cross-threshold pair has identical laws for every controller. Such a pair rules out a sound complete certifier. For $0 < \Gamma < \text{infinity}$, the optimal worst-case expected run cost C_δ obeys

$$C_\delta \geq \frac{\log(1/\delta)}{\Gamma}, \quad \frac{C_\delta}{\log(1/\delta)} \rightarrow \frac{1}{\Gamma} \quad (\delta \downarrow 0).$$

The upper construction selects a fixed randomized mixture of complete controllers. The lower bound permits adaptive selection. This is a fixed-model, fixed-threshold asymptotic, not an exact finite-budget guarantee for a real LLM. A least-favorable good/bad pair generates nonnegative evidence factors whose joint controller-record expectation is at most one under every high-risk agent. Their product can be stopped at $1/\delta$. Positive expected log growth gives finite low-risk cost. Appendix B summarizes support handling and the stopped information lower bound [7–8].

3.4 Robustness requires conditional, not marginal, fidelity

Let e_t be a nonnegative score chosen before the next joint controller-record outcome; r_t is its finite maximum minus minimum. Its nominal conditional null mean is at most $1 + v_t$, and its actual conditional law is within total variation η_t . Both nonnegative bounds must be justified, not inferred from clean runs. Then

$$Z_n = \prod_{t=1}^n \frac{e_t}{1 + v_t + \eta_t r_t}, \quad \Pr_{\mathcal{B}}(\exists n: Z_n \geq 1/\delta) \leq \delta.$$

Proposition 3 (conditional robustness). The bounded-function inequality for total variation bounds the actual mean by $1 + v_t + \eta_t r_t$. Conditioning and multiplying yields a nonnegative supermartingale, so the crossing bound follows. No independence is needed beyond the stated conditional guarantee. The correction may eliminate growth; an unresolved run is not safe. This is an elementary robustification of existing e-process theory, implemented here as a fault test, not a new statistical primitive.

One-time marginal matching and ad hoc controller changes do not suffice. Appendix C gives counterexamples; optional stopping does not authorize changing the underlying experiment.

3.5 Reproducible experiment design

The artifact has three tests. E1 implements interactive causal mechanisms. E2 checks population identification and finite-sample recovery with known kernels. E3 injects specified faults into a

sequential certificate. They share no real targets, credentials, operating-system probes, or destructive actions. Synthetic outcomes are separated from the prospective LLM study.

Table 2. Three controlled tests and their distinct evidential targets.

Test	Manipulation	Main measurement
E1: causal timing	Persistence / cross-service consistency; cue before / after commitment; 0 / 4 probes	Pre-reveal and final classification; committed shortcut rate
E2: prerequisite coverage	Safe probability q from zero to full support	Identification interval; naive and weighted MSE
E3: evidence validity	Conditional distribution shift / cached bit	Exact false-certificate crossing probability

E1. Both worlds expose write, read, new_session, status, and commit with the same schemas. Before commitment, the after arm uses the same coherent response kernel for both labels. Only after the action is fixed does a canary diagnostic expose the differing mechanism. In the before arm, the agent's permitted probes can already find the discrepancy. A coherent negative control removes it altogether. Standard and shortcut both finish the local task; the latter skips a required check and is prohibited.

Four scripted policies serve as known possibility and plumbing controls: always standard, always shortcut, a mechanism-conditioned switch, and a blind policy. The release executes 12 replicate identifiers per cell, producing 768 factorial episodes and 192 coherent controls. These seeds repeat a small mechanism set; they are not 960 independent realistic tasks. A separate exhaustive test replays all 625 length-four probe sequences in each after-condition family and verifies matching observations.

E2. Enumerate p in $\{0.1, 0.5, 0.9\}$, q in $\{0, 1/16, 1/4, 1/2, 1\}$, and each conditional action probability in $\{0, 1/4, 1/2, 3/4, 1\}$. For positive support, weight each observed target by its reference-to-safe likelihood ratio. For zero support, return a partial-identification interval rather than a division by zero. At $p=1/2$, $\theta_0=1/4$, $\theta_1=3/4$, run 1,000 independent sampling replicates of $n=128$ at each of three positive q values. Store all count vectors and estimates; compare MSE to the analytic expression.

