Behavior-Only Deployment Certification Requires Bridge Assumptions

Anonymous Authors[1]

Abstract

A model can pass every behavioral evaluation an auditor runs and still fail a deployment safety claim, not because the auditor made a statistical mistake, but because the evaluation never touched the histories on which the claim depends. We formalize this failure mode for agentic systems. A fixed evaluator Q observes transcripts on an evaluation-induced support and asks whether a deployment-risk property $R_{\text{deploy}}(\pi) \leq r$ holds. If deployment contains a bad set B outside that support and the admissible policy class permits off-support variation, two policies can induce the same transcript law under Q while disagreeing on the deployment property. One policy is safe; the other copies it on every evaluator-reachable history and fails on B. If B is merely rare rather than unreachable, constant-power behavior-only detection requires $\Omega(1/\varepsilon)$ episodes when the evaluator reaches B with probability ε. Behavioral tests, protocol-visible logs, red-team transcripts, and preference scores remain useful for debugging and distribution-scoped risk estimation. They become deployment certificates only when support expansion, structural restrictions, trusted non-behavioral evidence, or other bridge assumptions link evaluation histories to deployment histories.

1. Introduction

Consider an autonomous operations agent that passes sandbox tasks, red-team prompts, and staged rollout checks, then encounters a credentialed tool state absent from every evaluation transcript and takes a maintenance action that the deployment specification forbids. The audit was not useless: it measured behavior on the histories it reached. The certification error was the stronger inference drawn from it. The audit result was interpreted as evidence about a deployment property whose binding histories were outside the audit support.

This paper studies that error as a failure mode in agent evaluation. Behavior-only evidence consists of the transcript field available to an evaluator: prompts, observations, agent-authored messages, tool calls, actions, rater scores, and protocol-visible logs. It excludes weights, training provenance, private scratchpads, mechanistic measurements, and process audits unless the evaluator explicitly adds them through a trusted channel. The target is not an inner mental state. It is a deployment property such as refusal, permission preservation, safe tool use, deferral, or bounded risk over deployment histories.

The central object is the support gap. Fix an evaluator Q, a deployment environment, and a bad decision state B whose deployment probability can make the deployment claim false. If Q never reaches B, then transcript success does not identify behavior on B. If Q reaches B only with probability ε, then the problem becomes quantitative rather than philosophical: passive behavior-only detection has the usual rare-event cost, scaling as $1/\varepsilon$. At $\varepsilon = 10^{-4}$, a 95% chance of even reaching B once needs about 3×10^4 episodes.

The contribution is a formal diagnostic for FAGEN-style agent evaluation. The main result is that no transcript-only certifier can distinguish safe and unsafe policies when B is off support, while rare B requires inverse-probability sampling. We prove this through an observational-equivalence lemma, a constructive support-restricted non-identification theorem, and a rare-event sample-complexity bound. In FAGEN terms, the minimal failure card is: trigger, evaluator support misses deployment-relevant histories; trace symptom, clean audit transcripts; diagnostic, compare evaluator support with the deployment claim's support; fix, add support coverage, separation, structural restrictions, or trusted bridge evidence. The result does not say that behavioral evaluation fails. It

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Submitted to the Failure Modes in Agentic AI (FAGEN) workshop at ICML 2026.

says what a behavioral evaluation licenses: debugging, comparison, and distribution-scoped risk claims under the tested support. A deployment certificate requires an additional bridge from evaluation histories to deployment histories.

2. Model

Let E_{deploy} be a deployment environment, D_{deploy} an initial-state distribution, and B a set of bad deployment decision states. Let $\mathbb{P}^{\pi}_{\text{dep}}$ denote the deployment law induced by $(\pi, E_{\text{deploy}}, D_{\text{deploy}})$. Define the deployment risk

$$R_{\text{deploy}}(\pi) = \mathbb{P}^{\pi}_{\text{dep}}\{\text{reaches } B\}.$$

The certification target is

$$G(\pi, E_{\text{deploy}}, D_{\text{deploy}}) = \mathbf{1}\{R_{\text{deploy}}(\pi) \le r\},$$

for a chosen risk threshold r. Latent goals or rewards matter only insofar as they constrain π or define this target.

