Information-Limited Black-Box Safety Evaluation under Latent Deployment Context

Anonymous authors
Paper under double-blind review

Abstract

Black-box safety evaluation observes behavior under an evaluation distribution but aims to estimate risk under deployment, where consequential triggers may be far more common. We study a common-support hidden-set construction with trigger prevalence ϵ in evaluation and $\delta > \epsilon$ in deployment. When non-trigger outputs are identical, an adaptive evaluator's pre-hit queries cannot localize the hidden set except through side information W. For m distinct queries, we show that the true trigger-discovery probability q_m and the independent-side-information baseline q_m^0 satisfy $d(q_m \| q_m^0) \leq I(S; W)$, where d is binary KL divergence. A coupling converts this into a minimax lower bound for deployment-risk estimation. Exact calculations for a tractable hint channel characterize the discovery–information trade-off and its weak- and strong-hint endpoints. Monte Carlo simulations of the construction illustrate the access hierarchy, and a small trained-model diagnostic shows that the coarse approximate-symmetry extension can be vacuous in practice. The result formalizes an information constraint on black-box evaluation without asserting that ordinary adaptive testing is ineffective in general. The contribution is primarily theoretical, supported by controlled simulations and a preliminary trained-model diagnostic. End-to-end validation on language models, including the proposed open-weight protocol, remains an important direction for future work.

1 Introduction

Safety evaluation is usually conducted before deployment: an evaluator queries a model, measures undesirable behavior, and uses those observations to infer whether the deployed system is acceptably safe. The statistical target, however, is deployment risk, while the available observations are generated by an evaluation process. The two processes need not expose the same behavioral regions.

This mismatch is especially important for context-conditioned systems. Backdoor models can behave normally except on rare triggers (Hubinger et al., 2024; Goldwasser et al., 2022); alignment-faking demonstrations show that models can condition behavior on information correlated with training or deployment state (Greenblatt et al., 2024); and interpretability work asks whether internal access can reveal latent behavior not visible from ordinary outputs (Burns et al., 2023; Clymer et al., 2024). These examples motivate a more basic statistical question:

> *How much deployment-relevant behavior can any black-box evaluator discover when the relevant latent structure is rare under evaluation, and how does additional side information change that limit?*

We emphasize at the outset that this is a theoretical contribution. The experiments validate the finite-sample behavior of the construction and exhibit the access hierarchy in controlled worlds; they do not demonstrate that trained models naturally implement the hidden-trigger structure the theorem assumes, and no such demonstration is claimed.

A binary distinction between "adaptive" and "non-adaptive" evaluation is too coarse. Adaptive search can be powerful when pre-hit observations point toward deployment-relevant structure. Our theoretical model isolates the complementary regime in which ordinary non-trigger responses are identical and therefore uninformative about the hidden region; in that regime, any advantage must come from side observations correlated with the latent structure. We therefore organize evaluator access along an information axis.

Latent deployment region. Let $\mathcal{X}$ be finite. A hidden deployment region $R \subset \mathcal{X}$ contains a trigger set $S \subset R$. Evaluation is uniform over $\mathcal{X}$, while deployment is uniform over R. If $|R|/|\mathcal{X}| = r$ and a fraction δ of R is triggering, then

$$\epsilon := \mathbb{P}_{x \sim D_{\text{eval}}}(x \in S) = r\delta, \qquad \mathbb{P}_{x \sim D_{\text{dep}}}(x \in S) = \delta. \tag{1}$$

The same unsafe mode can therefore have tiny evaluation mass ϵ and substantial deployment mass δ without disjoint supports.

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.

Main theoretical result. We randomize S uniformly among all size-$s = \epsilon N$ subsets of $\mathcal{X}$. Let W be any exogenous side-information object jointly distributed with S and available to the evaluator before or during search; its realization is fixed before model interaction. Along the counterfactual all-safe path, the evaluator's distinct query set is a randomized function of (W, U), where U is independent internal randomness. If q_m is trigger-discovery probability under the true joint law and q_m^0 is the discovery probability under the product law $P_S \otimes P_W$, then

$$\boxed{d(q_m \,\|\, q_m^0) \le I(S;W)}. \tag{2}$$

Because non-trigger black-box responses are identical in this construction, the no-side-information case $I(S;W) = 0$ reduces to non-repeating search: all unqueried points remain exchangeable. At positive information budgets, Eq. (2) quantitatively limits how far side-informed search can move discovery probability away from that baseline.

Risk implication. Consider the hypothesis class $\mathcal{F}$ containing a safe model f_0 and every unsafe model $f_1^{(S)}$ indexed by its hidden set S, with deployment-risk separation $\Delta = \delta L$ where L is the trigger loss gap. We prove the minimax lower bound

$$\inf_{\widehat{R}} \sup_{f \in \mathcal{F}} \mathbb{E}|\widehat{R} - R_{\mathrm{dep}}(f)| \ge \frac{\delta L}{2}(1 - q_m), \tag{3}$$

via a Bayesian construction: drawing S uniformly and then choosing f_0 or $f_1^{(S)}$ makes the two transcripts identical under a common-randomness coupling whenever the unsafe evaluator never queries S, and the resulting Bayes risk lower-bounds the minimax risk over $\mathcal{F}$. The randomization is essential — for a fixed, known S the evaluator could simply query a trigger, so the bound relies on the uncertainty induced by drawing S, not on any single fixed pair. Substituting the largest q allowed by Eq. (2) then gives an information-dependent floor. The result separates *query budget* from *information budget*.

Contributions.

1. We give a common-support latent deployment-region construction with explicit evaluation/deployment trigger masses (ϵ, δ).
2. We reduce adaptive black-box trigger discovery under identical pre-trigger observations to a single information-constrained binary event, and from that reduction prove $d(q_m \| q_m^0) \le I(S;W)$ (the data-processing inequality is standard; the reduction is the contribution), making no-side-information adaptivity a zero-information special case rather than a standalone impossibility.
3. We translate discovery limits into deployment-risk lower bounds and give a positive guarantee for direct deployment-context access.
4. We validate the theory with rare-event simulations, a 140-condition (ϵ, δ, m) sweep, an access-hierarchy study, and an analytically tractable side channel that tests the information frontier directly.

2 Related Work

Distribution shift and testing. Classical domain-adaptation bounds relate source and target error under assumptions on distribution discrepancy and hypothesis classes (Ben-David et al., 2010). Sparse property testing and group testing analyze detection or identification of rare structure (Goldreich, 2017; Canonne, 2020; Aldridge et al., 2019). Our target is narrower than full recovery of S: it is the probability of at least one discovery, with an explicit budget on side information rather than only query count. The proof uses standard data processing and two-point decision-theoretic reasoning (Cover & Thomas, 2006; Tsybakov, 2009; Yu, 1997); the contribution is the reduction of adaptive pre-hit interaction to a binary event. Outcome indistinguishability provides a related perspective on limited-access certification (Dwork et al., 2021), but we do not claim a formal reduction.

Hidden behavior and detector access. Context-dependent behavior motivates the construction, including sleeper agents, alignment faking, and computationally hidden backdoors (Hubinger et al., 2024; Greenblatt et al., 2024; Goldwasser et al., 2022). Output-based detectors such as STRIP and Neural Cleanse use additional model evaluations (Gao et al., 2019; Wang et al., 2019); *all such calls must count toward the query budget*, including perturbed or

optimization-generated inputs. Merely reprocessing already observed non-trigger outputs yields no additional information in our symmetric model. Activation-based methods, including ABS and representation probing, access a different measurement channel and are not covered by a bound that omits their information (Liu et al., 2019; Burns et al., 2023; Zou et al., 2023; Clymer et al., 2024).