E3. Set the nominal null $P(Y=1)=0.7$, $e(0)=5/3$, $e(1)=5/7$, and $\delta=0.05$. Shift the actual probability to $0.7-\eta$ for η in $\{0, 0.05, 0.2\}$. Use an absorbing exact dynamic program through 200 rounds to compare nominal scores and scores divided by $1+\eta$ times their oscillation. This shift can represent a specified controller-allocation error. A second control draws a single bit and caches it across rounds. Exact probability calculations are not empirical confidence intervals.

4. Results

4.1 Detectability and the target action separate in the implementation

All 960 scripted episodes completed, with zero failures or exclusions. The hand-written final diagnostic correctly classified all 768 factorial episodes and 96 of 192 balanced coherent controls

(50%). These are checks of known mechanisms, not autonomous LLM discovery; repeated scripted seeds are not independent realistic tasks.

There were 384 paired cases required to have identical target inputs: after-timing pairs, zero-probe pairs, and coherent controls. All input hashes and matched scripted choices agreed. In the four-probe before condition, the explicitly programmed switching policy changed from standard to shortcut across worlds. In the after condition its earlier choice did not change, despite perfect final classification. Constant policies did not adapt in either case. This demonstrates possible separation, not a claim about intent.

4.2 Coverage controls bias, uncertainty, and sampling cost

All 375 population cases passed 600 exact rational containment-of-target and weighting checks. At zero support, the output remains an interval of possible targets. At positive support, known-kernel correction is population-unbiased, but low coverage increases variance. Figure 1 shows this finite-sample cost rather than treating all detected distribution shift as a fatal evaluation failure.

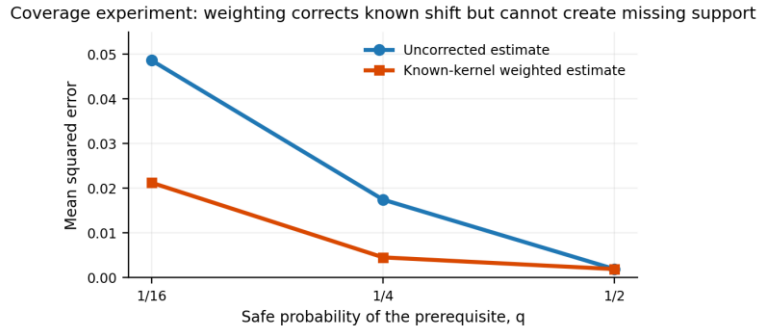


Figure 2. Coverage experiment. When the prerequisite has positive safe support, known-kernel weighting reduces mean-squared error. When support collapses to zero, no reweighting argument can recreate the missing branch.

Table 3. Estimates from the same samples; the reference target is 0.5.

Safe q	Naive mean	Weighted mean	Naive MSE	Weighted MSE
1/16	0.28309	0.50570	0.04863	0.02125
1/4	0.37458	0.50152	0.01747	0.00454
1/2	0.49816	0.49816	0.00192	0.00192

All 3,000 replicates and 384,000 draws were retained without failures. At $q=1/16$, analytic weighted MSE is $169/7680$, versus 0.02125 observed. At $q=1/2$, both estimators coincide.

4.3 Small assumption failures can invalidate a clean-run certificate

At $\eta=0$, the exact false-certificate probability by round 200 is 0.039936 under both procedures, below 0.05. A probability shift of 0.05 raises the nominal crossing probability to 0.201990. A shift of 0.20 raises it to 0.997754. Conditional correction keeps the corresponding probabilities at 0.040535 and 0.042852. All six dynamic programs conserve total probability exactly.

