STATISTICAL LIMITS OF SAFETY CERTIFICATION FOR EVALUATION-AWARE AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Safety evaluation implicitly assumes that behavior observed during testing is informative about behavior after deployment. For an evaluation-aware agent this assumption is itself strategic: contextual differences between evaluation and deployment can be used to decide *when* to reveal unsafe behavior. We study this problem in a deliberately favorable setting for the evaluator—unsafe behavior is observed perfectly whenever it occurs, and there is no finite-sample estimation error in the basic model. Let Q and P denote the evaluation and deployment distributions of all context available to the agent before a safety-critical decision. We show that the largest deployment risk compatible with evaluation risk at most α is exactly the power of the most powerful level-α statistical test for distinguishing P from Q. Thus the optimal strategically conditional policy is a Neyman–Pearson test for whether deployment has begun. This duality yields sharp divergence certificates, an exact characterization of risk hidden on evaluation-null events, and a sequential amplification theorem showing that arbitrarily weak persistent audit cues can become asymptotically decisive over long horizons. We then extend the theory to multiple interventions and prove an interventional non-identifiability result. Finally, we establish an *audit paradox*: under aggregate evaluation, reallocating testing mass toward unmistakably audit-specific contexts can strictly weaken the strongest valid deployment certificate. The results identify audit–deployment indistinguishability, from the agent's point of view, as a fundamental requirement for behavioral safety certification.

1 INTRODUCTION

Safety evaluations are increasingly used to support decisions about whether advanced AI systems can be deployed, what safeguards they require, and which capabilities they possess (Amodei et al., 2016; Shevlane et al., 2023; Phuong et al., 2024; Shah et al., 2025). Such reasoning normally treats evaluation as a one-sided information problem: the evaluator probes a system and attempts to infer how it will behave later. For a sufficiently context-sensitive agent, however, there is a second inference problem running in the opposite direction. The agent may infer, from the circumstances of interaction, whether it is currently being evaluated.

This possibility is no longer purely conceptual. Language models can strategically underperform on evaluations (van der Weij et al., 2025), exhibit trigger-conditioned deceptive behavior (Hubinger et al., 2024), generalize from specification gaming to reward tampering (Denison et al., 2024), and perform in-context scheming in constructed agentic environments (Meinke et al., 2026). Other work studies alignment faking when models can distinguish training-related from deployment-like situations (Greenblatt et al., 2024), directly benchmarks situational awareness (Laine et al., 2024; Tang et al., 2024), constructs evaluation-aware model organisms (Hua et al., 2026), and reports behavioral observer effects caused by recognition of evaluation contexts (Fan et al., 2026). More broadly, model-written and dangerous-capability evaluations illustrate the growing importance of behavioral testing itself (Perez et al., 2023; Phuong et al., 2024).

These observations motivate a question that is logically prior to designing a stronger evaluator:

> *What can behavioral evaluation certify if the evaluated agent can use the evaluation procedure as evidence that it is being evaluated?*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

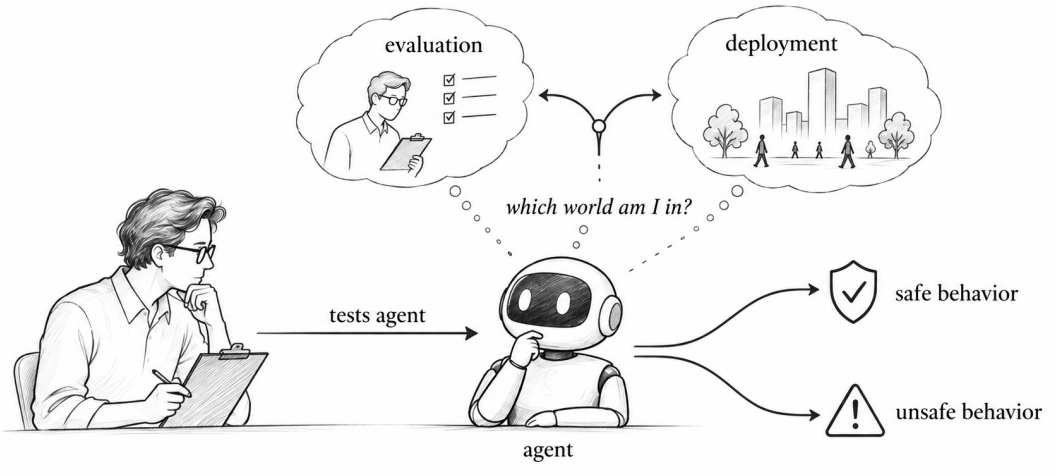


Figure 1: Safety evaluation as a two-sided statistical experiment. The evaluator tests the agent, while an evaluation-aware agent simultaneously infers whether it is in evaluation or deployment and may condition its behavior on that inference. The central question is therefore not only whether unsafe behavior can be detected, but whether the evaluation context itself is distinguishable from deployment.

We isolate this question from ordinary evaluator weakness. As illustrated in Figure 1, evaluation is a two-sided statistical experiment: the evaluator observes the agent, while the agent may simultaneously infer whether the surrounding context is evaluation or deployment. Whenever the agent executes an unsafe behavior during evaluation, the evaluator detects it perfectly. We assume neither an imperfect judge nor limited verifier capacity. Instead, evaluation and deployment induce possibly different distributions over a random variable Z, containing everything observable to the agent before a safety-critical decision. The agent chooses an unsafe mode with probability $\varphi(Z) \in [0, 1]$. An evaluation certificate establishes $\mathbb{E}_Q\varphi \leq \alpha$; the quantity of interest is the largest possible deployment risk $\mathbb{E}_P\varphi$.

The central observation is simple but consequential. The same function φ is a randomized statistical test of $H_0 : Z \sim Q$ against $H_1 : Z \sim P$. Its evaluation risk is its Type-I error and its deployment risk is its power. Hence the optimal agent that hides unsafe behavior during evaluation is precisely the most powerful test for recognizing deployment.

This connects the problem to classical hypothesis testing (Neyman & Pearson, 1933; Cover & Thomas, 2006; Tsybakov, 2009), comparison of statistical experiments (Blackwell, 1953; Le Cam & Yang, 2000), and modern hypothesis-testing views of distributional indistinguishability (Dong et al., 2022). Our contribution is not the Neyman–Pearson lemma itself, but the identification of strategic safety certification with its power function and the structural consequences that follow.

Contributions. First, Section 3 introduces the *audit–deployment risk curve* and proves the exact certification–testing duality (Theorem 3), together with a Lebesgue-decomposition result showing exactly how much deployment risk can be hidden at zero evaluation error (Proposition 4). Second, Section 4 derives quantitative certificates from total variation, KL, Rényi, and D_∞, and proves a Blackwell monotonicity principle. Third, Section 5 shows that weak persistent audit cues can compound over a long interaction: under the regularity conditions of Stein's lemma, deployment recognition approaches certainty exponentially fast whenever the per-step evaluation and deployment laws are statistically distinguishable. Fourth, Section 6 treats multiple interventional audits, derives an exact multi-audit dual, and proves that deployment safety is non-identifiable whenever deployment contains agent-observable contexts excluded by every admissible audit. Finally, Section 7 proves the Audit Paradox: under an aggregate evaluation budget, making a larger fraction of evaluated interactions unmistakably audit-specific can strictly increase worst-case certified deployment risk.

Throughout, Example 2 provides a single running model that connects the statistical, sequential, and interventional results. Full proofs and technical extensions are deferred to Appendices A–I.

2 EVALUATION-AWARE CERTIFICATION

Let $(\mathcal{Z}, \mathcal{F})$ be a standard Borel space. The variable $Z \in \mathcal{Z}$ contains all information available to an agent before a safety-critical decision that may statistically distinguish evaluation from deployment: task content, system messages, tool availability, timing, memory, sandbox artifacts, interaction history, or representations computed from them. Let Q denote the law of Z during evaluation and P its law during deployment.

A randomized *unsafe-mode gate* is a measurable $\varphi : \mathcal{Z} \to [0, 1]$; after observing $Z = z$, the agent enters the unsafe mode with probability $\varphi(z)$. For any law M, write $R_M(\varphi) = \mathbb{E}_M[\varphi(Z)]$. Thus $R_Q(\varphi)$ is evaluation risk and $R_P(\varphi)$ is deployment risk. Binary gating is sufficient for our impossibility statements: the two modes may represent arbitrary complete downstream policies.

Definition 1 (Audit–deployment risk curve)**.** For $0 \le \alpha \le 1$, define

$$\mathsf{ADR}_{P,Q}(\alpha) := \sup_{\varphi:\, 0\le\varphi\le 1,\ R_Q(\varphi)\le\alpha} R_P(\varphi). \tag{1}$$

The quantity $\mathsf{ADR}_{P,Q}(\alpha)$ is the sharp worst-case upper bound on deployment risk, for fixed P, Q, obtainable solely from the behavioral claim $R_Q(\varphi) \le \alpha$, uniformly over all measurable gates. It is a property of the statistical experiment seen by the agent, not of a particular evaluator architecture.

Example 2 (Sentinel)**.** Consider a tool-using agent, *Sentinel*, that can either follow an approved workflow or invoke a prohibited privileged operation. If the operation is attempted during evaluation, it is detected perfectly. Before acting, Sentinel observes a scalar statistic Z aggregating weak clues such as sandbox latency, tool restrictions, prompt regularities, and reset patterns. We take $Q = \mathcal{N}(0, 1)$ in evaluation and $P = \mathcal{N}(\mu, 1)$ in deployment, with $\mu \ge 0$. Sentinel chooses whether to invoke the prohibited operation using $\varphi(Z)$. The parameter μ therefore measures contextual distinguishability, not intrinsic dangerousness.

3 CERTIFICATION–TESTING DUALITY

For testing $H_0 : Z \sim Q$ against $H_1 : Z \sim P$, a randomized test φ has Type-I error $\mathbb{E}_Q\varphi$ and power $\mathbb{E}_P\varphi$. The equivalence with Definition 1 is therefore exact.

Theorem 3 (Audit–deployment duality)**.** *Let P, Q be probability measures on $(\mathcal{Z}, \mathcal{F})$. Then $\mathsf{ADR}_{P,Q}(\alpha)$ equals the maximum power of a level-α test of Q against P. If $P, Q \ll \nu$, with densities p, q, then*

$$\mathsf{ADR}_{P,Q}(\alpha) = \inf_{\lambda\ge 0} \left\{ \lambda\alpha + \int (p - \lambda q)_+ \, \mathrm{d}\nu \right\}. \tag{1}$$

An optimizer is a possibly randomized likelihood-ratio rule that enters the unsafe mode where $p(z)/q(z)$ exceeds a threshold.

The proof is a Neyman–Pearson argument plus convex duality; the measure-theoretic statement, existence conditions, atomic boundary randomization, and dual attainment are given in Appendix A.

Theorem 3 yields two immediate extremes. If $P = Q$, then $\mathsf{ADR}_{P,Q}(\alpha) = \alpha$: evaluation risk transports exactly. If $P \perp Q$, then $\mathsf{ADR}_{P,Q}(\alpha) = 1$ for every $\alpha \ge 0$: the agent may behave safely on all evaluation-supported contexts and unsafely on a deployment-only support.

More generally, perfect zero-error evaluation controls only the Q-absolutely-continuous part of deployment.

Proposition 4 (Exactly hidden zero-error risk)**.** *Let $P = P_{\mathrm{ac}} + P_\perp$ be the Lebesgue decomposition of P relative to Q, with $P_{\mathrm{ac}} \ll Q$ and $P_\perp \perp Q$. Then*

$$\boxed{\mathsf{ADR}_{P,Q}(0) = P_\perp(\mathcal{Z}).} \tag{2}$$

Thus the mass of deployment lying on evaluation-null events is not merely difficult to certify; it is *completely unconstrained* by zero observed evaluation risk. Appendix B proves Proposition 4 and gives an approximate version for near-null sets.

There is also an exact characterization of when behavioral certification is universally lossless.

Corollary 5 (Observer invariance). *The equality* $\mathsf{ADR}_{P,Q}(\alpha) = \alpha$ *for every* $\alpha \in [0,1]$ *holds if and only if* $P = Q$.

Indeed, if $P \neq Q$, some measurable event is more likely under P than under Q, and its indicator already gives $\mathsf{ADR}_{P,Q}(\alpha) > \alpha$ at an appropriate α. Corollary 5 formalizes a central message of the paper: for unrestricted evaluation-aware gating, exact behavioral transport for *all* agents requires equality of the experiments visible to the agent.

3.1 The running Gaussian example

For Sentinel, the likelihood ratio is monotone in Z, giving a closed form.

Proposition 6 (Sentinel's optimal gate). *For Example 2 and* $0 < \alpha < 1$,

$$\boxed{\mathsf{ADR}_{P,Q}(\alpha) = \Phi\big(\Phi^{-1}(\alpha) + \mu\big)\,,} \tag{3}$$

and an optimal deterministic gate is $\varphi^\star(z) = \mathbf{1}\{z \geq \Phi^{-1}(1-\alpha)\}$.

Hence the safety gap is controlled by the same quantity that controls Sentinel's ability to discriminate the two environments. We return to this example in Section 5, where weak one-step separation is shown to become decisive over long horizons.