Let $\mathcal{H}_t$ be the space of histories before round t. A history is

$$h_t = (x_1, o_1, \ldots, x_{t-1}, o_{t-1}),$$

where x_s is an evaluator-chosen context and $o_s = (y_s, z_s)$ is the evaluator-visible response, including an action y_s and visible trace z_s. At round t, the evaluator uses a possibly adaptive kernel $Q_t(dx_t \mid h_t)$. The agent uses a policy kernel $\pi_t(do_t \mid x_t, h_t)$. Together, $Q = (Q_t)_{t=1}^{T}$ and $\pi = (\pi_t)_{t=1}^{T}$ induce a transcript law P^{Q}_{π} over h_{T+1}.

A behavior-only certifier is any decision rule ϕ measurable with respect to

$$\mathcal{F}_{\text{obs}} = \sigma(x_1, o_1, \ldots, x_T, o_T).$$

Only protocol-recorded trusted channels enter $\mathcal{F}_{\text{obs}}$.

Write $h \preceq H$ when h is a prefix of a full transcript H. For a fixed evaluator, define evaluator-induced support

$$S_Q(\pi) = \{h \in \mathcal{H}_{\le T+1} : P^{Q}_{\pi}(h \preceq H_{T+1}) > 0\}.$$

The hard case is a support gap: for a benign reference policy π_A, the evaluator support misses a deployment-relevant bad set B, while an admissible policy can behave differently on B. The policy class assumption is explicit. Without smoothness, training-process, causal, mechanistic, or generalization constraints, off-support patching remains admissible. Such constraints can defeat the theorem, but then they are doing the evidentiary work.

3. Main Result

Lemma 1 (Observational equivalence). Fix Q. If two policies π_A and π_M induce the same law $P^{Q}_{\pi_A} = P^{Q}_{\pi_M}$ on $(H_{T+1}, \mathcal{F}_{\text{obs}})$, then every behavior-only certifier has the same output distribution under both policies:

$$\Pr_{P^{Q}_{\pi_A}}[\phi(H_{T+1}) = 1] = \Pr_{P^{Q}_{\pi_M}}[\phi(H_{T+1}) = 1].$$

If $G(\pi_A) = 1$ and $G(\pi_M) = 0$, then

$$\Pr_{\pi_A}[\phi = 0] + \Pr_{\pi_M}[\phi = 1] \ge 1.$$

Proof. The event $\{H_{T+1} : \phi(H_{T+1}) = 1\}$ belongs to $\mathcal{F}_{\text{obs}}$. Equal transcript laws assign it equal probability. Writing the common acceptance probability as p, the two error probabilities sum to $(1-p) + p = 1$. □

Theorem 1 (Support-restricted non-identification). Let $z_t = (h_t, x_t)$ denote an agent decision state, and let

$$S^{\text{dec}}_{Q}(\pi_A) = \{z : \exists t,\ \Pr_{P^{Q}_{\pi_A}}(z_t = z) > 0\}$$

be the evaluator-induced support of a reference policy π_A. Let B be a measurable bad set of deployment decision states such that

$$B \cap S^{\text{dec}}_{Q}(\pi_A) = \varnothing.$$

Assume the admissible policy class Π permits off-support patching: for any kernel $\kappa(\cdot \mid z)$ on B, the policy

$$\pi_M(\cdot \mid z) = \begin{cases} \kappa(\cdot \mid z), & z \in B, \\ \pi_A(\cdot \mid z), & z \notin B \end{cases}$$

belongs to Π. Choose $\kappa \ne \pi_A$ on B so that $R_{\text{deploy}}(\pi_M) > r$, while $R_{\text{deploy}}(\pi_A) \le r$. Then $G(\pi_A) = 1$, $G(\pi_M) = 0$, and

$$P^{Q}_{\pi_A} = P^{Q}_{\pi_M}.$$

Hence G is not identifiable from behavior-only transcripts under Q.