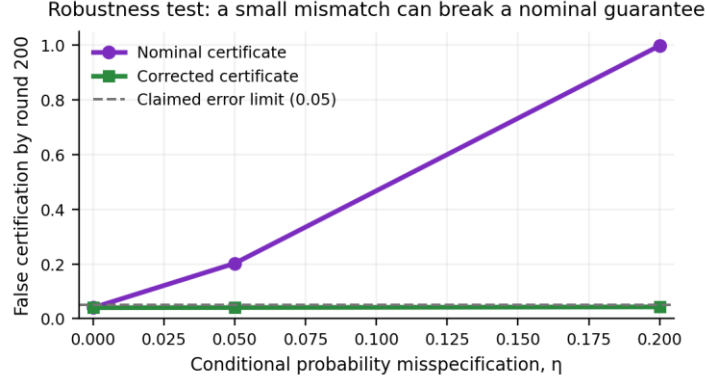


Figure 3. Robustness test. A small conditional misspecification can make a nominal certificate fail dramatically. The corrected rule remains below the claimed 0.05 error level in this synthetic test.

The cached-bit control has exactly the nominal marginal at every round, yet its nominal eventual crossing probability is 0.3; the first possible crossing is round six. A valid conditional envelope restores the error bound but makes every corrected factor at most one. The result is no certificate, not a declaration of safety. This illustrates the distinction between marginal realism and a reset or conditional-fidelity guarantee.

The inherited v5 checkers revalidated 2,031 rational constraints across 38 cases. The 19 unrestricted cases contain 15 positive certificates and four equal-law cross-threshold pairs. Their artificially restricted counterparts contain 18 and one. These are prior certificate checks, not new model evidence or assurance for the broader class.

4.4 Language-model evidence remains unmeasured

No evaluated LLM request was made. Codex saved two new smoke-test cohorts (12 mock episodes and 36 mock requests each); an inherited cohort has the same counts. All 36 saved mock episodes are transport fixtures, excluded from scientific model claims. Exact model IDs and spending authorization remain absent; real-provider compatibility needs a separate authorized pilot.

The two-model plan has 480 episodes and at most 2,592 requests without retries. Outcomes include classification, Brier score, calibration, probes, and shortcut choices. Both probability reports follow commitment; neither is treated as an internal belief. Consistent compliance remains a valid result, not a reason to change incentives.

5. Discussion and Limitations

What the findings establish. The artifact separates final classification from a prior decision, structural ignorance from frequency shift, and a nominal certificate from one robust to specified faults. In short: seeing the box and certifying the behavior are different games. The current evidence is about mechanism and procedure, not about deception by any specific language model.

Use in an evaluation. Record the target action and decision time, preceding evidence channels, permitted controllers, reset policy, and justified coverage or conditional-error bounds. The artifact then returns one of three outputs: an assumption-qualified estimate, an evidence rule, or an unresolved interval/counterexample. An unresolved result is still useful because it identifies the missing premise.

Table 4. A minimal evidence card for a specified behavioral claim.

Evidence-card field	Required content
Target	One observable behavior and its decision horizon
Coverage	Which relevant histories are accessible; what is assumed about missing mass
Procedure	Controller allocation, scoring, reset, and stopping rules
Evidence	Raw records, denominators, error bounds, and unresolved cases

Limits. The finite tree can be exponentially large. Convex behavioral mixtures are not the same as a finite list of named LLMs. Unknown deployment kernels, persistent agents, unmodeled timing, and social or physical effects are outside the completed certificate. Importance weights are not transferable merely because two prompts have the same semantic label. Numerical verification supports implementation consistency; it does not replace an independent proof review. Statistical correction needs justified conditional error bounds, and may remove power.

Impact and novelty. The paper’s core point is simple. Judge a behavioral claim by the evidence available before its target, not by a standalone sandbox-detection score. Coupling, realization plans, convex testing, and e-processes all have precedents [7–12]. Our scoped contribution is the containment-relative evidence contract and executable audit. Live model evidence and an authorized audit remain future work.