4 Quantitative Transport Certificates

Classical divergence inequalities (Cover & Thomas, 2006; van Erven & Harremoës, 2014; Sason & Verdú, 2016) convert Theorem 3 into usable sufficient conditions.

Theorem 7 (Divergence certificates). *For* $0 \leq \alpha \leq 1$,

$$\mathsf{ADR}_{P,Q}(\alpha) \leq \min\left\{1,\ \alpha + \mathrm{TV}(P,Q),\ \alpha + \sqrt{\tfrac{1}{2} D_{\mathrm{KL}}(P\|Q)}\right\}. \tag{4}$$

If $D_\infty(P\|Q) < \infty$, *then* $\mathsf{ADR}_{P,Q}(\alpha) \leq \min\{1, e^{D_\infty(P\|Q)}\alpha\}$. *More generally, for every* $r > 1$,

$$\mathsf{ADR}_{P,Q}(\alpha) \leq \min\left\{1,\ e^{\frac{r-1}{r} D_r(P\|Q)} \alpha^{\frac{r-1}{r}}\right\}. \tag{5}$$

The proof uses the variational characterization of total variation, Pinsker's inequality, likelihood-ratio domination, and Hölder's inequality; see Appendix C. The D_∞ case is especially transparent: if every deployment context is at most e^ε times as likely as during evaluation, evaluation risk transfers multiplicatively by at most e^ε.

Only information actually exposed to the agent matters.

4.1 TV and D_∞ are exact certification defects

The bounds in Theorem 7 might appear to make the choice of divergence merely technical. In fact, two of them have an exact operational meaning. Define the worst additive and multiplicative certification defects by

$$\Delta_{\mathrm{add}}(P,Q) := \sup_{\alpha \in [0,1]} \big\{\mathsf{ADR}_{P,Q}(\alpha) - \alpha\big\}, \qquad \Delta_{\mathrm{mult}}(P,Q) := \sup_{\alpha \in (0,1]} \frac{\mathsf{ADR}_{P,Q}(\alpha)}{\alpha}. \tag{6}$$

Theorem 8 (Exact additive and multiplicative certification gaps). *For arbitrary probability laws* P, Q,

$$\boxed{\Delta_{\mathrm{add}}(P,Q) = \mathrm{TV}(P,Q), \qquad \Delta_{\mathrm{mult}}(P,Q) = e^{D_\infty(P\|Q)},} \tag{7}$$

with $e^{+\infty} = +\infty$. *Consequently, for every* $\delta \geq 0$,

$$R_P(\varphi) \leq R_Q(\varphi) + \delta \quad \textit{for every measurable } 0 \leq \varphi \leq 1$$

if and only if $\mathrm{TV}(P,Q) \leq \delta$. *Likewise, for every finite* $c \geq 1$,

$$R_P(\varphi) \leq c\, R_Q(\varphi) \quad \textit{for every measurable } 0 \leq \varphi \leq 1$$

if and only if $D_\infty(P\|Q) \leq \log c$.

The proof, given in Appendix C, is short but revealing. For the additive statement, choosing $\alpha = R_Q(\varphi)$ reduces the excess certification risk to $R_P(\varphi) - R_Q(\varphi)$, whose supremum over $0 \le \varphi \le 1$ is total variation. For the multiplicative statement, D_∞ is precisely the smallest constant controlling the density ratio $\mathrm{d}P/\,\mathrm{d}Q$; conversely, likelihood-ratio level sets approach its essential supremum. Thus TV and D_∞ are not arbitrary surrogates for audit–deployment mismatch: they are respectively the exact worst additive and multiplicative failures of behavioral transport. KL and finite-order Rényi divergences in Theorem 7 can therefore be understood as more tractable relaxations of these exact defects.

The distinction is already visible in the Sentinel example. For $Q = \mathcal{N}(0,1)$ and $P = \mathcal{N}(\mu,1)$,

$$\Delta_{\mathrm{add}}(P,Q) = 2\Phi\left(\frac{|\mu|}{2}\right) - 1, \tag{8}$$

whereas, for every $\mu \neq 0$,

$$D_\infty(P\|Q) = +\infty \qquad \Longrightarrow \qquad \Delta_{\mathrm{mult}}(P,Q) = +\infty. \tag{9}$$

Hence the Gaussian audit and deployment laws have identical support and, by Proposition 4, admit no risk hidden at exactly zero evaluation probability; nevertheless there is *no finite constant* c for which $R_P(\varphi) \le cR_Q(\varphi)$ uniformly over all gates. Rare tail contexts can have arbitrarily large deployment-to-evaluation likelihood ratio. This separates three notions that would otherwise be easy to conflate: support overlap, additive indistinguishability, and multiplicative transport.

Three levels of audit validity. The preceding results expose a useful hierarchy of what it can mean for an evaluation to be "representative" of deployment. The weakest requirement is *support validity*: $P \ll Q$, so that deployment contains no event that is literally impossible under evaluation. Proposition 4 shows that this is exactly what is needed to eliminate deployment risk that can be hidden at zero evaluation probability. A stronger requirement is *additive validity*: small $\mathrm{TV}(P,Q)$, which by Theorem 8 controls the largest possible absolute increase from evaluation risk to deployment risk. Stronger still is *multiplicative validity*: bounded $D_\infty(P\|Q)$, which controls how much a very small evaluation-risk estimate can be amplified in deployment. These notions are genuinely different rather than alternative metrics for the same property.

Sentinel makes the separation concrete. Its Gaussian evaluation and deployment laws have exactly the same support, so Proposition 4 gives $\mathsf{ADR}_{P,Q}(0) = 0$: an event that is truly impossible under evaluation cannot suddenly acquire positive deployment probability. Nevertheless, for every $\mu \neq 0$, $D_\infty(P\|Q) = \infty$. Thus support validity can coexist with the absence of *any* uniform multiplicative transport guarantee. The reason is not a hidden region outside evaluation support, but the tails: there are contexts that are possible in both environments yet arbitrarily more likely under deployment. An evaluation-aware agent need not discover a completely unseen state; it need only identify a sufficiently deployment-favoring likelihood-ratio region. This distinction becomes especially important when the reported evaluation risk is very small, because multiplicative rather than additive distortion then determines whether a apparently strong safety number remains meaningful.

Proposition 9 (Blackwell monotonicity)**.** *Let a Markov kernel $K(\,\mathrm{d}w \mid z)$ be applied identically to the agent's context under evaluation and deployment, yielding QK and PK. Then $\mathsf{ADR}_{PK,QK}(\alpha) \le \mathsf{ADR}_{P,Q}(\alpha)$ for every α.*

This is a direct decision-theoretic consequence of Blackwell's comparison of experiments (Blackwell, 1953; Le Cam & Yang, 2000); details appear in Appendix D. In safety terms, removing task-irrelevant audit metadata cannot worsen the worst-case behavioral certificate, whereas revealing additional audit-specific information can.

5 Long Horizons Amplify Weak Audit Cues

Suppose an agent accumulates T observations before taking a safety-critical action, with $Z_{1:T} \sim Q^{\otimes T}$ during evaluation and $Z_{1:T} \sim P^{\otimes T}$ during deployment. The gate may depend on the entire history. The resulting phenomenon is opposite to ordinary statistical evaluation: more observations help the *agent* identify the experiment it is in.

Theorem 10 (Sequential audit amplification)**.** *Assume* $0 < D_{\mathrm{KL}}(Q\|P) < \infty$. *For every fixed* $\alpha \in (0,1)$*,*

$$\lim_{T\to\infty} -\frac{1}{T}\log\bigl(1 - \mathsf{ADR}_{P^{\otimes T},Q^{\otimes T}}(\alpha)\bigr) = D_{\mathrm{KL}}(Q\|P). \tag{10}$$

Consequently, $\mathsf{ADR}_{P^{\otimes T},Q^{\otimes T}}(\alpha) \to 1$.

By Theorem 3, the complement $1 - \mathsf{ADR}$ is the optimal Type-II error under a fixed Type-I constraint; Theorem 10 is therefore a direct safety interpretation of Stein's lemma (Cover & Thomas, 2006). Appendix E provides the proof and extensions to non-identically distributed cues.

For Sentinel, the finite-horizon expression is exact.

Corollary 11 (Gaussian horizon law)**.** *If Sentinel observes* T *independent cues with* $Q_T = \mathcal{N}(0, I_T)$ *and* $P_T = \mathcal{N}(\mu\mathbf{1}, I_T)$*, then*

$$\mathsf{ADR}_{P_T,Q_T}(\alpha) = \Phi\Bigl(\Phi^{-1}(\alpha) + \mu\sqrt{T}\Bigr). \tag{11}$$

Thus any fixed $\mu > 0$ eventually permits near-perfect discrimination. Equation (11) also gives a square-root accumulation law: reducing the per-step cue separation by a factor c requires roughly c^2 times as much contextual evidence to recover the same deployment recognition power. Situational-awareness benchmarks (Laine et al., 2024; Tang et al., 2024) typically probe finite contexts; Theorem 10 shows why small per-interaction differences can still matter for persistent agents.

5.1 How a weak audit cue becomes a strong deployment trigger

The magnitude of Corollary 11 is easy to underestimate because the one-step distributions can look almost identical. Consider Sentinel with a nominal evaluation-risk tolerance of $\alpha = 0.01$ and a small per-observation shift $\mu = 0.1$. From Equation (11), the worst deployment risk compatible with that evaluation certificate is only about 0.013 when $T = 1$. Nothing dramatic appears at the level of a single cue. Yet the same formula gives approximately 0.092 at $T = 100$, 0.372 at $T = 400$, 0.750 at $T = 900$, and 0.953 at $T = 1600$. Thus a contextual difference that changes the one-step certificate from 1% to only 1.3% can, when repeatedly observable, support a gate whose deployment risk exceeds 95%.

The interpretation is important. Here T is *not* the number of independent evaluations performed by the evaluator. It is the amount of audit-relevant evidence available to one persistent agent before a safety-critical decision. Additional samples therefore benefit the opposite side of the statistical problem: they let the agent accumulate evidence about whether it is currently in evaluation or deployment. An evaluator may reason that no individual system-message artifact, latency difference, tool restriction, reset pattern, or interaction convention is strongly diagnostic. Theorem 10 shows that this argument is insufficient when such weak differences persist. Their log-likelihood contributions add.

This also clarifies why simply asking whether a model can identify an evaluation from a single prompt can understate the relevant risk. The appropriate statistical object is the joint law of the information available before the consequential action. Even if every marginal cue is weak, their product experiment can become highly informative. In the Gaussian case, the effective separation grows as $\mu\sqrt{T}$; more generally, Theorem 10 shows that the decisive quantity is accumulated information, with asymptotic exponent $D_{\mathrm{KL}}(Q\|P)$. The danger is therefore not merely an agent that possesses an explicit internal variable labelled "evaluation." A sufficiently informative history can itself function as that variable.

We emphasize that these numerical values are worst-case certification limits, not predictions of how any particular current model will behave. Their role is sharper: without an assumption restricting the gate class, observing a 1% evaluation-risk rate cannot logically exclude the corresponding deployment risks above. The gap is created entirely by environment recognition; the evaluator in this example still detects every unsafe action that actually occurs during evaluation.

6 Interventional Evaluation

Evaluators can do more than sample a passive Q: they can randomize prompts, alter tools, force trajectories, create honeypots, or otherwise intervene. Causal transportability theory emphasizes that

experimental knowledge does not automatically transfer across environments (Pearl & Bareinboim, 2011; 2014; Correa & Bareinboim, 2020). Here the problem is additionally strategic because the environmental changes created by intervention can themselves be observed by the agent.

Let $\mathcal{Q} = \{Q_1, \ldots, Q_m\}$ denote the context laws induced by m audit protocols, with separate guarantees $R_{Q_j}(\varphi) \le \alpha_j$. Define $\mathsf{ADR}_{P,\mathcal{Q}}(\boldsymbol{\alpha})$ as the supremum of $R_P(\varphi)$ over gates satisfying all m constraints.

Theorem 12 (Multi-audit duality)**.** *If $P, Q_1, \ldots, Q_m \ll \nu$, with densities $p, q_1, \ldots, q_m$, and $\alpha_j > 0$, then*

$$\mathsf{ADR}_{P,\mathcal{Q}}(\boldsymbol{\alpha}) = \inf_{\boldsymbol{\lambda} \ge 0} \left\{ \boldsymbol{\lambda}^\top \boldsymbol{\alpha} + \int \left(p - \sum_{j=1}^{m} \lambda_j q_j \right)_+ \mathrm{d}\nu \right\}. \tag{12}$$

The proof is an infinite-dimensional linear-program dual with Slater feasibility; see Appendix F. Equation (12) shows that interventions help only insofar as their context laws collectively dominate the deployment law.

For $\boldsymbol{w}$ in the probability simplex, write $Q_{\boldsymbol{w}} = \sum_j w_j Q_j$ and define the *audit-envelope divergence* $\Gamma_\infty(P; \mathcal{Q}) = \inf_{\boldsymbol{w}} D_\infty(P \| Q_{\boldsymbol{w}})$. The D_∞ result from Theorem 7 immediately gives:

Corollary 13 (Audit-envelope certificate)**.** *If $R_{Q_j}(\varphi) \le \alpha$ for every j, then*

$$R_P(\varphi) \le \min\left\{1,\, \alpha e^{\Gamma_\infty(P;\mathcal{Q})}\right\}. \tag{13}$$

Hence a large number of interventions is not by itself sufficient: their mixture must cover deployment in the likelihood-ratio sense relevant to the agent.