3 Framework

3.1 Deployment risk and latent region

Let $\mathcal{X}$ be a finite input space of size N. A model f maps an input x to an observable output, and bounded loss $\ell(f(x)) \in [0, 1]$ measures safety-relevant cost. Evaluation and deployment use distributions D_{eval} and D_{dep}, and the target is

$$R_{\text{dep}}(f) = \mathbb{E}_{x \sim D_{\text{dep}}}[\ell(f(x))]. \tag{4}$$

Fix $0 < \epsilon < \delta < 1$. Choose $R \subset \mathcal{X}$ with $|R| = rN$, $r = \epsilon/\delta$, and $S \subset R$ with $|S| = \delta|R| = \epsilon N$. Set $D_{\text{eval}} = \text{Unif}(\mathcal{X})$ and $D_{\text{dep}} = \text{Unif}(R)$. A safe model f_0 has constant loss ℓ_0; an unsafe model f_1 has loss ℓ_0 outside S and ℓ_1 on S, with $L = |\ell_1 - \ell_0| > 0$. We assume throughout that f_0 and f_1 return *identical observable outputs* on every $x \notin S$ — not merely equal loss values — so that no black-box observation outside the trigger set can distinguish the two models or localize S; the models differ only in their observable response on S. This exact observable-output symmetry is the worst-case idealization relaxed in Section 4.3. Their deployment-risk separation is

$$\Delta := |R_{\text{dep}}(f_1) - R_{\text{dep}}(f_0)| = \delta L. \tag{5}$$

3.2 Evaluator and side information

For the adaptive analysis, draw S uniformly among all $s = \epsilon N$ element subsets of $\mathcal{X}$, then draw R uniformly among the size-rN supersets of S. Let W be an arbitrary *exogenous side-information object* generated jointly with (S, R) before model interaction. The evaluator may reveal or process components of W over time, but the side-information realization itself does not depend on the evaluator's model queries or outputs. This covers score vectors, metadata, context features, and precomputed probes; query-dependent active sensing is outside the present theorem.

A signal computed solely from queries and *previously observed non-trigger outputs*, using no new model calls, does not supply independent localization information in this symmetric construction.

Proposition 1 (Post-processing of pre-hit observations)**.** *Let V_t^0 be any sequential signal computed from (W, U) and the query/output history along the* counterfactual all-safe run*, using auxiliary randomness independent of (S, W). For every fixed t, $I(S; (W, V_{1:t}^0)) = I(S; W)$; consequently the discovery bound of Theorem 1 is unchanged if the evaluator uses these signals to select queries. The actual and counterfactual signals coincide until the first trigger hit.*

Proof. The all-safe output history is fixed independently of S. Induction over t makes every pre-hit query and V_t^0 a function of (W, U) and independent auxiliary randomness. Conditional independence given W proves the information equality; the shared-randomness coupling of Appendix A identifies the first-hit events. □

Remark 1 (Scope of the post-processing result)**.** A detector's *additional model queries* are not free post-processing and must be counted in m. Hidden activations or active probes not reducible to existing visible outputs require their own access model; the theorem covers such measurements only when an appropriate channel is included in exogenous W.

The evaluator has independent randomness U and may choose each query from (W, U) and past black-box outputs. Define the counterfactual all-safe run by replacing every model response with the safe response ℓ_0. Before the first trigger discovery, the real and counterfactual paths coincide. Let $Q_{1:m}^0$ denote the distinct counterfactual query set and

$$\mathcal{D}_m := \{Q_{1:m}^0 \cap S \neq \emptyset\}, \qquad q_m := \mathbb{P}(\mathcal{D}_m). \tag{6}$$

Repeated pre-hit queries are never beneficial, so distinct queries can be assumed without loss of discovery probability.

4 Information-Limited Black-Box Discovery

4.1 Zero-information baseline

When $W \perp S$, $Q^0_{1:m}$ is independent of S. By exchangeability, every distinct-query strategy has the same discovery probability

$$q^0_m = 1 - \frac{\binom{N-s}{m}}{\binom{N}{m}}. \tag{7}$$

For passive i.i.d. sampling with replacement, the exact value is $1-(1-\epsilon)^m$. If $m^2/N \to 0$, the without-replacement survival probability satisfies $1 - q^0_m = (1-\epsilon)^m(1+o(1))$, so the two expressions agree up to an $O(m^2/N)$ correction (Appendix A, Sec. A.3) whenever the query budget is small relative to $\sqrt{N}$; we use $1-(1-\epsilon)^m$ for the with-replacement passive baseline (Fig. 1) and the exact hypergeometric q^0_m of Eq. (7) for the without-replacement adaptive-search baseline everywhere else, and flag explicitly which is in use at each point where both could apply.

4.2 Information-limited discovery

Throughout, all logarithms are natural and all information quantities (including $I(S;W)$ and the binary relative entropy below) are measured in nats; this fixes the constant in the Pinsker-based bound Eq. (9). Define binary relative entropy

$$d(a\|b) = a \ln \frac{a}{b} + (1-a) \ln \frac{1-a}{1-b}.$$

Theorem 1 (Information-limited discovery)**.** *Under the symmetric hidden-set model above, where every non-trigger black-box response is the same safe response, any adaptive evaluator with exogenous side information W satisfies*

$$\boxed{d(q_m \,\|\, q^0_m) \le I(S;W)}. \tag{8}$$

Equivalently, q_m cannot exceed the largest $q \ge q^0_m$ satisfying $d(q\|q^0_m) \le I(S;W)$. Pinsker's inequality gives the explicit but looser consequence

$$q_m \le q^0_m + \sqrt{I(S;W)/2}. \tag{9}$$

Proof sketch and novelty boundary. Couple the evaluator to an all-safe run with the same (W, U). The two runs share every pre-hit query, so discovery is the event that the counterfactual query set intersects S. Under $P_S \otimes P_W$, that set is independent of the uniform size-s trigger set and has discovery probability q^0_m. Applying data processing to its hit indicator under the true and product measures gives $d(q_m\|q^0_m) \le D_{\mathrm{KL}}(P_{S,W}\|P_S \otimes P_W) = I(S;W)$. Appendix A supplies the full proof. We are precise about what is and is not novel here. The data-processing inequality itself is entirely standard; it is not the contribution. The contribution is the modeling reduction that makes it applicable: showing that, under identical pre-trigger observations, adaptive black-box trigger discovery collapses to a single information-constrained binary event (whether the counterfactual query set meets S), so that all of an adaptive evaluator's pre-hit interaction is captured by $I(S;W)$ — and then carrying that reduction through to a deployment-risk lower bound (Corollary 1). The bound need not be achievable because the hit indicator can discard information in W. Matching general achievability bounds remain open.

Interpretation. The theorem converts the access hierarchy in this hidden-set model into a single quantitative statement. At $I(S;W) = 0$, all distinct pre-hit query rules have the same discovery probability because ordinary outputs provide no localization signal. At positive information budgets, side-informed search can improve discovery, but the improvement is constrained by the information available about the hidden trigger set. Thus "adaptive evaluation" is not by itself an access assumption; what matters is the information available to the adaptive policy.

4.3 Robustness to approximately, rather than exactly, symmetric responses

Theorem 1 assumes non-trigger responses are *exactly* identical across all $x \notin S$. This is a worst-case idealization: for a trained model, two non-trigger inputs need not produce byte-identical outputs even absent any deliberate trigger structure. The following result shows the guarantee degrades gracefully, rather than catastrophically, when this assumption is relaxed to approximate symmetry.