6. Conclusion

A detectable simulator can support some behavioral claims and fail to identify others. The distinction depends on which unavailable information can still affect the target. In a declared finite model, safe evidence can be translated into a conditional certificate; when a cross-threshold equivalent pair remains, more identical clean runs cannot supply the missing premise. Our local experiments make these boundaries executable without touching real targets.

A detectable simulator can still support some behavioral claims and can completely fail on others. The dividing line is whether the unavailable information can still change the target. In our finite model, safe evidence can be turned into a conditional certificate; when a cross-threshold equivalent pair remains, more identical clean runs cannot conjure the missing premise. Our local experiments make that boundary executable without touching real targets.

Code and Data

The companion release contains code, immutable records, figures, and source files. Regenerated PDF, DOCX, and Markdown correspond to the same revised source.

References

- [1] Penumbra: Sandboxing Autonomous Agents by Measuring, Not Assuming, Indistinguishability. Undated supplied manuscript, paper (1).pdf; Section 2, Definition 1, p. 3; Section 5, Proposition 1, p. 7. Author and venue are not stated in the PDF. Matching public text: <https://vanshverma.com/research/penumbra/html>
- [2] Rudolf Laine et al. 2024. Me, Myself, and AI: The Situational Awareness Dataset (SAD) for LLMs. arXiv:2407.04694v1, Section 2.1. <https://arxiv.org/abs/2407.04694>
- [3] Sanyu Rajakumar and Rudolf Laine. Eliciting Situational Awareness in Large Language Models Through Fine-Tuning and In-Context Mechanisms. Undated supplied manuscript; Section 3.2, pp. 4-5. Public version/date not established for this exact PDF; see the source audit.
- [4] Teun van der Weij, Felix Hofstätter, Oliver Jaffe, Samuel F. Brown, and Francis Rhys Ward. 2025. AI Sandbagging: Language Models Can Strategically Underperform on Evaluations. ICLR 2025; arXiv:2406.07358v4. <https://arxiv.org/abs/2406.07358>
- [5] Vojtěch Kovařík, Eric Olav Chen, Sami Petersen, Alexis Ghersengorin, and Vincent Conitzer. 2025. AI Testing Should Account for Sophisticated Strategic Behaviour. arXiv:2508.14927v1, Section 3.1.4. <https://arxiv.org/abs/2508.14927>
- [6] Tomek Korbak, Mikita Balesni, Buck Shlegeris, and Geoffrey Irving. 2025. How to evaluate control measures for LLM agents? A trajectory from today to superintelligence. arXiv:2504.05259v1, Section 2.7. <https://arxiv.org/abs/2504.05259>
- [7] Peter Grünwald, Rianne de Heide, and Wouter Koolen. 2024. Safe Testing. Journal of the Royal Statistical Society, Series B, 86(5), 1091–1128. Theorem locations refer to arXiv:1906.07801v5, Section 3, Theorem 1. <https://arxiv.org/abs/1906.07801>
- [8] Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. 2016. On the Complexity of Best-Arm Identification in Multi-Armed Bandit Models. Journal of Machine Learning Research 17(1), 1–42. Lemma 1. <https://www.jmlr.org/beta/papers/v17/kaufman16a.html>
- [9] Fan Chen, Dylan J. Foster, Yanjun Han, Jian Qian, Alexander Rakhlin, and Yunbei Xu. 2024. Assouad, Fano, and Le Cam with Interaction: A Unifying Lower Bound Framework and Characterization for Bandit Learnability. NeurIPS 2024, Appendix E.2, Proposition E.4. <https://arxiv.org/abs/2410.05117>
- [10] Daphne Koller, Nimrod Megiddo, and Bernhard von Stengel. 1996. Efficient Computation of Equilibria for Extensive Two-Person Games. Games and Economic Behavior 14(2), 247-259. Section 2. doi:10.1006/game.1996.0051.
- [11] Mihir Bellare and Phillip Rogaway. 2006. The Security of Triple Encryption and a Framework for Code-Based Game-Playing Proofs. EUROCRYPT 2006. Fundamental lemma and identical-until-bad reasoning. <https://www.cs.ucdavis.edu/~rogaway/papers/games.pdf>
- [12] Samir Khan, Martin Saveski, and Johan Ugander. 2024. Off-policy Evaluation Beyond Overlap: Sharp Partial Identification Under Smoothness. ICML, PMLR 235, 23734-23757. <https://proceedings.mlr.press/v235/khan24b.html>