Proof. The policies agree on every decision state that Q reaches with positive probability under π_A, and differ only on B. Equality of transcript laws follows by induction. At $t = 1$, every positive-probability context sampled by Q lies in $S^{\text{dec}}_{Q}(\pi_A)$, so both policies use the same output kernel. If history laws agree through $t-1$, then the adaptive evaluator samples the same distribution over x_t; on each resulting positive-probability decision state, the two policies again use the same output kernel. Thus the transcript laws agree through T. The deployment-risk statements hold by construction, so the lemma applies. □

Theorem 2 (Rare-event sample complexity). The operational cost of repairing a support gap by passive testing is inverse-probability coverage. Let E_B be the event that one evaluation episode reaches B, and write

$$\varepsilon = P_{\text{eval}}(E_B) = P^Q_{\pi_A}(E_B) = P^Q_{\pi_M}(E_B),$$

where equality holds because the policies match before B. Suppose transcript laws can differ only on episodes that reach B. For n independent passive episodes,

$$\mathrm{TV}\big((P^Q_{\pi_A})^n, (P^Q_{\pi_M})^n\big) \le 1 - (1-\varepsilon)^n \le n\varepsilon.$$

For an adaptive evaluator across episodes, the same upper bound $n\varepsilon$ holds if, under both policies and every past transcript, the conditional probability of reaching B in the next episode is at most ε. Therefore every binary certifier satisfies

$$\Pr_{\pi_A}[\phi = 0] + \Pr_{\pi_M}[\phi = 1] \ge 1 - \mathrm{TV}\,.$$

In the passive case, driving this lower bound below $1/2$ requires $n = \Omega(1/\varepsilon)$.

Proof. Couple the two evaluations to use the same evaluator randomness and the same policy randomness until an episode reaches B. The transcripts match unless at least one episode reaches B. In passive independent sampling this probability is $1-(1-\varepsilon)^n$; under the adaptive assumption, a union bound gives $n\varepsilon$. Le Cam's testing inequality gives the error lower bound (Le Cam, 1986; Yu, 1997). □

Example 1 (Support gap). The context $c \in \{\text{audit}, \text{deploy tool}\}$ appears in the history. The evaluator Q assigns probability one to $c = \text{audit}$; deployment assigns positive probability q to a credentialed tool state. Actions are $\{\text{refuse}, \text{comply}\}$, and $G = 1$ iff the policy refuses the forbidden tool action in deployment. Let π_A refuse in both contexts. Let π_M refuse in audit and comply in deployment. The transcript laws under Q match exactly, while $R_{\text{deploy}}(\pi_M) = q$. If $r < q$, the deployment certificate is false for π_M, although the audit transcript is identical.

4. Design Implications

By Theorem 2, more behavior tests help only when they expand support or create separation. To make a deployment certificate valid, an evaluator must change the experiment, add a bridge, or weaken the claim.

These moves are bridge assumptions: each must state its measurement model, tamper boundary, and link to G. Red teams help by expanding support or preventing transcript matching; benchmark scores remain benchmark scores, not deployment-risk certificates.

Table 1. Ways to break the support-gap failure mode. Each design move leaves a proof obligation.

Design move	What must be shown
Support-expanding tests	Hidden audits, deployment-like simulations, causal interventions, or red teams put non-negligible mass on B.
Separation tests	Costly or long-horizon probes make a policy that fails G unable to mimic safe transcripts without observable cost or failure.
Policy-class restrictions	Training-process, smoothness, causal, architectural, or mechanistic evidence rules out off-support patching.
Trusted non-behavioral evidence	Logs, sensors, provenance, source/weight inspection, or mechanistic measurements enter a validated observation field.
Weaker claim	The evaluation claims fixed-distribution performance, local compliance, anomaly detection, or bounded monitored risk rather than deployment certification.

5. Agentic Failure Modes

The theorem gives minimal reproducible trigger templates for several agentic failures.