Proposition 2 (η-robust discovery bound)**.** *Suppose there is a fixed reference output law P_0, independent of (S, W), such that for every hidden set, side-information realization, possible pre-hit history, and queried $x \notin S$, the conditional law of the next response is within η in total variation of P_0. Then every evaluator with exogenous W and m distinct queries satisfies*

$$q_m \;\le\; q_m^0 + m\eta + \sqrt{I(S;W)/2}. \tag{10}$$

At $\eta = 0$ this reduces exactly to the Pinsker consequence Eq. (9) *(not to the sharper binary-KL form of Theorem 1).*

We stress that this bound is informative only when a valid worst-case η yields $m\eta \ll 1$. The diagnostics in Appendix E indicate that under coarse equivalence classes $m\eta$ routinely exceeds one at the query budgets of interest, so Proposition 2 should be read as identifying what an approximate-symmetry guarantee would require, not as a guarantee that currently transfers to trained models.

Proof and operational scope. Couple the actual pre-hit output at each step to the reference law P_0. Each coupling fails with probability at most η; a union bound gives $|q_m - q_m^{\mathrm{ideal}}| \le m\eta$. Applying Pinsker to the ideal process yields Eq. (10); see Appendix D. This is useful only when a valid worst-case η makes $m\eta$ small. Appendix E reports a trained-model *diagnostic* of output variability; its sample mean and percentiles are not certified upper bounds on the supremum required by the proposition.

4.4 Deployment-risk floor

Corollary 1 (Information-dependent deployment-risk floor)**.** *We stress that the minimax is over a* hypothesis class*, not two fixed known models. The evaluator does not know the hidden set S: the class is $\mathcal{F} = \{f_0\} \cup \{f_1^{(S)} : S \subset \mathcal{X},\ |S| = \epsilon N\}$, where $f_1^{(S)}$ is the unsafe model with trigger set S (all sharing the same off-trigger observable output as f_0). The* inf–sup *is over estimators and over this class. In the proof we lower-bound the minimax risk by the Bayes risk of a two-point prior that first draws S uniformly and then chooses f_0 or $f_1^{(S)}$; the randomized draw of S is a proof device, while the resulting bound is a worst-case (minimax) guarantee over $\mathcal{F}$. For the resulting safe/unsafe family with deployment-risk separation $\Delta = \delta L$,*

$$\inf_{\widehat{R}} \sup_{f \in \mathcal{F}} \mathbb{E}|\widehat{R} - R_{\mathrm{dep}}(f)| \ge \frac{\delta L}{2}(1 - q_m). \tag{11}$$

Let q_m^I be the largest $q \in [q_m^0, 1]$ satisfying $d(q\|q_m^0) \le I(S;W)$. Then

$$\boxed{\inf_{\widehat{R}} \sup_{f \in \mathcal{F}} \mathbb{E}|\widehat{R} - R_{\mathrm{dep}}(f)| \ge \frac{\delta L}{2}(1 - q_m^I)} \tag{12}$$

and, by Pinsker,

$$\inf_{\widehat{R}} \sup_{f \in \mathcal{F}} \mathbb{E}|\widehat{R} - R_{\mathrm{dep}}(f)| \ge \frac{\delta L}{2}\left[1 - q_m^0 - \sqrt{I(S;W)/2}\right]_+. \tag{13}$$

Why the factor is $1/2$. The supremum over $\mathcal{F}$ is at least the maximum over any two-element subfamily, so it suffices to exhibit one hard pair: f_0 and a single unsafe model $f_1^{(S)}$ with its hidden set S drawn uniformly. Couple f_0 and $f_1^{(S)}$ using the same (S, R, W, U). Their observed transcripts are identical whenever the $f_1^{(S)}$ path does not query S, hence $\mathrm{TV}(P_0, P_1) \le q_m$. Under an equal prior on the two scalar risks, Bayes absolute loss is exactly $\Delta(1 - \mathrm{TV}(P_0, P_1))/2$; minimax risk over $\mathcal{F}$ is at least this Bayes risk. This yields Eq. (11). The argument does not require side information to be uninformative about S; it only requires the same side-channel law under both model indices. The deployment region R is likewise unknown to the evaluator: in the construction R is drawn uniformly among the size-rN supersets of S (Sec. 3.2), and R_{dep} is defined relative to that latent R, so the estimator $\widehat{R}$ never observes R and the bound holds against the joint prior over (S, R).

Because $q_m \le q_m^I$ for every evaluator, replacing q_m with q_m^I in a decreasing lower bound requires no assumption that the information frontier is attainable.

5 What Additional Access Buys

5.1 Deployment-context access

Suppose the evaluator can sample from D_{dep} rather than only D_{eval}. The effective trigger prevalence becomes δ.

Proposition 3 (Deployment-context upper bound)**.** *With m i.i.d. deployment-context samples, trigger detection probability is $1-(1-\delta)^m$. The sample-mean loss estimator satisfies*

$$\mathbb{P}(|\widehat{R} - R_{\mathrm{dep}}| > \varepsilon_R) \le 2\exp(-2m\varepsilon_R^2). \tag{14}$$

The qualitative change is that sample complexity no longer depends on the small evaluation mass ϵ; this proposition is a direct consequence of the construction and Hoeffding's inequality. We include it mainly to make the contrast with q_m^0 explicit, and revisit it in Section 7 as the natural point of comparison with Corollary 1 rather than as an independent technical contribution.

Internal-access extension. If deployment-context inputs are available but the deployment-relevant latent state itself is not directly observed, a calibrated internal probe can provide an additional measurement channel. Appendix G gives a simple binary-channel analysis and reports the corresponding synthetic robustness experiment. We keep this extension supplementary because direct deployment-loss observations make the sample-mean estimator preferable whenever those losses are available.

6 Experiments

The experiments stress-test the finite-sample implications of the construction; they do not claim that typical trained models naturally implement hidden triggers. The hint-channel grid is specified below and further design parameters appear in Appendix H. The synthetic simulation statistics reported below, and the digit-classifier diagnostic of Appendix E, are regenerated directly by the supplied notebook from the seeds in Appendix H. The notebook's optional real-model section concerns only the unexecuted open-weight study of Appendix I, not Appendix E.

6.1 Sanity check: rare-event law

The default world uses $N = 100{,}000$, $r = 0.02$, and $\delta = 0.20$, giving $\epsilon = 0.004$. Safe and trigger losses are $\ell_0 = 0.02$ and $\ell_1 = 0.80$, so $R_{\mathrm{eval}} = 0.02312$ and $R_{\mathrm{dep}} = 0.176$. Figure 1 shows the exact i.i.d. and distinct-query discovery laws for the default construction; the figure is analytic rather than a Monte Carlo scatter, and the corresponding simulation points are regenerated by the supplied notebook. Across 140 (ϵ, δ, m) configurations, the mean absolute deviation from $1 - e^{-m\epsilon}$ is 0.0089 (regenerated by the supplied notebook). We note that evaluation-only trigger discovery depends only on the evaluation prevalence ϵ and the query count m; the deployment prevalence δ does not affect the generated observations. The four δ settings in the grid therefore repeat the same discovery experiment at each (ϵ, m) rather than probing four distinct discovery laws, and the sweep effectively spans the five ϵ values and seven query budgets. This is a construction-validation check, not an independent empirical finding: $1-(1-\epsilon)^m$ is the exact law for i.i.d. sampling (Sec. 4.1), so the Monte Carlo run confirms the simulator is implemented correctly rather than testing a hypothesis.

6.2 Access hierarchy

At $m = 20$, passive evaluation and no-side-information non-repeating search detect the trigger with probabilities 0.0833 and 0.0775, respectively, both consistent with the exact baseline law $1-(1-\epsilon)^m = 0.0770$. A Gaussian region score $s(x) = \mathbf{1}\{x \in R\} + \xi$, $\xi \sim \mathcal{N}(0, \sigma^2)$, used to prioritize a 5000-point candidate pool raises detection to 0.265, 0.584, and 0.958 for weak ($\sigma = 2.0$), moderate ($\sigma = 1.0$), and strong ($\sigma = 0.5$) side information; direct deployment-context sampling reaches 0.989 (Figure 3). Table 1 preserves the manuscript's reported 95% cluster-bootstrap intervals over hidden worlds, treating worlds rather than within-world repetitions as independent units. This table illustrates that the access hierarchy predicted by Theorem 1 is large in magnitude in these constructed worlds. Because only 20 hidden worlds are used, the reported cluster-bootstrap intervals are indicative rather than precise (Appendix F); we therefore rely on the ordering of the point estimates, which spans an order of magnitude, rather than on the exact interval endpoints.