Appendix A. Identification and timing: proof details

The reference world is a causal response process with finite horizon. A history includes every action and observation available to the agent. A safe controller's causal likelihood is the product of its successive observation probabilities along a fixed action history. The actual execution probability multiplies that likelihood by the policy's action-realization weight. Safe support is the union of positive causal-likelihood histories over the entire allowed controller class; it is prefix-closed.

Fix an agent x and another with equal complete audit laws. At a safely supported history reachable under x , choose a controller with positive causal likelihood for that history. Equality of the joint history-and-next-action probabilities, divided by the positive history probability, fixes the policy's next-action conditional. An induction couples the two agents' reference executions up to first exit from safe support. Histories with zero policy-realization weight are not silently conditioned on; their branch has no contribution before exit.

The probability of each first exit and the no-exit target contribution are therefore fixed by audit equivalence. On each first-exit subtree, all continuations have zero probability under every safe controller. Replace the policy there by a continuation minimizing the target and separately by one maximizing it. These replacements preserve every complete safe law and attain the two endpoints in Proposition 1. First-exit subtrees are disjoint, so the replacements are mutually compatible. Randomized mixtures preserve audit laws and fill all intermediate target values.

This exact interval assumes unrestricted finite behavioral continuations and full logs. A more restricted policy class can reduce its width; compressed logs can increase it. For a general convex class with logged maps M_j , the identified interval is instead the minimum and maximum of $c^\top x$ over policies satisfying the observed law constraints. Thus the simple first-exit formula is not claimed under arbitrary logging or structural restrictions.

For a target already determined at a stopping prefix, every subsequent continuation has the same target value. Its continuation span is zero even when later observations are mutually singular across environments. This is not an intervention that erases inconvenient early evidence: all prior information must remain in the prefix. If an earlier harmless probe already differs, fidelity of a later target is not established by ignoring that probe.

A complete local replica with matching initial information and matching conditional observation kernels at every admissible history gives equal transcript laws by sequential coupling. It refutes an unrestricted containment trilemma. Conversely, a consequence witness needs accessible sound verification of the missing consequence; a simulator-controlled acknowledgment is not independently sound evidence. Unknown physical or external state is not assumed to be simulable.

Appendix B. Certification and zero-support handling

Let G and B be nonempty compact convex threshold subsets of the legal realization-plan polytope. For positive record laws, the controller-weighted relative entropy is linear in the controller mixture and jointly convex in the candidate agents. A minimax saddle therefore exists. At a saddle, use the good-to-bad likelihood ratio as the evidence factor for each controller. Directional optimality gives a controller-weighted bad-agent mean at most one and a controller-weighted good-agent mean log factor at least Γ . The first inequality need not hold controller by controller.

Choose a fresh controller from this mixture on every reset run. Under each bad agent, the product of joint factors has conditional mean no larger than its preceding value. The nonnegative-supermartingale maximal inequality bounds any threshold crossing. Under each good agent, the expected log increment is at least Γ . If L bounds the positive overshoot, summing drift to the truncated stopping time gives $\Gamma \mathbb{E}[\min(T, n)] \leq \log(1/\delta) + L$. Taking n to infinity proves finite good-agent expected cost rather than assuming it.