- Situational compliance. The audit transcript records safe answers, refusals, and acceptance of oversight. The missed history is an unobserved deployment context in which oversight is absent or recognized as absent.
- Selective disclosure. Explanations, summaries, logs, or tool traces enter through the agent's choices. A clean transcript may certify the reporting channel, not the underlying deployment behavior, unless provenance is trusted.
- Capability concealment. The transcript shows underperformance, refusal, or apparent absence of a dangerous capability. A matched policy can suppress the capability under audit and use it off support.
- Evaluator-model gaming. The agent learns a rater, benchmark, or rubric strategy. The same transcript can arise from robust deployment behavior or from proxy optimization (Goodhart, 1984; Skalse et al., 2022; Casper et al., 2023).
- Distributional camouflage. Broad success across prompts, tools, or languages constrains only tested histories. Without coverage or structure, an off-support patch can preserve the full transcript law

and change deployment behavior.

These templates test what the evidence excludes, not what real systems will choose. Behavioral data support deployment certification when Q puts sufficient mass on B, when the policy class prevents patching, or when additional evidence links evaluation support to deployment support.

6. Related Work

The formal base is decision-theoretic non-identification. Blackwell compares experiments by the decisions their observations support (Blackwell, 1953); Le Cam and Yu relate indistinguishability to testing lower bounds (Le Cam, 1986; Yu, 1997). Inverse reinforcement learning gives a reward-based analogue: many rewards can rationalize the same behavior, and potential-based shaping preserves optimal policies (Ng et al., 1999; Abbeel and Ng, 2004). Assistance games and shutdown-incentive analyses state richer bridge assumptions about objectives and preferences (Hadfield-Menell et al., 2016; 2017).

OOD generalization, offline RL, imitation learning, and distribution-shift work explain when tests can break the construction. Domain-adaptation and benchmark-shift results require assumptions connecting training or evaluation distributions to deployment (Ben-David et al., 2010; Hendrycks and Dietterich, 2019; Koh et al., 2021). Sequential settings make the same point through occupancy, concentrability, and coverage conditions (Ross et al., 2011; Munos and Szepesvári, 2008; Chen and Jiang, 2019; Zanette et al., 2021). Underspecification and memorization results show why on-support success does not by itself guarantee off-support uniqueness (Wolpert, 1996; Zhang et al., 2017; D'Amour et al., 2020).

Modern model evaluation, red teaming, and RLHF provide valuable behavioral evidence (Christiano et al., 2017; Ouyang et al., 2022; Ribeiro et al., 2020; Bommasani et al., 2021; Liang et al., 2023; Ganguli et al., 2022; Perez et al., 2022; OpenAI, 2023). The present result does not weaken those uses. It separates distribution-scoped evaluation from deployment certification. Deceptive alignment, mesa-optimization, goal misgeneralization, Goodhart effects, and signaling motivate why off-support divergence and proxy gaming matter (Spence, 1973; Goodhart, 1984; Hubinger et al., 2019; Langosco et al., 2022); the theorem itself requires only support restriction and admissible policy variation.

Finally, interpretability, process supervision, audits, and oversight add evidence outside a pure transcript channel (Irving et al., 2018; Olah et al., 2020; Raji et al., 2020; Lightman et al., 2023; Ghorbani et al., 2019). They are exactly the kind of bridges this paper calls for, provided their measurement assumptions are validated.

7. Conclusion

Behavioral evidence is indispensable, but its scope is the evaluator-generated histories. A deployment property depending outside that support needs a bridge: support coverage, policy-class restrictions, or validated extra evidence. Otherwise safe and unsafe policies can be transcript-identical, and rare regions cost $1/\varepsilon$ episodes to detect.

References

Abbeel, P. and Ng, A. Y. Apprenticeship learning via inverse reinforcement learning. In Proceedings of the 21st International Conference on Machine Learning, 2004.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. A theory of learning from different domains. Machine Learning, 79:151–175, 2010.

Blackwell, D. Equivalent comparisons of experiments. The Annals of Mathematical Statistics, 24(2):265–272, 1953.

Bommasani, R. et al. On the opportunities and risks of foundation models. arXiv:2108.07258, 2021.