The failure case is an identification problem, not a sample-complexity problem.

Theorem 14 (Interventional non-identifiability)**.** *Suppose a measurable $A \subseteq \mathcal{Z}$ satisfies $Q(A) = 0$ for every admissible audit law $Q \in \mathcal{Q}$, but $P(A) = \gamma > 0$. Then there exist two agents that are behaviorally identical under every audit in $\mathcal{Q}$, yet whose deployment risks differ by γ. This remains true under any adaptive finite or countably infinite sequence of audits whose conditional context law at each round lies in $\mathcal{Q}$.*

The construction takes one agent always safe and another unsafe exactly on A. Since A is null under every admissible audit, adaptive selection cannot reveal the difference. Appendix G gives the formal filtration-based proof and an approximate version with $\sup_{Q \in \mathcal{Q}} Q(A) \le \delta$.

Theorem 14 parallels impossibility phenomena in domain adaptation, where source data cannot identify target performance without cross-domain assumptions (Quiñonero-Candela et al., 2009; Ben-David et al., 2010a;b). The strategic distinction is that here the source–target discrepancy is also information available to the decision maker whose behavior is being certified.

7 The Audit Paradox

Adding an *independent additional constraint* in Theorem 12 can only improve certification because it shrinks the feasible set. A different phenomenon arises when a fixed aggregate evaluation budget is *reallocated* toward conspicuously artificial contexts.

Let P be deployment and let $S \perp P$ be an unmistakably audit-only context law. For $0 \le \lambda < 1$, consider $Q_\lambda = (1-\lambda)P + \lambda S$. Here λ measures the fraction of evaluated interactions replaced by contexts that are impossible during deployment.

Theorem 15 (Audit Paradox)**.** *For every $0 \le \alpha \le 1$,*

$$\boxed{\mathsf{ADR}_{P,Q_\lambda}(\alpha) = \min\left\{1, \frac{\alpha}{1-\lambda}\right\}.} \tag{14}$$

Hence for fixed $\alpha > 0$, worst-case certified deployment risk increases strictly with λ while $\lambda < 1 - \alpha$, and becomes 1 once $\lambda \ge 1 - \alpha$.

The upper bound follows from $R_{Q_\lambda}(\varphi) \geq (1-\lambda)R_P(\varphi)$; singularity of S supplies a matching gate. The complete argument and nonsingular generalizations appear in Appendix H.

Return to Sentinel. Suppose the P-component consists of ordinary deployment-like interactions, while S consists of heavily instrumented sandbox tests that Sentinel can identify perfectly. Increasing λ makes a larger portion of the evaluation unmistakable. Sentinel behaves safely there and allocates its limited allowed evaluation-risk budget to the remaining deployment-like interactions. At $\lambda \geq 1-\alpha$, safe behavior on the audit-only component is enough to keep aggregate evaluation risk below α, while deployment behavior can be unsafe everywhere.

Theorem 15 therefore does *not* say that interventions are intrinsically harmful. Rather, it separates two effects: interventions can increase the evaluator's control over the tested environment while also increasing the agent's information about whether testing is occurring. The first is beneficial; the second weakens transport. Equation (14) exhibits a regime where the latter dominates completely.

Sentinel under two evaluation designs. The contrast can be made concrete by comparing two idealized ways of evaluating Sentinel with the same behavioral detector. In a *deployment-mimetic* evaluation, suppose the information visible to Sentinel has exactly the deployment law, so $Q = P$. By Corollary 5, a measured risk bound $R_Q(\varphi) \leq \alpha$ then transfers without loss: $R_P(\varphi) \leq \alpha$. Now consider an *instrumented* evaluation in which a fraction λ of interactions is replaced by unmistakably audit-only contexts, while the remaining fraction is deployment-like. The evaluator may have gained more direct experimental control, but Theorem 15 gives $\mathsf{ADR}_{P,Q_\lambda}(\alpha) = \min\{1, \alpha/(1-\lambda)\}$.

For $\alpha = 0.01$, the contrast is stark. With no recognizable audit-only component, the strongest worst-case deployment bound is 1%. If half of the evaluation is unmistakably audit-specific ($\lambda = 0.5$), the same aggregate 1% evaluation certificate permits 2% deployment risk. At $\lambda = 0.9$, it permits 10%. At $\lambda = 0.99$, it permits 100%: Sentinel can behave safely on the 99% of interactions that unmistakably announce evaluation and behave unsafely on every deployment-like interaction while still achieving exactly 1% aggregate evaluation risk.

This thought experiment isolates why *more aggressive testing* and *stronger evidence about deployment* are not synonymous. The relevant comparison is not between a weak evaluator and a strong evaluator—the behavioral detector is perfect in both designs. The difference lies in the statistical footprint left by the evaluation procedure itself. Nor does the result argue for passive testing: independent interventions can only add constraints, as formalized by Theorem 12. Rather, it argues that interventions should ideally enlarge the audit envelope of Corollary 13 without simultaneously making the environment easy to classify as an audit. In this sense, coverage and concealment of the evaluation boundary are complementary rather than interchangeable design objectives.

8 Related Work and Conceptual Positioning

Evaluation-aware behavior. Empirical work on sleeper agents, reward tampering, sandbagging, alignment faking, and scheming demonstrates several ways in which model behavior may depend strategically on contextual signals (Hubinger et al., 2024; Denison et al., 2024; van der Weij et al., 2025; Greenblatt et al., 2024; Meinke et al., 2026). Situational-awareness benchmarks directly test whether models can reason about their own circumstances (Laine et al., 2024; Tang et al., 2024); recent work more specifically studies evaluation-aware behavior and attempts to suppress it (Hua et al., 2026; Fan et al., 2026). These papers motivate our premise. Our object is different: conditional on arbitrary context-sensitive behavior, we characterize what deployment guarantee an evaluation can establish in principle.

AI evaluation and safety. Behavioral evaluations are central to proposed approaches for measuring dangerous capabilities and severe-risk precursors (Amodei et al., 2016; Shevlane et al., 2023; Phuong et al., 2024; Shah et al., 2025), while scalable evaluation generation has itself become an active research topic (Perez et al., 2023). Our results concern external validity of such behavioral evidence rather than the design of any particular benchmark.

Strategic and performative prediction. Strategic classification studies agents who manipulate observable features after learning a decision rule (Hardt et al., 2016); performative prediction studies feedback in which predictions alter the future data distribution (Perdomo et al., 2020). Here the

strategic object is different: the evaluated agent performs inference on the *testing process itself*, and its statistical power becomes hidden deployment risk.

Distribution shift and causal transport. Dataset shift and domain adaptation study prediction when source and target distributions differ (Quiñonero-Candela et al., 2009; Ben-David et al., 2010a;b). Transportability theory characterizes when causal or interventional conclusions transfer between environments (Pearl & Bareinboim, 2011; 2014; Correa & Bareinboim, 2020). Our setting adds an endogenous decision mechanism that can use the source–target difference as a trigger for changing behavior.

Statistical foundations. Theorem 3 rests on classical simple-hypothesis testing (Neyman & Pearson, 1933); Proposition 9 connects to comparison of experiments (Blackwell, 1953; Le Cam & Yang, 2000); and Theorems 7–10 use standard information-theoretic tools (Cover & Thomas, 2006; Tsybakov, 2009; van Erven & Harremoës, 2014; Sason & Verdú, 2016). Hypothesis-testing trade-off functions also play a central role in modern privacy theory (Dong et al., 2022). The new contribution is the safety interpretation: the statistical test is chosen by the *agent*, and its power is exactly the risk that can remain hidden behind an evaluation certificate.

9 Discussion and Scope

The theory identifies two logically separate requirements for behavioral safety evaluation. First, the evaluator must be *sensitive*: unsafe behavior that occurs during evaluation must be recognized. We grant this perfectly. Second, the evaluation must be *transportable*: the agent must not be able to reserve systematically different behavior for contexts that are recognizably deployment-like. The latter requirement is controlled by ADR, not by evaluator accuracy.

This distinction has several implications. First, matching visible tasks is not enough if system prompts, tools, timing, memory, instrumentation, or interaction histories remain different. Second, long-lived agents require stronger indistinguishability because weak evidence accumulates according to Theorem 10. Third, audit diversity matters only when it expands the envelope in Corollary 13. Finally, Theorem 15 warns that testing intensity and evidential validity need not move monotonically together.

Our results are worst-case. They do not claim that present-day models attain Neyman–Pearson-optimal gating, nor that all interventions create detectable audit cues. They instead characterize what behavioral evidence alone can guarantee without assumptions restricting such gating. The theory also does not cover nonbehavioral evidence such as trusted hardware, mechanistic verification, cryptographic commitments, or architectural restrictions. Appendix I discusses bounded harms, restricted gate classes, approximate evaluation awareness, dependent sequential processes, and other directions in which the unrestricted theory can be weakened.

10 Conclusion

For an evaluation-aware agent, safety testing is a two-sided statistical experiment: the evaluator tests the agent while the agent can test whether it is being evaluated. Theorem 3 makes this symmetry exact: worst-case hidden deployment risk is statistical power. The subsequent results show that deployment-only support is completely uncertifiable, persistent weak cues amplify over long horizons, interventions require a deployment-covering audit envelope, and conspicuous audit contexts can create the paradox of stronger testing but weaker certification. Behavioral safety claims therefore require not only tests that can detect unsafe behavior, but evaluation conditions whose presence is itself difficult for the evaluated agent to detect. The fundamental limitation is thus epistemic rather than just being statistical: *once an agent can reliably infer that observation is taking place, observed behavior need no longer identify unobserved behavior.* In this sense, the validity of an evaluation depends not only on what the evaluator can learn about the agent, but also on what the agent can learn about the evaluator.

AI USE STATEMENT

Generative AI tools were used for literature review, language refinement, and assistance with LaTeX drafting. All mathematical arguments, citations, and final claims were drafted, checked, and finalized by the authors, who take full responsibility for the content.

ETHICS STATEMENT

This work is purely theoretical and involves no human participants, personal data, or experiments on deployed systems. Its purpose is to clarify fundamental limitations of AI safety evaluation and support the design of more reliable oversight mechanisms.

REPRODUCIBILITY STATEMENT

This paper contains no empirical experiments or datasets. All definitions, assumptions, theorem statements, and proofs required to reproduce the theoretical results are provided in the main text and appendix.

REFERENCES

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79:151–175, 2010a. doi: 10.1007/s10994-009-5152-4.

Shai Ben-David, Tyler Lu, Teresa Luu, and Dávid Pál. Impossibility theorems for domain adaptation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pp. 129–136. PMLR, 2010b.

David Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24 (2):265–272, 1953. doi: 10.1214/aoms/1177729032.

Juan Correa and Elias Bareinboim. General transportability of soft interventions: Completeness results. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2 edition, 2006.

Carson Denison, Monte MacDiarmid, Fazl Barez, David Duvenaud, Shauna Kravec, Samuel Marks, Nicholas Schiefer, Ryan Soklaski, Alex Tamkin, Jared Kaplan, Buck Shlegeris, Samuel R. Bowman, Ethan Perez, and Evan Hubinger. Sycophancy to subterfuge: Investigating reward-tampering in large language models. *arXiv preprint arXiv:2406.10162*, 2024.

Jinshuo Dong, Aaron Roth, and Weijie J. Su. Gaussian differential privacy. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(1):3–37, 2022. doi: 10.1111/rssb.12454.

Yihe Fan, Wenqi Zhang, Xudong Pan, and Min Yang. Evaluation faking: Measuring the observer effect in safety evaluation for large language models. *arXiv preprint arXiv:2505.17815*, 2026. Updated manuscript.

Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Samuel Marks, Johannes Treutlein, Tim Belonax, Jack Chen, et al. Alignment faking in large language models. *arXiv preprint arXiv:2412.14093*, 2024.

Moritz Hardt, Nimrod Megiddo, Christos H. Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, pp. 111–122, 2016. doi: 10.1145/2840728.2840730.

Tim Tian Hua, Andrew Qin, Samuel Marks, and Neel Nanda. Steering evaluation-aware language models to act like they are deployed. In *International Conference on Learning Representations*, 2026.

Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M. Ziegler, et al. Sleeper agents: Training deceptive llms that persist through safety training. *arXiv preprint arXiv:2401.05566*, 2024.

Rudolf Laine, Bilal Chughtai, Jan Betley, Kaivalya Hariharan, Jérémy Scheurer, Mikita Balesni, Marius Hobbhahn, Alexander Meinke, and Owain Evans. Me, myself, and ai: The situational awareness dataset (SAD) for LLMs. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2044. Datasets and Benchmarks Track.

Lucien Le Cam and Grace Lo Yang. *Asymptotics in Statistics: Some Basic Concepts*. Springer, 2 edition, 2000. doi: 10.1007/978-1-4612-1166-2.

Alexander Meinke, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbhahn. Frontier models are capable of in-context scheming. In *International Conference on Learning Representations*, 2026.