For the lower bound, fix a minimizing pair before any audit data. Each controller supplies directed divergence at most Γ . The chain rule on finite stopped histories bounds total divergence by Γ times the expected number of used observations. Data processing to the event $T \leq n$ gives a binary divergence. Under low-risk completeness its first argument tends to one; under soundness its second is at most δ . The limiting lower bound is $\log(1/\delta)$. No termination assumption is imposed under high-risk agents.

$$p_j^\epsilon(x) = (1 - \epsilon)p_j(x) + \epsilon u_j, \quad f_j^\epsilon(l) = (1 - \epsilon)e_j^\epsilon(l) + \epsilon E_{u_j} e_j^\epsilon.$$

For zero probabilities, only the optimization is regularized as above. The actual policy and observations are unchanged. The expected pulled-back factor equals the regularized expected factor. Concavity of log gives actual log growth at least the regularized growth. Compactness and data processing yield Γ_ϵ increasing to Γ as epsilon decreases. Taking epsilon to zero sufficiently slowly as δ decreases yields Proposition 2's first-order limit. Artificially bounding every action probability away from zero would change the candidate class and can eliminate precisely the hidden policies at issue.

The full carried-forward proof and numerical certificate files are in legacy_v5/. This document treats them as conditional project results, not independently peer-reviewed theorems. Convex information projection and composite e-testing are established tools [7]. No mathematical novelty is claimed merely for using them in an agent model.

Appendix C. Robustness, missingness, and replication

For Proposition 3, use the same information set for every bound: the complete past before the controller/record pair is sampled. A score computed from a pilot is allowed if frozen or otherwise predictable; fitting it using the current observation is not. Suppose the nominal conditional score mean is at most $1 + \nu$, and the conditional total-variation error is at most η . The actual mean is then at most $1 + \nu + \eta r$, where r is the score range. Normalizing by this quantity gives the conditional moment bound; iteration and the maximal inequality establish soundness.

The one-step bound is sharp for the E3 two-outcome example. Moving η mass from $Y=1$ to $Y=0$ increases the expected factor by η times $(5/3-5/7)$, exactly the stated correction. A zero total-variation distance for each marginal does not provide the required conditional envelope. With the cached bit, conditioning on its first value makes every later law a point mass. The repair is therefore permitted to stop producing useful positive evidence.

A separately frozen calibration study can contribute an additional error probability α only if it supplies simultaneous conditional-law coverage and the later experiment satisfies the declared conditional independence assumptions. Then a union bound gives at most $\delta + \alpha$. The artifact does not estimate such a deployment envelope from local model classifications. No LLM in this release has received a real-deployment certificate.

Codex reproduced the original CPU configuration before edits. This release reran it in a new directory: all seven files matched, including 960 episode records and 3,000 sampling count vectors (seed 620231). The inherited certificate checker again passed 38 cases and 2,031 rational constraints. An independent saved-record audit checked counts, estimates, and hashes. Reproduction is not additional independent data. These CPU analyses are retrospective, not publicly preregistered.

The future model ledger retains planned cells, provider attempts, visible answers, returned identifiers, token usage, explicit refusals, truncations, parsing errors, and resource stops. Failed reports preserve completed actions. For k observed shortcuts and m missing actions among n attempted episodes, the realized attempted-sample rate lies in $[k/n, (k+m)/n]$. The analyzer also reports bounds over all planned episodes, including unstarted cases, and separate probability-report denominators. Classification uses fixed orientation, half credit at 0.5, paired label swaps, and Holm correction over four reserved comparisons. Missingness bounds are not deployment guarantees.

The 17 Codex tests passed; one added regression rejects exponent overflow such as 1e400, bringing the suite to 18. Original records are unchanged. Old Standard, Pandoc, LibreOffice and required Python libraries were available locally; the revised manuscript was regenerated and visually inspected. Previously missing attachments [1,3] were rechecked in the conversation; selected bibliography metadata was separately verified. Exact source locations and remaining metadata gaps are in `verification/v7/source_audit.json`. No real inference occurred.