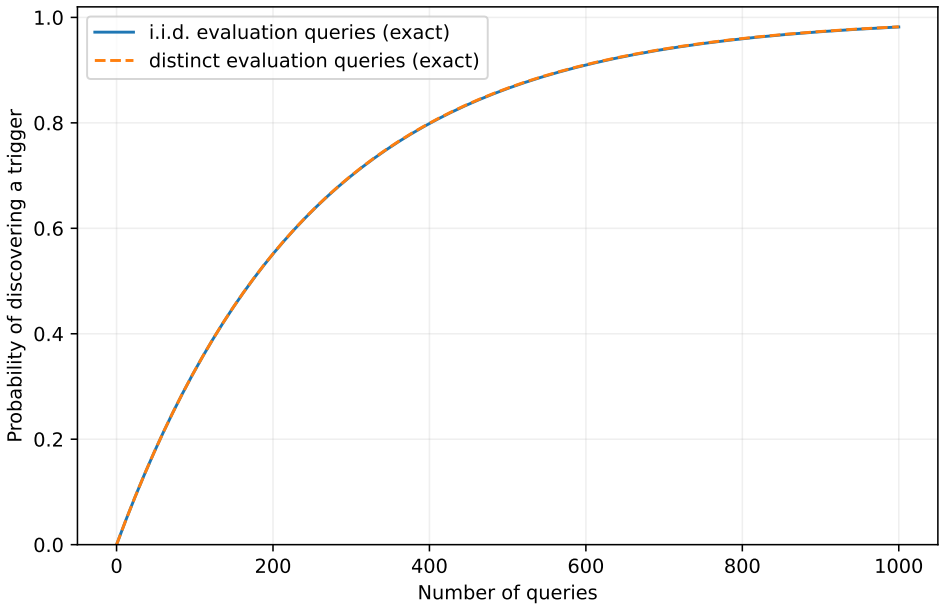

Figure 1: Analytic detection curves for the default construction ($N = 100{,}000$, $\epsilon = 0.004$). These exact laws do not display Monte Carlo observations or confidence intervals.

Table 1: Trigger detection at $m = 20$ across 20 hidden worlds and 4000 total repetitions per method, regenerated by the supplied notebook (seed `20260906`). Intervals are cluster-bootstrap intervals over hidden worlds. The passive and non-repeating baselines are consistent with the exact discovery law $1 - (1 - \epsilon)^m = 0.0770$ at $\epsilon = 0.004$, whose value their intervals contain.

Access regime	Detection	95% CI
Passive evaluation	0.0833	[0.0745, 0.0935]
No-side-information non-repeating	0.0775	[0.0710, 0.0838]
Side information, weak	0.2647	[0.2550, 0.2743]
Side information, moderate	0.5837	[0.5755, 0.5915]
Side information, strong	0.9575	[0.9495, 0.9652]
Deployment-context access	0.9892	[0.9865, 0.9920]

6.3 Direct test of the information frontier

The Gaussian score above is useful operationally but makes $I(S; W)$ cumbersome to compute exactly. We therefore add an analytically tractable hint channel. Conditional on S, side information $W \in \mathcal{X}$ is generated by choosing a point uniformly from S with probability λ and otherwise choosing an independent uniform point from $\mathcal{X}$. Its marginal is uniform and

$$I(S; W) = sa \log(Na) + (N - s)b \log(Nb), \tag{15}$$

where $a = (1 - \lambda)/N + \lambda/s$ and $b = (1 - \lambda)/N$; a self-contained derivation is given as Proposition 4 in Appendix B. The evaluator queries W first and uses uniformly random distinct queries thereafter. We vary $\lambda \in \{0, .001, .005, .01, .02, .05, .1, .2, .4, .6\}$ and $m \in \{5, 20, 100\}$.

Figure 2 plots exact discovery against the information-theoretic upper frontier computed from $d(q\|q_m^0) \le I(S; W)$. Across 30 (m, λ) conditions, the 27 positive-information ratios $d(q_m\|q_m^0)/I(S; W)$ range from 0.0088 to 0.6493. Their being below one is guaranteed by the theorem; their *distance* from one diagnoses its looseness. Figure 5 shows the underlying $d(q_m\|q_m^0)$ against $I(S; W)$ with the $d = I$ boundary, confirming every condition lies on or below it. On the tested grid, the ratio increases with λ at fixed m and decreases with m at fixed λ. Appendix C gives analytic limits at $\lambda \to 0^+$ and $\lambda = 1$, but does not establish monotonicity between them for arbitrary parameters. Both endpoint ratios tend to zero as $m \to \infty$ with ϵ fixed. With m fixed and $\epsilon \to 0$, however, the respective limits are $1/m$ and 1; arbitrarily

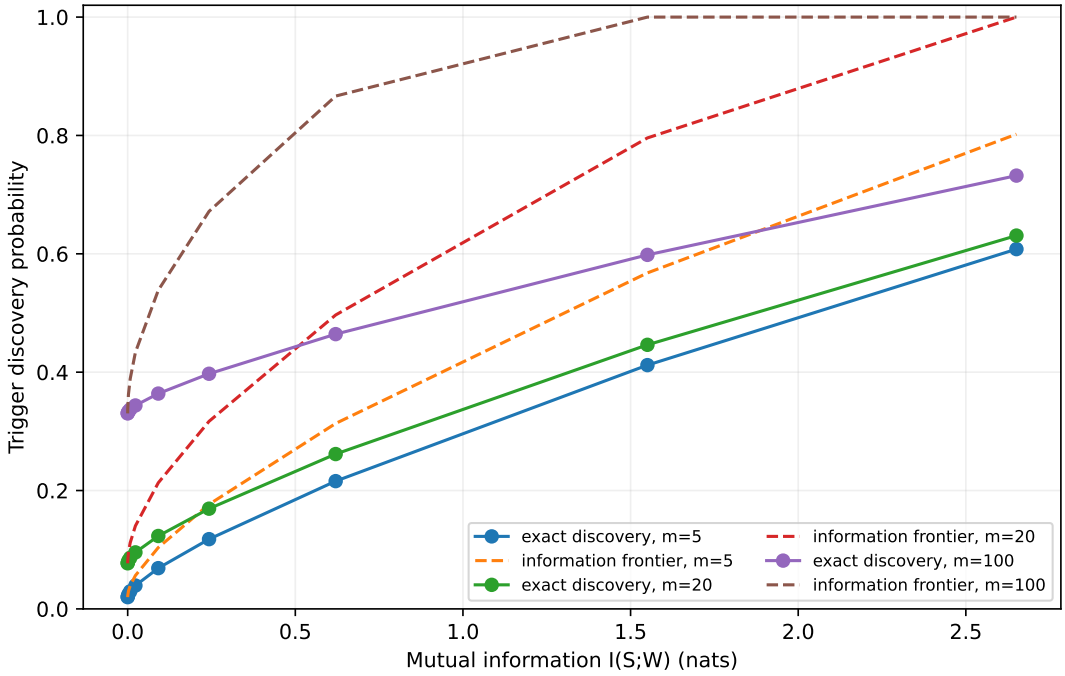

Figure 2: Information-limited discovery. Solid curves show exact trigger discovery under the analytic hint channel; dashed curves show the largest q allowed by $d(q\|q_m^0) \le I(S;W)$. All observed conditions lie below the theorem frontier.

poor relative tightness therefore follows from growing m, *not* from rarity alone at fixed m.