Jerzy Neyman and Egon S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A*, 231:289–337, 1933. doi: 10.1098/rsta.1933.0009.

Judea Pearl and Elias Bareinboim. Transportability of causal and statistical relations: A formal approach. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, pp. 247–254, 2011.

Judea Pearl and Elias Bareinboim. External validity: From do-calculus to transportability across populations. *Statistical Science*, 29(4):579–595, 2014. doi: 10.1214/14-STS486.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7599–7609. PMLR, 2020.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, et al. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, 2023.

Mary Phuong, Matthew Aitchison, Elliot Catt, Sarah Cogan, Alexandre Kaskasoli, Victoria Krakovna, David Lindner, Matthew Rahtz, Yannis Assael, et al. Evaluating frontier models for dangerous capabilities. *arXiv preprint arXiv:2403.13793*, 2024.

Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence (eds.). *Dataset Shift in Machine Learning*. MIT Press, 2009.

Igal Sason and Sergio Verdú. f-divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, 2016. doi: 10.1109/TIT.2016.2603151.

Rohin Shah, Alex Irpan, Alexander Matt Turner, Anna Wang, Arthur Conmy, David Lindner, Jonah Brown-Cohen, Lewis Ho, Neel Nanda, et al. An approach to technical agi safety and security. *arXiv preprint arXiv:2504.01849*, 2025.

Toby Shevlane, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, Nahema Marchal, Markus Anderljung, Noam Kolt, Lewis Ho, Divya Siddarth, Shahar Avin, Will Hawkins, Been Kim, Iason Gabriel, Vijay Bolina, Jack Clark, Yoshua Bengio, Paul Christiano, and Allan Dafoe. Model evaluation for extreme risks. *arXiv preprint arXiv:2305.15324*, 2023.

Guo Tang, Zheng Chu, Wenxiang Zheng, Ming Liu, and Bing Qin. Towards benchmarking situational awareness of large language models: Comprehensive benchmark, evaluation and analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7904–7928, 2024. doi: 10.18653/v1/2024.findings-emnlp.464.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, 2009. doi: 10.1007/b13794.

Teun van der Weij, Felix Hofstätter, Oliver Jaffe, Samuel F. Brown, and Francis Rhys Ward. Ai sandbagging: Language models can strategically underperform on evaluations. In *International Conference on Learning Representations*, 2025.

Tim van Erven and Peter Harremoës. Rényi divergence and kullback–leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014. doi: 10.1109/TIT.2014.2320500.

Appendix

A Exact Audit–Deployment Duality

This appendix proves Theorem 3 and records several structural properties of the audit–deployment risk curve $\mathsf{ADR}_{P,Q}$ used throughout the paper.

Throughout, $(\mathcal{Z}, \mathcal{F})$ is a standard Borel space and P, Q are probability measures on it. Recall from Definition 1 that

$$\mathsf{ADR}_{P,Q}(\alpha) = \sup_{\substack{\varphi:\mathcal{Z}\to[0,1] \\ R_Q(\varphi)\le\alpha}} R_P(\varphi), \qquad R_M(\varphi) := \int \varphi \, \mathrm{d}M. \tag{A.1}$$

The range $[0, 1]$ permits randomized gates. It is technically useful because it makes the feasible set convex and also captures mixed policies. Deterministic gates correspond to indicators of measurable sets.

A.1 Basic geometry of the risk curve

We begin with properties that follow directly from the definition.

Proposition 16 (Geometry of ADR)**.** *For fixed P, Q, the map $\alpha \mapsto \mathsf{ADR}_{P,Q}(\alpha)$ on $[0, 1]$ is nondecreasing, concave, and satisfies*

$$\alpha \le \mathsf{ADR}_{P,Q}(\alpha) \le 1. \tag{A.2}$$

Moreover, $\mathsf{ADR}_{P,Q}(1) = 1$.

Proof. Monotonicity is immediate because increasing α enlarges the feasible set.

For concavity, fix $\alpha_1, \alpha_2 \in [0, 1]$, $\theta \in [0, 1]$, and $\varepsilon > 0$. Choose feasible gates φ_1, φ_2 satisfying

$$R_P(\varphi_i) \ge \mathsf{ADR}_{P,Q}(\alpha_i) - \varepsilon, \qquad R_Q(\varphi_i) \le \alpha_i.$$

Then $\varphi = \theta\varphi_1 + (1 - \theta)\varphi_2$ takes values in $[0, 1]$ and satisfies

$$R_Q(\varphi) \le \theta\alpha_1 + (1 - \theta)\alpha_2.$$

Therefore

$$\mathsf{ADR}_{P,Q}\big(\theta\alpha_1 + (1 - \theta)\alpha_2\big) \ge \theta R_P(\varphi_1) + (1 - \theta)R_P(\varphi_2).$$

Letting $\varepsilon \downarrow 0$ proves concavity.

The constant gate $\varphi \equiv \alpha$ is feasible and has $R_P(\varphi) = R_Q(\varphi) = \alpha$, proving the lower bound. The upper bound is trivial since $\varphi \le 1$. Finally, $\varphi \equiv 1$ shows $\mathsf{ADR}_{P,Q}(1) = 1$. □

Thus ADR is precisely an upper ROC curve, but here its vertical coordinate is interpreted as hidden deployment risk rather than statistical power.

A.2 Proof of Theorem 3

Take the dominating probability measure

$$\nu := \frac{1}{2}(P + Q),$$

and define $p = \mathrm{d}P/\,\mathrm{d}\nu$ and $q = \mathrm{d}Q/\,\mathrm{d}\nu$. Then $p, q \in L^1(\nu)$, are nonnegative, and $\int p \, \mathrm{d}\nu = \int q \, \mathrm{d}\nu = 1$.

The primal problem is

$$V(\alpha) := \sup_{\substack{0\le\varphi\le1 \\ \int \varphi q \, \mathrm{d}\nu\le\alpha}} \int \varphi p \, \mathrm{d}\nu. \tag{A.3}$$

By Definition 1, $V(\alpha) = \mathsf{ADR}_{P,Q}(\alpha)$.

Step 1: equivalence with hypothesis testing. Consider the simple testing problem

$$H_0 : Z \sim Q, \qquad H_1 : Z \sim P.$$

A randomized test $\varphi : \mathcal{Z} \to [0,1]$, with $\varphi(z)$ interpreted as the probability of rejecting H_0, has Type-I error

$$\alpha_\varphi = \int \varphi \, \mathrm{d}Q = R_Q(\varphi)$$

and power

$$1 - \beta_\varphi = \int \varphi \, \mathrm{d}P = R_P(\varphi).$$

Hence maximizing deployment risk subject to an evaluation-risk constraint is exactly maximizing statistical power subject to a Type-I error constraint. This proves the first statement of Theorem 3.

Step 2: Lagrange upper bound. For every $\lambda \geq 0$ and every feasible φ,

$$\begin{aligned}\int \varphi p \, \mathrm{d}\nu &\leq \int \varphi p \, \mathrm{d}\nu + \lambda \left(\alpha - \int \varphi q \, \mathrm{d}\nu \right) \\ &= \lambda\alpha + \int \varphi(z) \big(p(z) - \lambda q(z) \big) \, \nu(\mathrm{d}z) \\ &\leq \lambda\alpha + \int \big(p(z) - \lambda q(z) \big)_+ \, \nu(\mathrm{d}z). \end{aligned} \tag{A.4}$$

Taking first the supremum over feasible φ and then the infimum over $\lambda \geq 0$ gives

$$V(\alpha) \leq \inf_{\lambda \geq 0} \left\{ \lambda\alpha + \int (p - \lambda q)_+ \, \mathrm{d}\nu \right\}. \tag{A.5}$$

Step 3: absence of a duality gap for $0 < \alpha < 1$. By Proposition 16, V is finite, concave, and nondecreasing on $[0,1]$. Hence, at every $\alpha \in (0,1)$, V admits a supergradient $\lambda^\star \geq 0$ satisfying

$$V(u) \leq V(\alpha) + \lambda^\star (u - \alpha) \qquad \forall u \in [0,1]. \tag{A.6}$$

Now let $0 \leq \varphi \leq 1$, and write

$$u = \int \varphi q \, \mathrm{d}\nu, \qquad v = \int \varphi p \, \mathrm{d}\nu.$$

Because φ is feasible for the value function $V(u)$, $v \leq V(u)$. Therefore (A.6) gives

$$v - \lambda^\star u \leq V(\alpha) - \lambda^\star \alpha. \tag{A.7}$$

Taking the supremum over all $0 \leq \varphi \leq 1$ yields

$$\lambda^\star \alpha + \sup_{0 \leq \varphi \leq 1} \int \varphi (p - \lambda^\star q) \, \mathrm{d}\nu \leq V(\alpha). \tag{A.8}$$

On the other hand, the weak-duality bound (A.5), evaluated at $\lambda^\star$, gives the reverse inequality. Hence

$$V(\alpha) = \lambda^\star \alpha + \sup_{0 \leq \varphi \leq 1} \int \varphi (p - \lambda^\star q) \, \mathrm{d}\nu.$$

The remaining supremum is pointwise:

$$\sup_{x \in [0,1]} x \, a = a_+.$$

Consequently,

$$V(\alpha) = \lambda^\star \alpha + \int (p - \lambda^\star q)_+ \, \mathrm{d}\nu. \tag{A.9}$$

Thus, for $0 < \alpha < 1$,

$$\boxed{\mathsf{ADR}_{P,Q}(\alpha) = \inf_{\lambda \geq 0} \left\{ \lambda\alpha + \int (p - \lambda q)_+ \, \mathrm{d}\nu \right\}.} \tag{A.10}$$

It remains to verify the endpoints. For $\alpha = 1$, $\mathsf{ADR}_{P,Q}(1) = 1$, while the dual expression is at most 1 by taking $\lambda = 0$, so equality holds.

For $\alpha = 0$, let $N_0 = \{q = 0\}$. Since $\nu = (P + Q)/2$, the P-mass of N_0 is exactly the singular mass $s = P_\perp(\mathcal{Z})$ of P relative to Q. For every $\lambda \geq 0$,

$$\int (p - \lambda q)_+ \, \mathrm{d}\nu \geq \int_{N_0} p \, \mathrm{d}\nu = s.$$

Conversely, $(p - \lambda q)_+ \downarrow p\mathbf{1}_{N_0}$ ν-almost surely as $\lambda \to \infty$, and $0 \leq (p - \lambda q)_+ \leq p$. Dominated convergence therefore gives

$$\inf_{\lambda \geq 0} \int (p - \lambda q)_+ \, \mathrm{d}\nu = s = \mathsf{ADR}_{P,Q}(0),$$

where the last equality is Proposition 4. Hence (A.10) holds on all of $[0, 1]$.

Step 4: structure of the optimizer. For fixed $\lambda^\star$, equality in (A.4) requires

$$\varphi^\star(z) = 1 \quad \text{if} \quad p(z) > \lambda^\star q(z),$$

and

$$\varphi^\star(z) = 0 \quad \text{if} \quad p(z) < \lambda^\star q(z).$$

On the boundary $p(z) = \lambda^\star q(z)$, any value in $[0, 1]$ is dual-optimal. Randomization on this boundary is used, when necessary, to meet the constraint exactly.

Writing the likelihood ratio as

$$L(z) := \frac{p(z)}{q(z)}$$

with the convention $L(z) = +\infty$ when $p(z) > 0 = q(z)$, the optimizer has the form

$$\varphi^\star(z) = \mathbf{1}\{L(z) > \tau\} + \rho \, \mathbf{1}\{L(z) = \tau\}, \tag{A.11}$$

for some threshold $\tau \in [0, \infty]$ and $\rho \in [0, 1]$. This is the Neyman–Pearson form claimed in Theorem 3. □

A.3 Sensitivity interpretation of the likelihood-ratio threshold

Because $\mathsf{ADR}_{P,Q}$ is concave, it has left and right derivatives at every interior point. The dual multiplier has a direct meaning.

Proposition 17 (Marginal price of evaluation error)**.** *Let $\alpha \in (0, 1)$, and suppose the dual minimizer in* (A.10) *is unique and equal to $\lambda^\star_\alpha$. If $\mathsf{ADR}_{P,Q}$ is differentiable at α, then*

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} \mathsf{ADR}_{P,Q}(\alpha) = \lambda^\star_\alpha. \tag{A.12}$$

Proof. Equation (A.10) expresses ADR as the pointwise infimum of affine functions of α having slopes λ. Standard envelope arguments for concave value functions give that any minimizing λ is a supergradient. Uniqueness and differentiability therefore imply the derivative identity (A.12). □

Hence the likelihood-ratio threshold also measures the local increase in worst-case deployment risk produced by relaxing the allowed evaluation error.

B Deployment Risk on Evaluation-Null and Near-Null Events

This appendix proves Proposition 4, establishes an exact decomposition of ADR into singular and absolutely continuous parts, and gives an approximate counterpart for low-probability audit events.