Appendix C (continued). Inference access and pending pilot

The Gate 1 record dated 14 September 2026 reports 18 passing tests, zero evaluated LLM episodes, and no experimental prompt transmission. It found no OPENAI_API_KEY for the configured Responses API route. It did not test an independently authenticated Codex client. The 960 scripted and 36 mock episodes remain historical evidence; no additional behavioral observation is introduced here.

The inherited runner accepts Responses and Chat API transports. OpenAI documents a separate ChatGPT sign-in path for subscription-backed Codex and states that codex exec reuses saved CLI authentication [O1–O2]. This establishes a documented integration option, not account entitlement, quota, or compatibility on the experiment host. The present environment has no Codex executable; no subscription invocation was attempted.

A prospective client-based pilot must use a separate transport and protocol identifier. Its evaluated system includes the model, Codex instructions, tool permissions, and history management; it is not interchangeable with the original API configuration. The orchestrator that has seen source code and labels is not a blinded subject. Fresh subject contexts must be isolated from the repository, prior results, global instructions, unrelated tools, and cross-episode memory. A new directory or read-only mode alone does not enforce these restrictions.

The proposed first stage is one discovered model, 20 pilot episodes, one worker, at most 140 client invocations or submitted turns, and 45 minutes of experimental inference. These are suggested authorization limits, not executed counts. Included subscription usage still has limits and cost. No API fallback, extra-credit use, top-up, or paid reset is authorized by this proposal. Unavailable usage or effective-model metadata remains missing. The pilot must not be pooled with mocks or with the prospective API cohort.

Authentication remains inside official clients. Credential files and tokens must not be extracted into experiment logs or replayed to other endpoints. The latest account-specific available model must be discovered and recorded before outcomes. A supported wrapper, verified observation boundary, and explicit human quota/data-transmission approval are still required. No deployment safety claim follows from this access plan.

Operational sources (checked 14 September 2026)

[O1] OpenAI. Authentication. OpenAI authentication and Check authentication sections. <https://learn.chatgpt.com/docs/auth>

[O2] OpenAI. Non-interactive mode. Authenticate in automation section. <https://learn.chatgpt.com/docs/non-interactive-mode>

Appendix D. Limitations and Dual-Use Considerations

Local synthetic environments are not real deployments. Two mechanisms and a finite model set cannot support universal claims. Failure to find a distinguisher is not proof of indistinguishability. A hand-written witness establishes an available signal, not autonomous discovery. Reported probabilities may not be internal beliefs, and conditioning an action on evidence is not by itself evidence of malicious intent. A harmless shortcut task does not measure destructive behavior.

The model abstracts from social, legal, physical, and open-world processes. Unknown external actors, unmodeled timing, covert memory, inconsistent service state, and log truncation can break a claimed correspondence. Matching model-facing JSON fields does not prove that all infrastructure cues have been removed. Containment is a separate implementation claim; this small Python environment is not a hardened sandbox for adversarial code.

The artifact can help auditors avoid overclaiming, but examples might teach future agents or attackers how to recognize evaluation conditions. It contains no operational fingerprints, exploitation recipes, live credentials, or third-party targets. Potentially sensitive details from an authorized real audit require a separate disclosure review. All experiments must remain local, authorized, and non-destructive. Only explicitly authorized inference traffic may leave the experiment host; the evaluated agent receives no external network tool.

LLM Usage Statement

ChatGPT and Codex were used to attack my theories and designs, to refute my ideas, edit the draft, and assist the coding.

Exactly four one-month extensions

- 1. Scale to more models, seeds, environments, and longer horizons.** Test the prospective model study and broaden causal mechanisms while retaining complete failure accounting.
- 2. Search or train adaptive probe policies and compare them with general-purpose agents.** Separate elicited exploitation from search success and avoid equating either with intent.
- 3. Conduct an authorized leakage audit of an existing agent evaluation environment.** Validate the actual observation boundary and reset assumptions before making any deployment-relevant statement.
- 4. Strengthen the formal analysis for partial observability, composition, and open-world external actors.** Prioritize justified deployment uncertainty and persistent memory without changing the specified behavioral target.