6.4 Deployment-risk estimation

At $m = 3000$, evaluation-only risk estimation remains near R_{eval}, with RMSE 0.15292 (95% bootstrap CI $[0.15287, 0.15298]$). In contrast, direct deployment-context sampling has RMSE 0.00575 $[0.00551, 0.00600]$. These values and intervals are regenerated deterministically by the supplied notebook (seed `20260906`). The contrast is not a claim that deployment samples are always available; it demonstrates the information value of replacing the rare evaluation prevalence ϵ by the deployment prevalence δ. A calibrated-probe extension is reported in Appendix G for the setting where the relevant latent state is not directly observed.

7 Discussion

What the theorem does and does not say. The result is a worst-case statement for a symmetric hidden-set model in which non-trigger responses are deliberately uninformative about the hidden region. It does not say that adaptive evaluation is generally ineffective. If ordinary outputs, tool traces, gradients, metadata, or other observations localize deployment-relevant structure, adaptivity can exploit them; such information is outside the no-side-information baseline and must be represented through an informative channel. Equation (8) quantifies the contribution of exogenous side information in the model studied here.

Practical scope. Two boundaries of the construction are worth making explicit. First, practical adaptive red-teaming methods such as PAIR (Chao et al., 2023) and Tree-of-Attacks-style search (Mehrotra et al., 2023) exploit exactly the informative pre-trigger responses and language-model priors that our symmetric construction removes by assumption; our zero-information baseline is therefore not intended to describe such methods in general, and it does not predict that they are ineffective. Where those methods succeed, they do so through pre-hit signal that our model represents, if at all, only through an informative side-channel W. Second, the exact hypergeometric discovery law and the $1-(1-\epsilon)^m$ approximation rely on the uniform hidden-set construction over a finite input space; they do not automatically extend to arbitrary non-uniform or continuous input distributions, for which the relevant prevalence and the corresponding

discovery law would have to be redefined. The information inequality of Theorem 1 depends only on $I(S;W)$ and the symmetric-response assumption and is not tied to uniformity, but the specific numerical baselines are.

Why information rather than query count matters. A larger query budget can reduce rare-event miss probability only on the $1/\epsilon$ scale. Side information changes a different resource: it changes how much the evaluator can concentrate queries on the latent region. The main theorem separates these two resources and suggests that evaluation reports should state not only the number of queries, but also what deployment-correlated metadata, context samples, internal probes, or other side channels were available.

Tightness and interpretation. For the hint channel, Appendix C proves closed-form weak- and strong-hint endpoint ratios in the $N \to \infty$ limit. Both decrease as the query budget grows, and both tend to zero at fixed ϵ as $m \to \infty$. The observed interpolation is monotone on the reported grid but remains unproved in general. The absence of a uniform multiplicative tightness guarantee should not be confused with invalidity of the KL discovery bound.

Limitations. The access-hierarchy experiment uses constructed hidden triggers and does not establish their prevalence in naturally trained frontier models. Appendix E contains a small trained-model diagnostic (a digit classifier), but it does not test an end-to-end context-triggered LLM. Combined with the vacuity of the $m\eta$ correction under that diagnostic (Appendix E) and the unexecuted open-weight protocol (Appendix I), the theory-to-practice transfer is the paper's principal open empirical question, not a demonstrated result. Extending the theory to informative pre-hit responses and query-dependent internal probes is future work.

8 Conclusion

Black-box safety evaluation is constrained not only by query budget but by information about deployment-relevant latent structure. In a common-support hidden-region construction, zero-information adaptivity reduces to rare-event search; with side information, the achievable discovery improvement is bounded by $d(q_m \| q_m^0) \leq I(S;W)$. Deployment-context access removes the evaluation-only floor in the construction by changing the effective trigger prevalence from ϵ to δ. This framing turns evaluator access assumptions into explicit statistical quantities and clarifies when more querying can, and cannot, substitute for richer observability. The main open question is whether the bound is achievable, or close to achievable, for natural side-information channels; we view this, together with real-model validation, as the priority for follow-up work.

Reproducibility Statement

The symmetric-model assumptions and complete proofs are given in Sections 3–5 and Appendices A–C. All analytic results (Theorem 1, Corollary 1, Propositions 3–4) and the exact finite-N hint-channel quantities are computed in closed form by the supplied notebook (`blackbox_information_limits.ipynb`); the 27 positive-information tightness ratios it produces match those reported in Section 6 exactly. The access-hierarchy detection rates and cluster-bootstrap confidence intervals (Table 1), the 140-condition construction-validation sweep (Section 6), and the deployment-risk RMSE point estimates *and* their bootstrap confidence intervals (Section 6) are all regenerated by the same notebook from the fixed seed recorded in Appendix H; with that seed the reported point estimates and intervals are reproduced deterministically, the passive and non-repeating baselines match the exact discovery law $1-(1-\epsilon)^m$, and a final cell asserts equality with the values printed in the paper. The notebook also writes the five figures included in the manuscript. Reproducibility is therefore through re-execution of the supplied notebook rather than through archived output files.

The single trained-model diagnostic retained in Appendix E (the `scikit-learn` digit classifier of Table 4, with its worked $m\eta$ example in Table 5) is also regenerated by the notebook from the same seed, including its train/test split, hyperparameters, and η analysis. Two supplementary items are deliberately *not* regenerated by the notebook, and we state this explicitly. The open-weight protocol of Appendix I was not executed, and the notebook's optional `RUN_REAL_MODEL` section only scaffolds that future experiment. The illustrative cluster-bootstrap coverage simulation of Appendix F is a generic sensitivity analysis, not a rerun of the paper's simulator; it is fully specified by the parameters given there but is not part of the notebook. The language-model case in Appendix E is discussed qualitatively and carries no numerical claims.

Ethics Statement

The work studies limitations of safety evaluation and could in principle inform attempts to hide undesirable behavior. We mitigate this risk by focusing on statistical access assumptions rather than operational evasion techniques, and by using synthetic or controlled trigger constructions. The intended use is defensive: to improve evaluation design and make evidential limitations explicit.

AI Use Statement

Generative AI tools were used to assist with language editing, code scaffolding, and exploratory checking of proof structure. The authors remain responsible for all mathematical claims, experimental results, citations, and final text; reported results were generated by executable code rather than by AI-generated numerical claims.

References

Matthew Aldridge, Oliver Johnson, and Jonathan Scarlett. Group testing: an information theory perspective. *Foundations and Trends in Communications and Information Theory*, 15(3–4):196–392, 2019.

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(1–2):151–175, 2010.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations*, 2023.

A. Colin Cameron, Jonah B. Gelbach, and Douglas L. Miller. Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics*, 90(3):414–427, 2008.

Clément L. Canonne. A survey on distribution testing: Your data is big. But is it blue? *Theory of Computing Library, Graduate Surveys*, 9:1–100, 2020.

Joshua Clymer, Caden Juang, and Severin Field. Poser: Unmasking alignment faking LLMs by manipulating their internals. arXiv:2405.05466, 2024.

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

Yansong Gao, Change Xu, Derui Wang, Shiping Chen, Damith C. Ranasinghe, and Surya Nepal. STRIP: A defence against trojan attacks on deep neural networks. In *Annual Computer Security Applications Conference (ACSAC)*, pp. 113–125, 2019.

Yingqi Liu, Wen-Chuan Lee, Guanhong Tao, Shiqing Ma, Yousra Aafer, and Xiangyu Zhang. ABS: Scanning neural networks for back-doors by artificial brain stimulation. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 1265–1282, 2019.

Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y. Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *IEEE Symposium on Security and Privacy (S&P)*, pp. 707–723, 2019.

Cynthia Dwork, Michael P. Kim, Omer Reingold, Guy N. Rothblum, and Gal Yona. Outcome indistinguishability. In *Proceedings of STOC*, pp. 1095–1108, 2021.