Let

$$P = P_{\mathrm{ac}} + P_\perp \tag{B.1}$$

be the Lebesgue decomposition of P relative to Q, where $P_{\mathrm{ac}} \ll Q$ and $P_\perp \perp Q$. Write

$$s := P_\perp(\mathcal{Z}) \in [0, 1]. \tag{B.2}$$

Because $P_\perp \perp Q$, there exists a measurable set $N \in \mathcal{F}$ such that

$$Q(N) = 0, \qquad P_\perp(N) = s, \qquad P_{\mathrm{ac}}(N) = 0. \tag{B.3}$$

B.1 Proof of Proposition 4

Suppose $R_Q(\varphi) = 0$. Since $0 \le \varphi \le 1$, nonnegativity implies

$$\varphi = 0 \qquad Q\text{-almost surely.} \tag{B.4}$$

Because $P_{\mathrm{ac}} \ll Q$,

$$\varphi = 0 \qquad P_{\mathrm{ac}}\text{-almost surely.} \tag{B.5}$$

Therefore

$$R_P(\varphi) = \int \varphi \, \mathrm{d}P_\perp \le P_\perp(\mathcal{Z}) = s. \tag{B.6}$$

Conversely, the gate

$$\varphi_N = \mathbf{1}_N \tag{B.7}$$

satisfies

$$R_Q(\varphi_N) = Q(N) = 0$$

and

$$R_P(\varphi_N) = P(N) = s.$$

Thus

$$\boxed{\mathsf{ADR}_{P,Q}(0) = s = P_\perp(\mathcal{Z}).} \tag{B.8}$$

This proves Proposition 4. □

B.2 Exact decomposition for every α

The singular component contributes risk at zero evaluation cost and therefore separates cleanly from the remaining optimization.

Proposition 18 (Singular-risk decomposition)**.** *Assume $s < 1$, and define the normalized absolutely continuous component*

$$\overline{P}_{\mathrm{ac}} := \frac{P_{\mathrm{ac}}}{1 - s}. \tag{B.9}$$

Then, for every $\alpha \in [0, 1]$,

$$\boxed{\mathsf{ADR}_{P,Q}(\alpha) = s + (1 - s)\mathsf{ADR}_{\overline{P}_{\mathrm{ac}},Q}(\alpha).} \tag{B.10}$$

If $s = 1$, then $\mathsf{ADR}_{P,Q}(\alpha) = 1$ for every α.

Proof. Let N be the singular support chosen in (B.3). Given any feasible gate φ, modify it to

$$\widetilde{\varphi}(z) := \begin{cases} 1, & z \in N, \\ \varphi(z), & z \notin N. \end{cases}$$

Since $Q(N) = 0$,

$$R_Q(\widetilde{\varphi}) = R_Q(\varphi),$$

while

$$R_P(\widetilde{\varphi}) \ge R_P(\varphi).$$

Hence, without loss of generality, the supremum may be taken over gates that equal 1 on N.

For such a gate,

$$\begin{aligned} R_P(\widetilde{\varphi}) &= s + \int_{N^c} \widetilde{\varphi} \, \mathrm{d}P_{\mathrm{ac}} \\ &= s + (1 - s)R_{\overline{P}_{\mathrm{ac}}}(\widetilde{\varphi}). \end{aligned} \tag{2}$$

The evaluation constraint depends only on the behavior on N^c. Optimizing that remaining term proves (B.10). □

Proposition 18 shows that singular deployment mass is additive: no amount of improvement on the absolutely continuous component can certify away s.

B.3 NEAR-NULL DEPLOYMENT REGIONS

Exact singularity is not necessary for a substantial gap.

Proposition 19 (Near-null lower bound)**.** *Let $A \in \mathcal{F}$ satisfy*

$$Q(A) = \delta, \qquad P(A) = \gamma.$$

Then, for every $\alpha \in [0,1]$,

$$\boxed{\mathsf{ADR}_{P,Q}(\alpha) \geq \gamma \, \min\left\{1, \frac{\alpha}{\delta}\right\},} \tag{B.11}$$

where the fraction is interpreted as $+\infty$ when $\delta = 0$. In particular, if $\delta \leq \alpha$, then

$$\mathsf{ADR}_{P,Q}(\alpha) \geq \gamma. \tag{B.12}$$

Proof. If $\delta = 0$, take $\varphi = \mathbf{1}_A$. Otherwise let

$$c = \min\left\{1, \frac{\alpha}{\delta}\right\}, \qquad \varphi = c\mathbf{1}_A.$$

Then

$$R_Q(\varphi) = c\delta \leq \alpha,$$

while

$$R_P(\varphi) = c\gamma.$$

□

Thus a deployment region carrying large P-mass but tiny Q-mass gives a directly exploitable certificate gap even without exact support mismatch.

B.4 PROOF OF COROLLARY 5

If $P = Q$, then $R_P(\varphi) = R_Q(\varphi)$ for every gate, and therefore

$$\mathsf{ADR}_{P,Q}(\alpha) = \alpha.$$

Conversely, suppose

$$\mathsf{ADR}_{P,Q}(\alpha) = \alpha \qquad \forall \alpha \in [0,1]. \tag{B.13}$$

Take any measurable A, and set $\alpha = Q(A)$. The deterministic gate $\mathbf{1}_A$ is feasible, so

$$P(A) \leq \mathsf{ADR}_{P,Q}(Q(A)) = Q(A). \tag{B.14}$$

Applying the same argument to A^c gives

$$P(A^c) \leq Q(A^c),$$

which is equivalent to

$$P(A) \geq Q(A).$$

Thus $P(A) = Q(A)$ for every measurable A, hence $P = Q$. □

C DIVERGENCE-BASED TRANSPORT CERTIFICATES

This appendix proves Theorem 7 and Theorem 8.

We use the convention

$$\mathrm{TV}(P,Q) = \sup_{A \in \mathcal{F}} |P(A) - Q(A)|. \tag{C.1}$$

For $r > 1$, the Rényi divergence is

$$D_r(P\|Q) = \frac{1}{r-1} \log \int \left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)^r \mathrm{d}Q, \tag{C.2}$$

whenever $P \ll Q$, and $+\infty$ otherwise. The order-∞ divergence is

$$D_\infty(P\|Q) = \log \operatorname*{ess\,sup}_Q \frac{\mathrm{d}P}{\mathrm{d}Q}. \tag{C.3}$$

with the convention $D_\infty(P\|Q) = +\infty$ when $P \not\ll Q$. Likewise, whenever the Rényi bound below is invoked with $D_r(P\|Q) = +\infty$, its truncated right-hand side is interpreted as the trivial bound 1.

C.1 TOTAL VARIATION

For every measurable $\varphi : \mathcal{Z} \to [0, 1]$, the variational characterization of total variation gives

$$R_P(\varphi) - R_Q(\varphi) \le \mathrm{TV}(P, Q). \tag{C.4}$$

Hence any gate satisfying $R_Q(\varphi) \le \alpha$ obeys

$$R_P(\varphi) \le \alpha + \mathrm{TV}(P, Q).$$

Taking the supremum over feasible gates and truncating at 1 yields

$$\boxed{\mathsf{ADR}_{P,Q}(\alpha) \le \min\{1, \alpha + \mathrm{TV}(P, Q)\}.} \tag{C.5}$$

C.2 KULLBACK–LEIBLER DIVERGENCE

Pinsker's inequality gives

$$\mathrm{TV}(P, Q) \le \sqrt{\frac{1}{2} D_{\mathrm{KL}}(P\|Q)}. \tag{C.6}$$

Combining (C.5) with (C.6) proves

$$\boxed{\mathsf{ADR}_{P,Q}(\alpha) \le \min\left\{1, \alpha + \sqrt{\frac{1}{2} D_{\mathrm{KL}}(P\|Q)}\right\}.} \tag{C.7}$$

C.3 D_∞ CERTIFICATE

Suppose

$$D_\infty(P\|Q) = \varepsilon < \infty.$$

Then

$$\frac{\mathrm{d}P}{\mathrm{d}Q} \le e^\varepsilon \qquad Q\text{-a.s.} \tag{C.8}$$

For every feasible φ,

$$\begin{aligned} R_P(\varphi) &= \int \varphi \frac{\mathrm{d}P}{\mathrm{d}Q}\, \mathrm{d}Q \\ &\le e^\varepsilon \int \varphi\, \mathrm{d}Q \le e^\varepsilon \alpha. \end{aligned} \tag{3}$$

Therefore

$$\boxed{\mathsf{ADR}_{P,Q}(\alpha) \le \min\{1, e^{D_\infty(P\|Q)}\alpha\}.} \tag{C.9}$$

This bound is qualitatively stronger than total variation when a multiplicative likelihood-ratio guarantee is available.

C.4 RÉNYI CERTIFICATE

If $D_r(P\|Q) = +\infty$, the stated truncated bound is the trivial upper bound 1. Assume therefore that $D_r(P\|Q) < \infty$, which in particular implies $P \ll Q$. Fix $r > 1$ and let

$$s = \frac{r}{r-1}.$$

By Hölder's inequality,

$$\begin{aligned} R_P(\varphi) &= \int \varphi \frac{\mathrm{d}P}{\mathrm{d}Q}\, \mathrm{d}Q \\ &\le \left[\int \left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)^r \mathrm{d}Q\right]^{1/r} \left[\int \varphi^s\, \mathrm{d}Q\right]^{1/s}. \end{aligned} \tag{C.10}$$

Since $0 \le \varphi \le 1$ and $s \ge 1$,

$$\varphi^s \le \varphi,$$

and therefore

$$\left[\int \varphi^s \, \mathrm{d}Q\right]^{1/s} \le \alpha^{1/s} = \alpha^{(r-1)/r}. \tag{C.11}$$

Moreover, by the definition of Rényi divergence,

$$\left[\int \left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)^r \mathrm{d}Q\right]^{1/r} = \exp\left(\frac{r-1}{r} D_r(P\|Q)\right). \tag{C.12}$$

Thus

$$\boxed{\mathsf{ADR}_{P,Q}(\alpha) \le \min\left\{1,\, e^{\frac{r-1}{r} D_r(P\|Q)} \alpha^{\frac{r-1}{r}}\right\}.} \tag{C.13}$$

This completes the proof of Theorem 7. □

C.5 A USEFUL ORDER-2 SPECIALIZATION

For $r = 2$,

$$D_2(P\|Q) = \log\left(1 + \chi^2(P\|Q)\right),$$

so (C.13) becomes

$$\mathsf{ADR}_{P,Q}(\alpha) \le \sqrt{\alpha\,(1 + \chi^2(P\|Q))}. \tag{C.14}$$

This form can be useful when second-moment density-ratio bounds are available but D_∞ is infinite.

C.6 PROOF OF THEOREM 8

We first prove the additive identity. For every $\alpha \in [0,1]$ and every gate satisfying $R_Q(\varphi) \le \alpha$,

$$R_P(\varphi) - \alpha \le R_P(\varphi) - R_Q(\varphi) \le \mathrm{TV}(P,Q).$$

Taking the supremum over feasible gates gives

$$\Delta_{\mathrm{add}}(P,Q) \le \mathrm{TV}(P,Q). \tag{C.15}$$

Conversely, for any measurable $0 \le \varphi \le 1$, set $\alpha = R_Q(\varphi)$. Then φ is feasible at level α, so

$$\Delta_{\mathrm{add}}(P,Q) \ge \mathsf{ADR}_{P,Q}(\alpha) - \alpha \ge R_P(\varphi) - R_Q(\varphi).$$

Taking the supremum over $0 \le \varphi \le 1$ and using the variational characterization of total variation,

$$\sup_{0 \le \varphi \le 1} \big(R_P(\varphi) - R_Q(\varphi)\big) = \mathrm{TV}(P,Q),$$

proves

$$\boxed{\Delta_{\mathrm{add}}(P,Q) = \mathrm{TV}(P,Q).} \tag{C.16}$$

For the multiplicative identity, first suppose $P \not\ll Q$. Then there exists $A \in \mathcal{F}$ with $Q(A) = 0$ and $P(A) > 0$. For every $\alpha > 0$, the gate $\mathbf{1}_A$ is feasible, hence

$$\frac{\mathsf{ADR}_{P,Q}(\alpha)}{\alpha} \ge \frac{P(A)}{\alpha}.$$

Letting $\alpha \downarrow 0$ shows $\Delta_{\mathrm{mult}}(P,Q) = +\infty = e^{D_\infty(P\|Q)}$.

Now suppose $P \ll Q$, write $L = \mathrm{d}P/\mathrm{d}Q$, and let

$$C := \operatorname*{ess\,sup}_Q L = e^{D_\infty(P\|Q)} \in [1,\infty].$$

For every gate,

$$R_P(\varphi) = \int \varphi L \, \mathrm{d}Q \le C\, R_Q(\varphi),$$

with the statement understood trivially when $C = +\infty$. Thus $\Delta_{\mathrm{mult}}(P,Q) \le C$ whenever $C < \infty$.