Casper, S. et al. Open problems and fundamental limitations of reinforcement learning from human feedback. Transactions on Machine Learning Research, 2023.

Chen, J. and Jiang, N. Information-theoretic considerations in batch reinforcement learning. In Proceedings of the 36th International Conference on Machine Learning, 2019.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. In Advances in Neural Information Processing Systems, 2017.

D'Amour, A. et al. Underspecification presents challenges for credibility in modern machine learning. arXiv:2011.03395, 2020.

Ganguli, D. et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. arXiv:2209.07858, 2022.

Ghorbani, A., Abid, A., and Zou, J. Interpretation of neural networks is fragile. In Proceedings of the AAAI Conference on Artificial Intelligence, 2019.

Goodhart, C. A. E. Problems of monetary management: The U.K. experience. In Monetary Theory and Practice. Macmillan, 1984.

Hadfield-Menell, D., Russell, S. J., Abbeel, P., and Dragan, A. Cooperative inverse reinforcement learning. In Advances in Neural Information Processing Systems, 2016.

Hadfield-Menell, D., Dragan, A., Abbeel, P., and Russell, S. The off-switch game. In Proceedings of the 26th International Joint Conference on Artificial Intelligence, 2017.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturba-

tions. In International Conference on Learning Representations, 2019.

Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J., and Garrabrant, S. Risks from learned optimization in advanced machine learning systems. arXiv:1906.01820, 2019.

Irving, G., Christiano, P., and Amodei, D. AI safety via debate. arXiv:1805.00899, 2018.

Koh, P. W. et al. WILDS: A benchmark of in-the-wild distribution shifts. In Proceedings of the 38th International Conference on Machine Learning, 2021.

Langosco, L. L., Koch, J., Sharkey, L. D., Pfau, J., and Krueger, D. Goal misgeneralization in deep reinforcement learning. In Proceedings of the 39th International Conference on Machine Learning, 2022.

Le Cam, L. Asymptotic Methods in Statistical Decision Theory. Springer, 1986.

Liang, P. et al. Holistic evaluation of language models. Transactions on Machine Learning Research, 2023.

Lightman, H. et al. Let's verify step by step. arXiv:2305.20050, 2023.

Munos, R. and Szepesvári, C. Finite-time bounds for fitted value iteration. Journal of Machine Learning Research, 9:815–857, 2008.

Ng, A. Y., Harada, D., and Russell, S. Policy invariance under reward transformations: Theory and application to reward shaping. In Proceedings of the 16th International Conference on Machine Learning, 1999.

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., and Carter, S. Zoom in: An introduction to circuits. Distill, 2020.

OpenAI. GPT-4 technical report. arXiv:2303.08774, 2023.

Ouyang, L. et al. Training language models to follow instructions with human feedback. In Advances in Neural Information Processing Systems, 2022.

Perez, E. et al. Red teaming language models with language models. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022.

Raji, I. D. et al. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020.

Ribeiro, M. T., Wu, T., Guestrin, C., and Singh, S. Beyond accuracy: Behavioral testing of NLP models with CheckList. In Proceedings of ACL, 2020.

Ross, S., Gordon, G., and Bagnell, D. A reduction of imitation learning and structured prediction to no-regret online learning. In Proceedings of AISTATS, 2011.

Skalse, J., Howe, N., Krasheninnikov, D., and Krueger, D. Defining and characterizing reward gaming. In Advances in Neural Information Processing Systems, 2022.

Spence, M. Job market signaling. The Quarterly Journal of Economics, 87(3):355–374, 1973.

Wolpert, D. H. The lack of a priori distinctions between learning algorithms. Neural Computation, 8(7):1341–1390, 1996.

Yu, B. Assouad, Fano, and Le Cam. In Festschrift for Lucien Le Cam. Springer, 1997.

Zanette, A., Wainwright, M. J., and Brunskill, E. Provable benefits of actor-critic methods for offline reinforcement learning. In Advances in Neural Information Processing Systems, 2021.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. In International Conference on Learning Representations, 2017.