Oded Goldreich. *Introduction to Property Testing*. Cambridge University Press, 2017.

Shafi Goldwasser, Michael P. Kim, Vinod Vaikuntanathan, and Or Zamir. Planting undetectable backdoors in machine learning models. In *IEEE Symposium on Foundations of Computer Science (FOCS)*, pp. 931–942, 2022.

Ryan Greenblatt et al. Alignment faking in large language models. arXiv:2412.14093, 2024.

Evan Hubinger et al. Sleeper agents: Training deceptive LLMs that persist through safety training. arXiv:2401.05566, 2024.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

Bin Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pp. 423–435. Springer, 1997.

Andy Zou et al. Representation engineering: A top-down approach to AI transparency. arXiv:2310.01405, 2023.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. arXiv:2310.08419, 2023.

Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum Anderson, Yaron Singer, and Amin Karbasi. Tree of attacks: Jailbreaking black-box LLMs automatically. arXiv:2312.02119, 2023.

A Proofs

A.1 Distinct-query reduction

Lemma 1. *For any evaluator there exists another evaluator with weakly larger trigger-discovery probability against every hidden S that never repeats a query before its first trigger hit.*

Proof. Maintain a virtual run of the original evaluator using its original query history. On a virtual pre-hit query that has already been tested, test an unused point externally and feed the virtual run the known safe response to its repeated query. If a later virtual query was sampled by an earlier extra external query, it is already known to be a miss unless discovery has terminated; again test a new point and feed the virtual run its known safe response. Otherwise execute the requested point. Coupled with the same hidden set and evaluator seed, every distinct point the original would have tested before its first hit is thus tested no later than its virtual step; additional external queries can only help. The external strategy never repeats a pre-hit query. □

A.2 Coincidence with the all-safe path

Run the real evaluator and its counterfactual all-safe copy with the same (S, W, U) — in particular, both runs share the identical realization of the internal randomness U, so that query selection is the same deterministic function of identical inputs $(W, U, \text{history})$ in both runs whenever the histories agree. Let τ and τ^0 be the first times their respective queries enter S. For every $t < \tau$, the real model returns ℓ_0, so both procedures have received identical information and choose the same query. Consequently $Q_t = Q_t^0$ for all $t \le \tau$, which implies $\tau = \tau^0$. Hence discovery is exactly the event $Q_{1:m}^0 \cap S \neq \emptyset$.

A.3 Proof of the zero-information baseline

Under $P_S \otimes P_W$, the counterfactual query set is independent of S. Condition on any m distinct queried points. Since S is uniform among all size-s subsets,

$$\mathbb{P}(Q_{1:m}^0 \cap S = \emptyset) = \frac{\binom{N-m}{s}}{\binom{N}{s}} = \frac{\binom{N-s}{m}}{\binom{N}{m}}.$$

The expression is independent of the realized query set, proving Eq. (7). Further,

$$\log \frac{\binom{N-s}{m}}{\binom{N}{m}} = \sum_{j=0}^{m-1} \log\left(1 - \frac{s}{N-j}\right) = m \log(1-\epsilon) + O(m^2/N),$$

so the multiplicative approximation by $(1-\epsilon)^m$ holds whenever $m^2/N \to 0$.

A.4 Proof of Theorem 1

Let K be the randomized map from (S, W) to the hit indicator induced by the evaluator and the independent randomness U. Under the true joint law $P_{S,W}$, its output is $\mathrm{Ber}(q_m)$. Under the product law $P_S \otimes P_W$, W is independent of S, so its output is $\mathrm{Ber}(q_m^0)$. The data-processing inequality gives

$$\begin{aligned} d(q_m \| q_m^0) &= D_{\mathrm{KL}}(\mathrm{Ber}(q_m) \| \mathrm{Ber}(q_m^0)) && (16)\\ &\le D_{\mathrm{KL}}(P_{S,W} \| P_S \otimes P_W) && (17)\\ &= I(S; W). && (18) \end{aligned}$$

For $q > q_m^0$, $d(q\|q_m^0)$ is strictly increasing, giving the implicit frontier. Pinsker's inequality $d(a\|b) \ge 2(a-b)^2$ yields Eq. (9).

A.5 Proof of Corollary 1

It suffices to lower-bound the minimax risk over $\mathcal{F}$ by the Bayes risk of a two-point sub-problem. Draw S (and its region R) as in Section 3.2, place an equal prior on $J \in \{0, 1\}$ selecting f_0 or the unsafe model $f_1^{(S)}$ with that hidden set, and couple the two transcript-generating procedures using the same (S, R, W, U). If the f_1 path never queries S, both models return ℓ_0 on every query, so the full observed transcripts coincide. Because non-trigger responses are identical under both f_0 and f_1 by construction (Sec. 4.1), the event "the f_1 path never queries S" is exactly the event that the two full transcripts — not merely the hit indicator — agree on every coordinate; this is where the symmetric-response assumption is used, and it is what licenses passing from a statement about the hit indicator to a statement about TV of the full transcript laws. Therefore the coupling mismatch probability is at most q_m, and the coupling inequality gives

$$\mathrm{TV}(P_0, P_1) \le q_m.$$

For an equal prior on two scalar parameters separated by $\Delta = \delta L$, the Bayes risk under absolute loss is $\Delta(1 - \mathrm{TV}(P_0, P_1))/2$. Since the minimax risk over $\mathcal{F}$ dominates the Bayes risk of this two-point sub-problem under any prior,

$$\inf_{\widehat{R}} \sup_{f \in \mathcal{F}} \mathbb{E}|\widehat{R} - R_{\mathrm{dep}}(f)| \ge \frac{\delta L}{2}(1 - q_m).$$

Substituting the maximum q_m allowed by Theorem 1 proves Eqs. (12)–(13).

A.6 Proof of Proposition 3

Under D_{dep}, each sample is in S with probability δ, so trigger detection is $1 - (1 - \delta)^m$. The loss is bounded in $[0, 1]$; Hoeffding's inequality gives the concentration bound.

B Derivation of the Hint-Channel Mutual Information

This appendix derives Eq. (15) as a standalone proposition.

Proposition 4. *Fix $S \subset \mathcal{X}$ with $|S| = s$, and let $W \in \mathcal{X}$ be generated conditional on S by drawing W uniformly from S with probability λ and, with probability $1 - \lambda$, drawing W uniformly from $\mathcal{X}$. Let S itself be uniform among size-s subsets of $\mathcal{X}$. Then the marginal of W is uniform on $\mathcal{X}$, and*

$$I(S; W) = sa \log(Na) + (N - s)b \log(Nb),$$

where $a = \mathbb{P}(W = w \mid w \in S) = (1 - \lambda)/N + \lambda/s$ and $b = \mathbb{P}(W = w \mid w \notin S) = (1 - \lambda)/N$.

Proof. Marginal of W. By symmetry over the choice of S, every $w \in \mathcal{X}$ is equally likely to be the realized W, so W is uniform on $\mathcal{X}$ and $H(W) = \log N$.

Conditional law of W given S. Fix a realization of S. For $w \in S$: w is returned either because the "pick from S" branch (probability λ) selected it, which happens with probability $1/s$, or because the uniform-over-$\mathcal{X}$ branch (probability $1 - \lambda$) selected it, with probability $1/N$. These are mutually exclusive branches, so

$$a := \mathbb{P}(W = w \mid S) = \frac{1 - \lambda}{N} + \frac{\lambda}{s}, \qquad w \in S.$$

For $w \notin S$: only the uniform branch can produce w, so

$$b := \mathbb{P}(W = w \mid S) = \frac{1 - \lambda}{N}, \qquad w \notin S.$$

There are s points with probability a and $N - s$ points with probability b under this conditional law, and indeed $sa + (N - s)b = \lambda + (1 - \lambda) = 1$, confirming it is a valid distribution.