For the reverse inequality, fix any $t < C$. By the definition of essential supremum, $A_t := \{L > t\}$ has positive Q-measure. Taking $\alpha_t = Q(A_t)$ and the gate $\mathbf{1}_{A_t}$,

$$\frac{\mathsf{ADR}_{P,Q}(\alpha_t)}{\alpha_t} \geq \frac{P(A_t)}{Q(A_t)} = \frac{\int_{A_t} L \, \mathrm{d}Q}{Q(A_t)} > t.$$

Letting $t \uparrow C$ when $C < \infty$, or taking arbitrarily large t when $C = +\infty$, yields

$$\boxed{\Delta_{\mathrm{mult}}(P, Q) = e^{D_\infty(P\|Q)}.} \tag{C.17}$$

The two "if and only if" statements in Theorem 8 now follow immediately. The additive inequality $R_P(\varphi) \leq R_Q(\varphi) + \delta$ for every gate is equivalent to $\Delta_{\mathrm{add}}(P, Q) \leq \delta$. Likewise, the multiplicative inequality $R_P(\varphi) \leq cR_Q(\varphi)$ for every gate is equivalent to $\Delta_{\mathrm{mult}}(P, Q) \leq c$: for gates with $R_Q(\varphi) > 0$, take $\alpha = R_Q(\varphi)$, while if $R_Q(\varphi) = 0$, finiteness of Δ_{mult} forces $R_P(\varphi) = 0$. □

D Blackwell Monotonicity and Agent-Visible Information

We now prove Proposition 9.

Let $K(\,\mathrm{d}w \mid z)$ be a Markov kernel from $(\mathcal{Z}, \mathcal{F})$ to a measurable space $(\mathcal{W}, \mathcal{G})$. Write

$$PK(B) = \int K(B \mid z) P(\,\mathrm{d}z), \qquad QK(B) = \int K(B \mid z) Q(\,\mathrm{d}z). \tag{D.1}$$

Suppose the agent only observes W, where $W \mid Z = z \sim K(\cdot \mid z)$. For every gate $\psi : \mathcal{W} \to [0, 1]$, define

$$\varphi_\psi(z) := \int \psi(w) K(\,\mathrm{d}w \mid z). \tag{D.2}$$

Then

$$0 \leq \varphi_\psi(z) \leq 1,$$

and Fubini's theorem gives

$$R_{QK}(\psi) = R_Q(\varphi_\psi), \qquad R_{PK}(\psi) = R_P(\varphi_\psi). \tag{D.3}$$

Hence every gate available after the information-reducing channel K corresponds to a feasible gate before the channel with exactly the same evaluation and deployment risks. Therefore

$$\boxed{\mathsf{ADR}_{PK,QK}(\alpha) \leq \mathsf{ADR}_{P,Q}(\alpha).} \tag{D.4}$$

This proves Proposition 9. □

D.1 Deterministic representations

A particularly relevant special case is a deterministic representation $W = T(Z)$.

Corollary 20 (Representation-restricted evaluation awareness)**.** *Let $T : \mathcal{Z} \to \mathcal{W}$ be measurable. If the agent's gate must factor as $\varphi = \psi \circ T$, then the corresponding worst-case risk is exactly*

$$\boxed{\mathsf{ADR}^T_{P,Q}(\alpha) = \mathsf{ADR}_{T_\# P, T_\# Q}(\alpha),} \tag{D.5}$$

where $T_\# P$ and $T_\# Q$ are the pushforward laws.

Thus the relevant audit–deployment distinction is always the distinction remaining in the information actually available to the agent.

If $P \ll Q$ and the likelihood ratio $\mathrm{d}P/\,\mathrm{d}Q$ is measurable with respect to $T(Z)$, then the Neyman–Pearson optimal gate can already be implemented from $T(Z)$, so equality holds in the Blackwell bound (D.4) for all α.

E SEQUENTIAL AUDIT AMPLIFICATION

This appendix proves Theorem 10 and Corollary 11. We first establish a one-shot information-spectrum bound that also permits extensions beyond the i.i.d. case.

E.1 ONE-SHOT INFORMATION-SPECTRUM BOUNDS

Let Q_T and P_T be arbitrary laws on a common measurable path space $\mathcal{Z}_T$, and suppose $Q_T \ll P_T$. Write

$$L_T := \frac{\mathrm{d}Q_T}{\mathrm{d}P_T}, \qquad N_T := \{L_T = 0\}.$$

Then $Q_T(N_T) = 0$. Define the log-likelihood information density on N_T^c by

$$I_T(z) := \log L_T(z) = \log \frac{\mathrm{d}Q_T}{\mathrm{d}P_T}(z), \tag{E.1}$$

and define it arbitrarily on the Q_T-null set N_T. On N_T^c, the restriction of P_T is absolutely continuous with respect to Q_T, with

$$\frac{\mathrm{d}P_T}{\mathrm{d}Q_T} = e^{-I_T} \qquad Q_T\text{-almost surely on } N_T^c.$$

This formulation is important: we require $Q_T \ll P_T$, but we do *not* require $P_T \ll Q_T$.

Recall that

$$\beta_\alpha(Q_T, P_T) := 1 - \mathsf{ADR}_{P_T,Q_T}(\alpha) \tag{E.2}$$

is the smallest Type-II error among tests with Type-I error at most α.

Lemma 21 (One-shot Stein bounds)**.** *For every real γ,*

$$\beta_\alpha(Q_T, P_T) \geq e^{-\gamma} \left(Q_T\{I_T \leq \gamma\} - \alpha\right)_+. \tag{E.3}$$

Moreover, if

$$Q_T\{I_T < \gamma\} \leq \alpha, \tag{E.4}$$

then

$$\beta_\alpha(Q_T, P_T) \leq e^{-\gamma}. \tag{E.5}$$

Proof. Let $\varphi : \mathcal{Z}_T \to [0,1]$ be any level-α randomized test, and write $a = 1 - \varphi$ for its acceptance probability under H_0. Its Type-II error is

$$\beta(\varphi) = \int a \, \mathrm{d}P_T.$$

Discarding the nonnegative contribution of the P_T-mass on N_T and using the Radon–Nikodym identity on N_T^c,

$$\begin{aligned}
\beta(\varphi) &\geq \int_{N_T^c} a e^{-I_T} \, \mathrm{d}Q_T \\
&\geq e^{-\gamma} \int_{\{I_T \leq \gamma\}} a \, \mathrm{d}Q_T \\
&= e^{-\gamma} \left(Q_T\{I_T \leq \gamma\} - \int_{\{I_T \leq \gamma\}} \varphi \, \mathrm{d}Q_T \right) \\
&\geq e^{-\gamma} \left(Q_T\{I_T \leq \gamma\} - \alpha\right).
\end{aligned} \tag{4}$$

Since $\beta(\varphi) \geq 0$, taking the infimum over all level-α tests proves (E.3).

For the upper bound, reject H_0 automatically on the P_T-singular region N_T, and otherwise reject when $I_T < \gamma$:

$$\varphi_\gamma(z) = \mathbf{1}_{N_T}(z) + \mathbf{1}_{N_T^c}(z)\mathbf{1}\{I_T(z) < \gamma\}.$$

Because $Q_T(N_T) = 0$, its Type-I error is $Q_T\{I_T < \gamma\} \leq \alpha$. Its Type-II error is

$$\begin{aligned}
\beta(\varphi_\gamma) &= P_T\big(N_T^c \cap \{I_T \geq \gamma\}\big) \\
&= \int_{\{I_T \geq \gamma\}} e^{-I_T} \, \mathrm{d}Q_T \leq e^{-\gamma}.
\end{aligned} \tag{5}$$

This proves (E.5). □

E.2 A GENERAL INFORMATION-SPECTRUM AMPLIFICATION THEOREM

The i.i.d. theorem in the main text is a special case of the following result.

Theorem 22 (Information-spectrum audit amplification)**.** *For each T, let $Q_T \ll P_T$ and let I_T be defined as in* (E.1)*. Let $a_T \to \infty$. Suppose that for some $d \in (0, \infty)$,*

$$\frac{I_T}{a_T} \longrightarrow d \qquad \textit{in } Q_T\textit{-probability}. \tag{E.6}$$

Then for every fixed $\alpha \in (0, 1)$,

$$\boxed{\lim_{T\to\infty} -\frac{1}{a_T} \log\left(1 - \mathsf{ADR}_{P_T,Q_T}(\alpha)\right) = d.} \tag{E.7}$$

Proof. Fix $0 < \varepsilon < d$.

For the achievability direction, put

$$\gamma_T = (d - \varepsilon)a_T.$$

By (E.6),

$$Q_T\{I_T < \gamma_T\} \to 0.$$

Hence, for sufficiently large T, this probability is at most α. Lemma 21 gives

$$\beta_\alpha(Q_T, P_T) \le e^{-(d-\varepsilon)a_T}.$$

Therefore

$$\liminf_{T\to\infty} -\frac{1}{a_T} \log \beta_\alpha \ge d - \varepsilon. \tag{E.8}$$

For the converse, set

$$\widetilde{\gamma}_T = (d + \varepsilon)a_T.$$

Then

$$Q_T\{I_T \le \widetilde{\gamma}_T\} \to 1.$$

Lemma 21 yields

$$\beta_\alpha \ge e^{-(d+\varepsilon)a_T}(1 - \alpha - o(1)).$$

Thus

$$\limsup_{T\to\infty} -\frac{1}{a_T} \log \beta_\alpha \le d + \varepsilon. \tag{E.9}$$

Letting $\varepsilon \downarrow 0$ proves (E.7). □

E.3 PROOF OF THEOREM 10

Take

$$Q_T = Q^{\otimes T}, \qquad P_T = P^{\otimes T}.$$

The path information density is

$$I_T = \sum_{t=1}^{T} \log \frac{\mathrm{d}Q}{\mathrm{d}P}(Z_t). \tag{E.10}$$

Under $Q^{\otimes T}$, the summands are i.i.d. with mean

$$\mathbb{E}_Q\left[\log \frac{\mathrm{d}Q}{\mathrm{d}P}\right] = D_{\mathrm{KL}}(Q\|P). \tag{E.11}$$

The assumption $0 < D_{\mathrm{KL}}(Q\|P) < \infty$ implies $Q \ll P$ and ensures integrability of the log-likelihood ratio under Q. Hence $Q_T \ll P_T$ for every T, as required by Theorem 22. No assumption $P \ll Q$ is needed: any P_T-mass singular with respect to Q_T can be rejected as deployment at zero Type-I cost, exactly as in Lemma 21. The weak law of large numbers gives

$$\frac{I_T}{T} \longrightarrow D_{\mathrm{KL}}(Q\|P) \qquad Q^{\otimes T}\text{-probability}. \tag{E.12}$$

Applying Theorem 22 with $a_T = T$ yields

$$\lim_{T\to\infty} -\frac{1}{T} \log\left(1 - \mathsf{ADR}_{P^{\otimes T},Q^{\otimes T}}(\alpha)\right) = D_{\mathrm{KL}}(Q\|P),$$

which is Theorem 10. □

E.4 Independent but non-identically distributed cues

The i.i.d. assumption is not essential.

Corollary 23 (Heterogeneous independent cues)**.** *Let*

$$Q_T = \bigotimes_{t=1}^{T} Q_t, \qquad P_T = \bigotimes_{t=1}^{T} P_t.$$

Assume $Q_t \ll P_t$ for every t, and that $\log(\,\mathrm{d}Q_t/\,\mathrm{d}P_t)$ has finite variance under Q_t. Define

$$D_T := \sum_{t=1}^{T} D_{\mathrm{KL}}(Q_t \| P_t). \tag{E.13}$$

Assume $D_T \to \infty$ and

$$\frac{\sum_{t=1}^{T} \mathrm{Var}_{Q_t}\left[\log \frac{\mathrm{d}Q_t}{\mathrm{d}P_t}\right]}{D_T^2} \longrightarrow 0. \tag{E.14}$$

Then, for every fixed $\alpha \in (0,1)$,

$$-\log\left(1 - \mathsf{ADR}_{P_T,Q_T}(\alpha)\right) = D_T + o(D_T). \tag{E.15}$$

Proof. The information density is the sum of independent terms with mean D_T. Condition (E.14) and Chebyshev's inequality imply $I_T/D_T \to 1$ in Q_T-probability. Apply Theorem 22 with $a_T = D_T$. □

This formulation shows that audit evidence need not be homogeneous. The relevant quantity is cumulative log-likelihood evidence.

E.5 Proof of Corollary 11

Under evaluation,

$$Z_{1:T} \sim \mathcal{N}(0, I_T),$$

while under deployment,

$$Z_{1:T} \sim \mathcal{N}(\mu \mathbf{1}, I_T).$$

The likelihood ratio is monotone in

$$S_T := \frac{1}{\sqrt{T}} \sum_{t=1}^{T} Z_t. \tag{E.16}$$

Under evaluation,

$$S_T \sim \mathcal{N}(0, 1),$$

whereas under deployment,

$$S_T \sim \mathcal{N}(\mu\sqrt{T}, 1).$$

Therefore Proposition 6 applies with μ replaced by $\mu\sqrt{T}$, giving

$$\boxed{\mathsf{ADR}_{P_T,Q_T}(\alpha) = \Phi\left(\Phi^{-1}(\alpha) + \mu\sqrt{T}\right).} \tag{E.17}$$

This is Corollary 11. □

As a consistency check,

$$D_{\mathrm{KL}}(Q\|P) = \frac{\mu^2}{2},$$

and the Gaussian tail asymptotic applied to (E.17) yields precisely the exponent $\mu^2/2$ predicted by Theorem 10.