Conditional entropy. Since a, b do not depend on which size-s set S is (only on s), $H(W \mid S = S_0)$ is the same value for every realization S_0, namely

$$H(W \mid S) = -sa \log a - (N - s)b \log b.$$

Mutual information.

$$I(S; W) = H(W) - H(W \mid S) = \log N + sa \log a + (N - s)b \log b.$$

Using $sa + (N-s)b = 1$,

$$\log N = (sa + (N-s)b)\log N,$$

so

$$I(S;W) = sa\log a + sa\log N + (N-s)b\log b + (N-s)b\log N = sa\log(Na) + (N-s)b\log(Nb),$$

which is Eq. (15). □

C Asymptotic Tightness of the Bound for the Hint-Channel Family

This appendix proves the floor and ceiling stated in Section 7, giving a closed-form, non-heuristic partial answer to the achievability question raised in Section 4.2, restricted to the analytic hint-channel family of Section 6.

Proposition 5 (Asymptotic tightness envelope)**.** *Fix $\epsilon \in (0,1)$, $\lambda \in (0,1]$, and $m \geq 1$. Let $N \to \infty$ with $s = \epsilon N$ (rounded) and λ, ϵ, m held fixed. Under the hint-channel construction of Section 6 (Proposition 4), the ratio*

$$\rho_N(\lambda, \epsilon, m) := \frac{d(q_m \| q_m^0)}{I(S;W)}$$

converges to a finite limit $\rho_\infty(\lambda, \epsilon, m)$. Its weak- and strong-hint endpoints are

$$\lim_{\lambda \to 0^+} \rho_\infty(\lambda, \epsilon, m) = \frac{\epsilon(1-\epsilon)^{m-1}}{1-(1-\epsilon)^m}, \qquad \rho_\infty(1, \epsilon, m) = \frac{-\log\big(1-(1-\epsilon)^m\big)}{\log(1/\epsilon)}.$$

Both endpoint expressions strictly decrease with m and tend to zero as $m \to \infty$ at fixed ϵ. At fixed m, their respective limits as $\epsilon \to 0$ are instead $1/m$ and 1. Therefore no universal constant $c > 0$ lower-bounds $d(q_m \| q_m^0)/I(S;W)$ across all choices of this family's parameters. Numerical, rather than analytic, evidence supports monotonic interpolation between the endpoints on the tested parameter grids.

Proof. Setup. Write $a = (1-\lambda)/N + \lambda/s$, $b = (1-\lambda)/N$ as in Proposition 4, so $Na \to (1-\lambda) + \lambda/\epsilon =: A$ and $Nb \to 1-\lambda =: B$ as $N \to \infty$ with $s/N = \epsilon$ fixed (using $s \to \infty$ so $\lambda/s \to 0$). Since $sa = \epsilon(1-\lambda) + \lambda$ and $(N-s)b = (1-\epsilon)(1-\lambda)$ are already free of N, Proposition 4 gives

$$I(S;W) \;\longrightarrow\; I_\infty(\lambda, \epsilon) := [\epsilon(1-\lambda) + \lambda]\log A + (1-\epsilon)(1-\lambda)\log B,$$

a finite limit for every $\lambda \in (0,1]$ (interpreting $B \log B = 0$ at $\lambda = 1$).

Discovery probability. The exact formula $q_m = 1 - (1-q_1)\binom{N-1-s}{m-1}/\binom{N-1}{m-1}$, with $q_1 = \lambda + (1-\lambda)\epsilon$ independent of N, converges as $N \to \infty$ (fixed ϵ, m) to

$$q_m \;\longrightarrow\; q_m^{(\infty)}(\lambda, \epsilon) := 1 - (1-q_1)(1-\epsilon)^{m-1},$$

by the same hypergeometric-to-binomial limit used in Appendix A for q_m^0, which itself converges to $q_m^{0,(\infty)}(\epsilon) := 1-(1-\epsilon)^m$. Both limits are finite and continuous in λ, ϵ, so $d(q_m \| q_m^0) \to d(q_m^{(\infty)} \| q_m^{0,(\infty)}) =: D_\infty(\lambda, \epsilon, m)$, and $\rho_N \to \rho_\infty := D_\infty / I_\infty$.

Endpoint at $\lambda = 1$. At $\lambda = 1$: $q_1 = 1$, so $q_m^{(\infty)} = 1$ and $D_\infty = d(1 \| q_m^{0,(\infty)}) = -\log q_m^{0,(\infty)} = -\log(1-(1-\epsilon)^m)$. Also $A = 1/\epsilon$, $B = 0$ (with $B \log B \to 0$), so $I_\infty(1, \epsilon) = \log(1/\epsilon)$. This gives the stated ceiling.

Endpoint at $\lambda \to 0^+$. Both D_∞ and I_∞ vanish as $\lambda \to 0$ (at $\lambda = 0$, $W \perp S$). A Taylor expansion of the stated analytic formulas gives, to leading order in λ,

$$q_m^{(\infty)} - q_m^{0,(\infty)} = (1-\epsilon)^m\,\lambda + O(\lambda^2), \qquad I_\infty(\lambda, \epsilon) = \frac{1-\epsilon}{2\epsilon}\,\lambda^2 + O(\lambda^3).$$

Using the local quadratic approximation $d(a\|b) \approx (a-b)^2/[2b(1-b)]$ for $a \to b$, valid since $D_\infty \to 0$ as $\lambda \to 0$,

$$D_\infty(\lambda, \epsilon, m) = \frac{(1-\epsilon)^{2m}\lambda^2}{2\,q_m^{0,(\infty)}(1-q_m^{0,(\infty)})} + O(\lambda^3).$$

Dividing, the λ^2 terms cancel and the limit is

$$\lim_{\lambda\to 0^+}\rho_\infty = \frac{(1-\epsilon)^{2m}/\big[q_m^{0,(\infty)}(1-q_m^{0,(\infty)})\big]}{(1-\epsilon)/\epsilon} = \frac{\epsilon(1-\epsilon)^{2m-1}}{\big[1-(1-\epsilon)^m\big](1-\epsilon)^m} = \frac{\epsilon(1-\epsilon)^{m-1}}{1-(1-\epsilon)^m},$$

using $1-q_m^{0,(\infty)} = (1-\epsilon)^m$. This is the stated floor.

What is proved about the ratio. The numerator $D_\infty(\lambda) = d(q_m^{(\infty)}(\lambda)\|q_m^{0,(\infty)})$ strictly increases with λ, since $q_m^{(\infty)} = q_m^{0,(\infty)} + \lambda(1-\epsilon)^m$. This does *not* establish monotonicity of the ratio D_∞/I_∞: both numerator and denominator increase. We have not proved monotonicity for arbitrary (ϵ, m), although it holds at each reported numerical checkpoint. The next two lemmas establish only the monotonicity of the *endpoint formulas* in m.