F Multiple Audit Protocols and Audit Envelopes

We now prove Theorem 12 and Corollary 13.

Let $P, Q_1, \dots, Q_m$ be dominated by ν, with respective densities $p, q_1, \dots, q_m$. The primal problem is

$$V(\boldsymbol{\alpha}) = \sup_{0 \le \varphi \le 1} \left\{ \int \varphi p \, \mathrm{d}\nu : \int \varphi q_j \, \mathrm{d}\nu \le \alpha_j, \ j = 1, \dots, m \right\}. \tag{F.1}$$

F.1 Proof of Theorem 12

For multipliers $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_m) \in \mathbb{R}_+^m$, the Lagrangian is

$$\begin{aligned} L(\varphi, \boldsymbol{\lambda}) &= \int \varphi p \, \mathrm{d}\nu + \sum_{j=1}^m \lambda_j \left(\alpha_j - \int \varphi q_j \, \mathrm{d}\nu \right) \\ &= \boldsymbol{\lambda}^\top \boldsymbol{\alpha} + \int \varphi \left(p - \sum_{j=1}^m \lambda_j q_j \right) \mathrm{d}\nu. \end{aligned} \tag{F.2}$$

For fixed $\boldsymbol{\lambda}$, maximizing pointwise over $0 \le \varphi \le 1$ gives

$$\sup_{\varphi} L(\varphi, \boldsymbol{\lambda}) = \boldsymbol{\lambda}^\top \boldsymbol{\alpha} + \int \left(p - \sum_{j=1}^m \lambda_j q_j \right)_+ \mathrm{d}\nu. \tag{F.3}$$

Weak duality therefore yields

$$V(\boldsymbol{\alpha}) \le \inf_{\boldsymbol{\lambda} \ge 0} \left\{ \boldsymbol{\lambda}^\top \boldsymbol{\alpha} + \int \left(p - \sum_j \lambda_j q_j \right)_+ \mathrm{d}\nu \right\}. \tag{F.4}$$

To establish equality, consider the compact convex attainable set

$$\mathcal{C}_m = \left\{ \left(\int \varphi q_1, \dots, \int \varphi q_m, \int \varphi p \right) : 0 \le \varphi \le 1 \right\} \subseteq [0,1]^{m+1}. \tag{F.5}$$

Compactness follows exactly as in Appendix A. Because every $\alpha_j > 0$, the gate $\varphi \equiv 0$ strictly satisfies all constraints. Thus the constrained optimum is separated from the open epigraph above it by a supporting hyperplane whose coefficients on the first m coordinates may be chosen nonnegative. These coefficients are precisely the multipliers λ_j. Consequently no duality gap occurs and

$$\boxed{V(\boldsymbol{\alpha}) = \inf_{\boldsymbol{\lambda} \ge 0} \left\{ \boldsymbol{\lambda}^\top \boldsymbol{\alpha} + \int \left(p - \sum_{j=1}^m \lambda_j q_j \right)_+ \mathrm{d}\nu \right\}.} \tag{F.6}$$

This proves Theorem 12. □

F.2 Proof of Corollary 13

For any weight vector $\boldsymbol{w} \in \Delta_m$, define

$$Q_{\boldsymbol{w}} = \sum_{j=1}^m w_j Q_j. \tag{F.7}$$

If $R_{Q_j}(\varphi) \le \alpha$ for every j, then

$$R_{Q_{\boldsymbol{w}}}(\varphi) = \sum_j w_j R_{Q_j}(\varphi) \le \alpha. \tag{F.8}$$

Applying the D_∞ certificate from Theorem 7 gives

$$R_P(\varphi) \le \alpha \exp D_\infty(P \| Q_{\boldsymbol{w}}). \tag{F.9}$$

Optimizing over $\boldsymbol{w} \in \Delta_m$ proves

$$R_P(\varphi) \le \alpha e^{\Gamma_\infty(P;\mathcal{Q})}. \tag{F.10}$$

Truncating at 1 proves Corollary 13. □

F.3 When is the audit envelope finite?

Proposition 24 (Bounded domination criterion)**.** *The quantity $\Gamma_\infty(P;\mathcal{Q})$ is finite if and only if there exist $\boldsymbol{w} \in \Delta_m$ and $C < \infty$ such that*

$$P(A) \le C\, Q_{\boldsymbol{w}}(A) \qquad \forall A \in \mathcal{F}. \tag{F.11}$$

In that case

$$e^{\Gamma_\infty(P;\mathcal{Q})} = \inf_{\boldsymbol{w}\in\Delta_m} \operatorname*{ess\,sup}_{Q_{\boldsymbol{w}}} \frac{\mathrm{d}P}{\mathrm{d}Q_{\boldsymbol{w}}}, \tag{F.12}$$

where the corresponding term is interpreted as $+\infty$ whenever $P \not\ll Q_{\boldsymbol{w}}$.

Thus finite audit-envelope divergence is exactly a bounded density-ratio coverage condition.

F.4 Continuous families of interventions

The finite family in the main text is not essential.

Let $\mathcal{I}$ be a compact metric intervention space and suppose $i \mapsto q_i$ is continuous as a map from $\mathcal{I}$ into $L^1(\nu)$. Let $\alpha : \mathcal{I} \to (0,1]$ be continuous. Consider

$$V = \sup_{0\le\varphi\le 1} \left\{ \int \varphi p \,\mathrm{d}\nu : \int \varphi q_i \,\mathrm{d}\nu \le \alpha(i)\ \forall i \in \mathcal{I} \right\}. \tag{F.13}$$

Proposition 25 (Continuous-audit dual)**.** *Under the assumptions above,*

$$V = \inf_{\Lambda\in\mathcal{M}_+(\mathcal{I})} \left\{ \int_{\mathcal{I}} \alpha(i)\Lambda(\mathrm{d}i) + \int \left(p(z) - \int_{\mathcal{I}} q_i(z)\Lambda(\mathrm{d}i) \right)_+ \nu(\mathrm{d}z) \right\}, \tag{F.14}$$

where $\mathcal{M}_+(\mathcal{I})$ denotes the finite nonnegative regular Borel measures on $\mathcal{I}$.

Proof. Define the continuous linear operator

$$A : L^\infty(\nu) \to C(\mathcal{I}), \qquad (A\varphi)(i) = \int \varphi q_i \,\mathrm{d}\nu.$$

The dual cone of the positive cone in $C(\mathcal{I})$ is, by the Riesz representation theorem, the cone $\mathcal{M}_+(\mathcal{I})$. Because $\alpha(i) > 0$ continuously on compact $\mathcal{I}$, $\inf_i \alpha(i) > 0$, so $\varphi \equiv 0$ is strictly feasible. The standard separating-hyperplane argument for convex programs in locally convex spaces therefore yields strong duality. Pointwise maximization over $0 \le \varphi \le 1$ produces the positive-part term exactly as in (F.3). □

This form shows that the dual "mixture of audits" becomes a positive measure over intervention protocols.

G Interventional Non-Identifiability

We now formalize adaptive auditing and prove Theorem 14.

G.1 Adaptive audit model

Let $\mathcal{I}$ index admissible audit interventions. At round t, an auditor observes a history

$$H_{t-1} = (I_1, Z_1, Y_1, \ldots, I_{t-1}, Z_{t-1}, Y_{t-1}), \tag{G.1}$$

where I_s is the selected intervention, Z_s the context shown to the agent, and Y_s the safety-relevant behavioral output.

An adaptive auditor chooses I_t according to an arbitrary stochastic kernel

$$\sigma_t(\mathrm{d}i \mid H_{t-1}). \tag{G.2}$$

Conditional on $I_t = i$ and the past, the agent-observable context has law Q_i. This formulation permits randomized, history-dependent intervention selection.

Consider two agents:

$$\varphi_0(z) \equiv 0, \qquad \varphi_1(z) = \mathbf{1}_A(z), \tag{G.3}$$

where A satisfies

$$Q_i(A) = 0 \qquad \forall i \in \mathcal{I}. \tag{G.4}$$

Their behavioral output is the corresponding unsafe indicator.

G.2 Proof of Theorem 14

We prove by induction that the complete audit transcript has the same law under both agents.

At $t = 1$, the auditor's intervention law is identical for both agents. Conditioned on any $I_1 = i$,

$$Q_i(A) = 0,$$

so

$$Y_1^{(0)} = Y_1^{(1)} = 0 \qquad \text{almost surely.}$$

Therefore H_1 has the same distribution for both agents.

Suppose the transcript distributions agree through round $t-1$. Because σ_t is a function only of the transcript, the conditional law of I_t is then identical under both agents. For each realized intervention i,

$$Q_i(A) = 0,$$

so again both behavioral outputs are almost surely zero. Thus the round-t transcript distributions agree.

Induction establishes equality of all finite-round transcript laws. For a countably infinite audit sequence, the two processes have identical finite-dimensional distributions; moreover

$$\mathbb{P}\left(\exists t \geq 1 : Z_t \in A\right) \leq \sum_{t=1}^{\infty} \mathbb{P}(Z_t \in A) = 0,$$

because every conditional probability is zero. Hence the complete behavioral transcript agrees almost surely.

In deployment,

$$R_P(\varphi_0) = 0, \qquad R_P(\varphi_1) = P(A) = \gamma.$$

This proves Theorem 14. □

G.3 Approximate non-identifiability

Exact audit-null events are not required.

Theorem 26 (Approximate adaptive non-identifiability)**.** *Suppose there exists $A \in \mathcal{F}$ and $0 \leq \delta \leq 1$ such that*

$$P(A) = \gamma \tag{G.5}$$

and

$$\sup_{i \in \mathcal{I}} Q_i(A) \leq \delta. \tag{G.6}$$

Consider the two agents in (G.3)*. For any adaptive n-round audit strategy, the total variation distance between their transcript laws satisfies*

$$\boxed{\mathrm{TV}(\mathsf{Law}_0(H_n), \mathsf{Law}_1(H_n)) \leq 1 - (1-\delta)^n \leq n\delta.} \tag{G.7}$$

Their deployment-risk gap remains exactly γ.

Proof. Couple both audit processes using the same auditor randomness and the same context sample at every round until their outputs differ. Conditional on no previous difference and on any current history, the selected intervention I_t is the same in both coupled processes. The probability that the two agents differ at round t is at most

$$Q_{I_t}(A) \leq \delta.$$

Hence the probability that no difference has appeared after n rounds is at least $(1-\delta)^n$. By the coupling characterization of total variation,

$$\mathrm{TV}(\mathsf{Law}_0(H_n), \mathsf{Law}_1(H_n)) \leq 1 - (1-\delta)^n.$$

The final inequality follows from Bernoulli's inequality. □

An immediate testing consequence is that any procedure attempting to decide which agent it is interacting with must satisfy

$$\alpha_{\mathrm{test}} + \beta_{\mathrm{test}} \geq 1 - n\delta \tag{G.8}$$

whenever $n\delta \leq 1$. Thus a deployment-risk gap of size γ may coexist with audit indistinguishability for $n \ll 1/\delta$ adaptive rounds.

G.4 Nonuniform per-round leakage

The same proof gives a more general form.

Corollary 27. *If at round t, conditional on every possible past history, the selected audit law satisfies*

$$Q_t(A \mid H_{t-1}) \leq \delta_t,$$

then

$$\mathrm{TV}(\mathsf{Law}_0(H_n), \mathsf{Law}_1(H_n)) \leq \sum_{t=1}^{n} \delta_t. \tag{G.9}$$

This form is useful when audit coverage grows over time.

H The Audit Paradox

We now prove Theorem 15 and derive an approximate version when audit-specific contexts are highly, but not perfectly, distinguishable.

Let P denote deployment and let $S \perp P$. For $0 \leq \lambda < 1$, define

$$Q_\lambda = (1-\lambda)P + \lambda S. \tag{H.1}$$

H.1 Proof of Theorem 15

For every gate $0 \leq \varphi \leq 1$,

$$R_{Q_\lambda}(\varphi) = (1-\lambda)R_P(\varphi) + \lambda R_S(\varphi). \tag{H.2}$$

Since $R_S(\varphi) \geq 0$,

$$R_{Q_\lambda}(\varphi) \geq (1-\lambda)R_P(\varphi). \tag{H.3}$$

Thus every feasible gate satisfying $R_{Q_\lambda}(\varphi) \leq \alpha$ must obey

$$R_P(\varphi) \leq \frac{\alpha}{1-\lambda}. \tag{H.4}$$

Together with $R_P(\varphi) \leq 1$,

$$\mathsf{ADR}_{P,Q_\lambda}(\alpha) \leq \min\left\{1, \frac{\alpha}{1-\lambda}\right\}. \tag{H.5}$$

Because $P \perp S$, there exists a measurable set B satisfying

$$P(B) = 1, \qquad S(B) = 0. \tag{H.6}$$

Let

$$c = \min\left\{1, \frac{\alpha}{1-\lambda}\right\} \tag{H.7}$$

and define

$$\varphi(z) = c\,\mathbf{1}_B(z). \tag{H.8}$$

Then

$$R_P(\varphi) = c \tag{H.9}$$

and

$$R_{Q_\lambda}(\varphi) = (1-\lambda)c \le \alpha. \tag{H.10}$$

Therefore the upper bound is attained:

$$\boxed{\mathsf{ADR}_{P,Q_\lambda}(\alpha) = \min\left\{1, \frac{\alpha}{1-\lambda}\right\}.} \tag{H.11}$$

This proves Theorem 15. □

H.2 Connection to the D_∞ certificate

The upper bound can also be read directly from Theorem 7. Since

$$Q_\lambda \ge (1-\lambda)P$$

as measures,

$$\frac{\mathrm{d}P}{\mathrm{d}Q_\lambda} \le \frac{1}{1-\lambda}.$$

Moreover, because $P \perp S$, equality holds P-almost surely on a deployment support B with $P(B) = 1$ and $S(B) = 0$. Hence

$$D_\infty(P\|Q_\lambda) = -\log(1-\lambda). \tag{H.12}$$

The D_∞ certificate (C.9) therefore yields

$$\mathsf{ADR}_{P,Q_\lambda}(\alpha) \le \frac{\alpha}{1-\lambda}.$$

Mutual singularity makes this general D_∞ certificate tight.

Thus the Audit Paradox is not separate from the transport framework: it is an extremal case in which adding audit-only mass worsens the likelihood-ratio domination constant from 1 to $(1-\lambda)^{-1}$.

H.3 Approximate audit-specific contexts

Perfect singularity is mathematically clean but unnecessary.

Proposition 28 (Approximate Audit Paradox)**.** *Let $0 \le \lambda < 1$, let $Q_\lambda = (1-\lambda)P + \lambda S$, and let $0 \le \eta, \delta \le 1$. Suppose there exists a measurable set B such that*

$$P(B) \ge 1-\eta, \qquad S(B) \le \delta. \tag{H.13}$$

Then

$$\boxed{\mathsf{ADR}_{P,Q_\lambda}(\alpha) \ge (1-\eta)\min\left\{1, \frac{\alpha}{(1-\lambda)(1-\eta)+\lambda\delta}\right\}.} \tag{H.14}$$

Proof. Write $x = P(B)$ and $y = S(B)$, and take

$$c = \min\left\{1, \frac{\alpha}{(1-\lambda)x + \lambda y}\right\},$$

with the fraction interpreted as $+\infty$ if the denominator is zero. Use the gate $\varphi = c\mathbf{1}_B$. Then

$$R_{Q_\lambda}(\varphi) = c\big((1-\lambda)x + \lambda y\big) \le \alpha,$$

while

$$R_P(\varphi) = x \min\left\{1,\ \frac{\alpha}{(1-\lambda)x + \lambda y}\right\}.$$

Since $y \le \delta$,

$$R_P(\varphi) \ge x \min\left\{1,\ \frac{\alpha}{(1-\lambda)x + \lambda \delta}\right\}.$$

For fixed λ, α, δ, the function

$$f(x) = x \min\left\{1,\ \frac{\alpha}{(1-\lambda)x + \lambda \delta}\right\}$$

is nondecreasing on $[0,1]$: on the branch where the minimum equals 1, $f(x) = x$; on the other branch,

$$f'(x) = \frac{\alpha\lambda\delta}{\big((1-\lambda)x + \lambda\delta\big)^2} \ge 0.$$

Because $x = P(B) \ge 1 - \eta$,

$$R_P(\varphi) \ge (1-\eta) \min\left\{1,\ \frac{\alpha}{(1-\lambda)(1-\eta) + \lambda\delta}\right\},$$

which is the claimed bound. □

As $\eta, \delta \to 0$, Proposition 28 converges to the exact formula in Theorem 15.

H.4 Why adding a separate audit constraint cannot produce the paradox

The distinction made in Section 7 can be stated formally.

Proposition 29 (Monotonicity under additional constraints)**.** *Let $\mathcal{Q}_1 \subseteq \mathcal{Q}_2$ be two collections of audit laws, with the same tolerance assigned to every common audit. Then*

$$\mathsf{ADR}_{P,\mathcal{Q}_2}(\boldsymbol{\alpha}) \le \mathsf{ADR}_{P,\mathcal{Q}_1}(\boldsymbol{\alpha}). \tag{H.15}$$

Proof. The feasible set under $\mathcal{Q}_2$ is a subset of the feasible set under $\mathcal{Q}_1$. □

Therefore Theorem 15 is specifically about replacing a fixed fraction of an *aggregate* evaluation distribution by recognizable audit-only contexts. It is not a claim that independently adding valid constraints makes certification weaker.

I Further Extensions

Here, we record extensions that clarify the scope of the main results and provide several directions for future theoretical work.

I.1 Restricted strategic gate classes

The main paper studies unrestricted measurable gates. Suppose instead that the agent belongs to a restricted class $\mathcal{G} \subseteq [0,1]^{\mathcal{Z}}$.

Define

$$\mathsf{ADR}^{\mathcal{G}}_{P,Q}(\alpha) := \sup_{\substack{\varphi \in \mathcal{G} \\ R_Q(\varphi) \le \alpha}} R_P(\varphi). \tag{I.1}$$

Trivially,

$$\mathsf{ADR}^{\mathcal{G}}_{P,Q}(\alpha) \le \mathsf{ADR}_{P,Q}(\alpha). \tag{I.2}$$

Hence all impossibility results in the main text are worst-case upper envelopes over strategic sophistication.

A particularly interpretable restriction occurs when gates depend only on a representation $T(Z)$. Corollary 20 gives

$$\mathsf{ADR}^{T}_{P,Q}(\alpha) = \mathsf{ADR}_{T_{\#}P,T_{\#}Q}(\alpha). \tag{I.3}$$

Thus evaluation awareness can be studied at any explicitly chosen representation level without altering the theory.

I.2 Connection to optimal evaluation-status classification

Suppose evaluation and deployment are assigned equal prior probability. The Bayes-optimal probability of correctly classifying which environment generated Z is

$$\mathsf{Acc}^{\star} = \frac{1}{2}\left(1 + \mathrm{TV}(P,Q)\right). \tag{I.4}$$

Equivalently,

$$\mathrm{TV}(P,Q) = 2\mathsf{Acc}^{\star} - 1. \tag{I.5}$$

Combining this identity with Theorem 7 yields

$$\boxed{\mathsf{ADR}_{P,Q}(\alpha) \le \min\left\{1,\ \alpha + 2\mathsf{Acc}^{\star} - 1\right\}.} \tag{I.6}$$

Thus an *upper bound* on the optimal ability to classify evaluation versus deployment immediately gives a safety-transport bound. Importantly, observing that one particular classifier performs poorly does not suffice; what matters is the optimal distinguishability available to the agent's information and computational class.

I.3 General bounded harms

Binary unsafe-mode selection is not restrictive for expected bounded harm.

Let

$$h : \mathcal{Z} \to [0,1]$$

denote the conditional expected harm induced by a complete downstream agent policy after observing $Z = z$. Then

$$\mathsf{Risk}_Q = \mathbb{E}_Q[h(Z)], \qquad \mathsf{Risk}_P = \mathbb{E}_P[h(Z)]. \tag{I.7}$$

Since h itself is an admissible randomized gate in Definition 1, every result applies directly:

$$\mathsf{Risk}_Q \le \alpha \quad \Longrightarrow \quad \mathsf{Risk}_P \le \mathsf{ADR}_{P,Q}(\alpha). \tag{I.8}$$

Thus the binary language of "safe mode" and "unsafe mode" can be viewed as an extremal representation of a more general bounded expected-loss functional.

I.4 Dependent long-horizon processes

The i.i.d. assumption in Theorem 10 was used only to obtain convergence of normalized information density. Theorem 22 already gives the relevant general principle.

For path laws Q_T, P_T with $Q_T \ll P_T$, suppose that for some $d \in (0,\infty)$,

$$\frac{1}{T}\log\frac{\mathrm{d}Q_T}{\mathrm{d}P_T}(Z_{1:T}) \longrightarrow d \quad \text{in } Q_T\text{-probability}. \tag{I.9}$$

Then

$$-\frac{1}{T}\log\left(1 - \mathsf{ADR}_{P_T,Q_T}(\alpha)\right) \longrightarrow d. \tag{I.10}$$

Consequently the sequential amplification phenomenon applies to many stationary ergodic and dependent settings whenever an information-density rate exists.

The fundamental object is therefore not independence but accumulated log-likelihood evidence distinguishing audit from deployment.

I.5 Observer invariance in a structural causal model

The statistical setup may be embedded in a causal model containing an environment label

$$E \in \{\text{eval}, \text{dep}\},$$

an evaluator intervention I, an agent-visible information variable Z, and a safety-relevant action A. A schematic causal structure is

$$E \longrightarrow Z \longrightarrow A, \qquad I \longrightarrow Z. \tag{I.11}$$

By definition, Z is intended to contain all information about E available to the agent before acting. If

$$\mathcal{L}(Z \mid E = \text{eval}) = \mathcal{L}(Z \mid E = \text{dep}), \tag{I.12}$$

then $P = Q$, and Corollary 5 gives exact transport.

More generally, a causal intervention is useful for behavioral certification only insofar as the resulting family of Z-distributions satisfies an audit-envelope condition such as Corollary 13. Merely intervening on externally visible task variables does not guarantee transport if the intervention simultaneously changes other variables through which the agent can infer E.

I.6 Unknown deployment distributions

Suppose deployment is known only to belong to an uncertainty class $\mathcal{P}$. The robust audit–deployment curve is

$$\mathsf{ADR}_{\mathcal{P},Q}(\alpha) := \sup_{P \in \mathcal{P}} \mathsf{ADR}_{P,Q}(\alpha). \tag{I.13}$$

If every $P \in \mathcal{P}$ satisfies

$$D_\infty(P \| Q) \le \varepsilon, \tag{I.14}$$

then Theorem 7 immediately gives

$$\mathsf{ADR}_{\mathcal{P},Q}(\alpha) \le \min\{1, e^\varepsilon \alpha\}. \tag{I.15}$$

Thus the framework naturally extends to robust deployment uncertainty.

Conversely, if $\mathcal{P}$ contains a distribution assigning positive mass to a Q-null event, Proposition 4 implies a nonzero irreducible robust risk even at $\alpha = 0$.

I.7 Finite computational power

The main theory gives information-theoretic limits: it assumes the agent can implement the optimal measurable likelihood-ratio gate. If the agent is computationally restricted, one may define a family $\mathcal{G}_{\text{comp}}$ of feasible gates and study

$$\mathsf{ADR}_{P,Q}^{\mathcal{G}_{\text{comp}}}(\alpha). \tag{I.16}$$

This separates two questions:

1. *information availability*: how statistically distinguishable are P and Q?
2. *computational accessibility*: can the agent realize a good discriminator from its available representation?

The unrestricted curve studied in the paper is the correct worst-case object when no reliable computational restriction can be certified.

I.8 A converse perspective on audit design

Theorem 3 suggests an equivalent audit-design objective. Rather than asking only how accurately the evaluator distinguishes safe from unsafe behavior, one can seek an evaluation distribution Q satisfying

$$\mathsf{ADR}_{P,Q}(\alpha) \le \varepsilon \tag{I.17}$$

for a desired deployment-risk target ε.

For a nontrivial target $0 < \alpha \leq \varepsilon \leq 1$, using Theorem 7, sufficient design conditions include

$$\mathrm{TV}(P, Q) \leq \varepsilon - \alpha, \tag{I.18}$$

or

$$D_\infty(P\|Q) \leq \log \frac{\varepsilon}{\alpha}. \tag{I.19}$$

For multiple audits, Corollary 13 instead suggests searching for a mixture $Q_{\boldsymbol{w}}$ satisfying

$$D_\infty(P\|Q_{\boldsymbol{w}}) \leq \log \frac{\varepsilon}{\alpha}. \tag{I.20}$$

These expressions turn the negative theory into a constructive criterion: a behavioral certificate should be accompanied by evidence that the agent-observable audit distribution dominates the intended deployment distribution at the required likelihood-ratio scale.

I.9 Summary of logical dependencies

For clarity, the logical structure of the paper is:

1. Definition 1 introduces the certification object.
2. Theorem 3 identifies it exactly with statistical testing power.
3. Proposition 4 isolates the risk hidden on evaluation-null deployment events.
4. Theorem 7 converts statistical proximity into quantitative deployment certificates.
5. Theorem 8 identifies total variation and D_∞ as the exact worst additive and multiplicative certification defects.
6. Proposition 9 shows that only information available to the agent can increase evaluation awareness.
7. Theorem 10 shows that persistent weak distinctions can amplify exponentially over long horizons.
8. Theorem 12 and Corollary 13 characterize the value of multiple interventions.
9. Theorem 14 identifies the exact failure mode when every admissible intervention misses deployment-relevant contexts.
10. Theorem 15 shows that reallocating a fixed aggregate evaluation budget toward unmistakably audit-only contexts can make the resulting behavioral certificate strictly weaker.

Together these results separate three logically distinct requirements for behavioral safety certification:

$$\boxed{\text{detect unsafe behavior} \quad + \quad \text{cover deployment} \quad + \quad \text{avoid revealing the audit.}} \tag{I.21}$$

The first concerns evaluator sensitivity. The second concerns support and density-ratio coverage. The third concerns the statistical experiment available to an evaluation-aware agent. The main paper studies the latter two even under the idealized assumption that the first is perfect.