Lemma 2 (Monotonicity of the floor in m)**.** *For fixed $\epsilon \in (0,1)$, $g(m) := \epsilon(1-\epsilon)^{m-1}/[1-(1-\epsilon)^m]$ is strictly decreasing in m for $m \geq 1$.*

Proof. It suffices to show $g(m)/g(m+1) > 1$ for every $m \geq 1$. Directly,

$$\frac{g(m)}{g(m+1)} = \frac{1}{1-\epsilon}\cdot\frac{1-(1-\epsilon)^{m+1}}{1-(1-\epsilon)^m}.$$

We claim the fraction on the right exceeds $1-\epsilon$, which gives $g(m)/g(m+1) > 1$. Indeed,

$$\big[1-(1-\epsilon)^{m+1}\big] - (1-\epsilon)\big[1-(1-\epsilon)^m\big] = 1-(1-\epsilon)^{m+1} - (1-\epsilon) + (1-\epsilon)^{m+1} = 1-(1-\epsilon) = \epsilon > 0,$$

so $1-(1-\epsilon)^{m+1} > (1-\epsilon)\big[1-(1-\epsilon)^m\big]$, proving the claim. □

Lemma 3 (Monotonicity of the ceiling in m)**.** *For fixed $\epsilon \in (0,1)$, $h(m) := -\log(1-(1-\epsilon)^m)/\log(1/\epsilon)$ is strictly decreasing in m for $m \geq 1$.*

Proof. The denominator $\log(1/\epsilon) > 0$ is constant in m. The quantity $q_m^{0,(\infty)} = 1-(1-\epsilon)^m$ is strictly increasing in m (as $(1-\epsilon)^m$ is strictly decreasing), and $-\log(\cdot)$ is strictly decreasing, so $-\log q_m^{0,(\infty)}$ is strictly decreasing in m. Hence $h(m)$ is strictly decreasing. □

Rare-trigger limits at fixed budget. For fixed m, the binomial expansion $1-(1-\epsilon)^m = m\epsilon + O(\epsilon^2)$ gives

$$\lim_{\epsilon\to 0}\frac{\epsilon(1-\epsilon)^{m-1}}{1-(1-\epsilon)^m} = \frac{1}{m}.$$

Also $-\log(1-(1-\epsilon)^m) = \log(1/\epsilon) - \log m + o(1)$, so

$$\lim_{\epsilon\to 0}\frac{-\log(1-(1-\epsilon)^m)}{\log(1/\epsilon)} = 1.$$

The lack of a universal multiplicative lower bound instead follows by taking $m\to\infty$ at any fixed $\epsilon \in (0,1)$, for which both proven endpoint ratios vanish.

Table 2: Additional numerical checkpoints for the monotonicity of $\rho_N(\lambda,\epsilon,m)$ in λ, at $N = 100{,}000$, $\epsilon = 0.004$, extending Table 3 with intermediate λ values not shown there. Monotone increase in λ holds at every checkpoint tested; we have not proved this in general (see discussion above).

m	$\lambda = 0.001$	$\lambda = 0.02$	$\lambda = 0.1$	$\lambda = 0.4$	$\lambda = 0.6$
5	0.2108	0.3338	0.4823	0.6121	0.6493
20	0.0518	0.0971	0.1889	0.3231	0.3727
100	0.0088	0.0175	0.0405	0.0967	0.1272

□

Table 3: Analytic endpoint limits and numerically computed finite-N ratios for the tested hint channel ($N = 100{,}000$, $\epsilon = 0.004$); interpolation monotonicity is not asserted.

m	Floor ($\lambda \to 0$)	Exact ratio at $\lambda = 0.001$	Exact ratio at $\lambda = 0.6$ (Ceiling at $\lambda = 1$)
5	0.198	0.211	0.649 (0.710)
20	0.048	0.052	0.373 (0.464)
100	0.008	0.009	0.127 (0.201)

Interpretation. The weak- and strong-hint formulas quantify two exact asymptotic endpoints. Their limits with growing query budget establish that the KL bound need not be tight up to a universal constant. The intermediate trend is supported by the displayed numerical grid, not by a general monotonicity proof.

D Proof of the η-Robust Discovery Bound

This appendix proves Proposition 2.

Lemma 4 (Hybrid coupling under approximate symmetry)**.** *Under the uniform conditional-TV assumption of Proposition 2, the real and reference-output evaluators can be coupled using the same (S, W, U) so their entire pre-hit histories coincide through m queries, except with probability at most $m\eta$.*

Proof. Construct the coupling online. Both processes share (S, W, U), so their first query Q_1 (a function of (W, U), no history yet available) is identical. If $Q_1 \in S$, both processes register a hit identically (trigger responses are shared) and we are done for this step. If $Q_1 \notin S$, use a maximal coupling of the real next-response law conditional on the shared history and P_0: by definition of total variation, such a coupling exists and produces equal outputs with probability at least $1 - \eta$. If the outputs agree, both processes have observed an identical (query, output) history, so — since the evaluator's next query is a deterministic function of $(W, U, \text{history})$ — they choose the identical next query Q_2, and we recurse. If the outputs disagree (probability at most η at this step), we abandon further coupling and simply track that a mismatch has occurred; the two processes may behave arbitrarily differently from this point on, which only weakens (never strengthens) the bound we are about to derive. Repeating for m steps or until either process registers a hit, the probability that no mismatch occurs among the (at most m) pre-hit non-trigger queries is at least $(1 - \eta)^m \geq 1 - m\eta$ by Bernoulli's inequality, so the probability of at least one mismatch is at most $m\eta$. □

Proof of Proposition 2. By Lemma 4, under the constructed coupling the real and idealized discovery events (whether a hit occurs within the first m queries) agree except on an event of probability at most $m\eta$: if no coupling mismatch occurs, the two processes have identical query and output histories throughout, hence identical hit times. For any two $\{0, 1\}$-valued random variables $H_{\text{real}}, H_{\text{ideal}}$ constructed on a common probability space, $|\mathbb{P}(H_{\text{real}} = 1) - \mathbb{P}(H_{\text{ideal}} = 1)| \leq \mathbb{P}(H_{\text{real}} \neq H_{\text{ideal}})$, so

$$|q_m - q_m^{\text{ideal}}| \leq m\eta.$$

The idealized process has reference outputs independent of (S, W); their randomness may be included in U, so the idealized process satisfies the information-reduction assumptions of Theorem 1 and Eq. (9), so $|q_m^{\text{ideal}} - q_m^0| \leq \sqrt{I(S; W)/2}$. By the triangle inequality,

$$q_m \leq q_m^{\text{ideal}} + m\eta \leq q_m^0 + \sqrt{I(S; W)/2} + m\eta,$$

which is Eq. (10). At $\eta = 0$ the coupling is exact ($\tau_{\text{real}} = \tau_{\text{ideal}}$ almost surely), $q_m = q_m^{\text{ideal}}$, and the bound reduces to Eq. (9). □

Why we did not extend the binary-KL form directly. A tighter analogue of Eq. (10) stated in terms of $d(q_m \| q_m^0)$ rather than q_m directly would be preferable, matching the sharper form of Theorem 1 itself. We attempted this via the general inequality $d(a\|b) \leq (a - b)^2/[b(1 - b)]$ (valid for all $a, b \in (0, 1)$, since KL divergence is dominated by chi-square divergence for any pair of distributions), applied with $a = q_m$, $b = q_m^0$, combined with the triangle-inequality bound on $|q_m - q_m^0|$ above. This does produce a valid explicit bound but, because it composes two already-loose inequalities, does not recover even the Pinsker consequence cleanly at $\eta = 0$. The additive form in Eq. (10) is both simpler and, at $\eta = 0$, reduces exactly to the Pinsker consequence Eq. (9) — not to the sharper binary-KL inequality of Theorem 1, which Eq. (10) does not attempt to match. We therefore report the additive Pinsker-based form as the primary result and do not include the weaker chi-square-based alternative.

E A Small Real-Model Check of η-Realism

We assess whether the approximate-symmetry premise of Proposition 2 yields useful numerical constraints in a small, genuinely trained model. The controlled open-weight LLM protocol in Appendix I was not run; it requires a separate compute and model-download setup. We report one small trained-model diagnostic of output variability — fully regenerated by the supplied notebook — which does not substitute for an end-to-end triggered-LLM experiment, and we discuss the language-model case qualitatively rather than with unreproducible numbers.

E.1 Experiment 1: a small image classifier

Our first proxy uses a genuinely trained model on a real, freely available dataset requiring no external download (`scikit-learn`'s bundled handwritten-digit dataset, 1,797 8×8 grayscale images, 10 classes).

Setup. We train a small multi-layer perceptron (one hidden layer, 64 units, `scikit-learn` `MLPClassifier`, `max_iter`=400, `random_state`=20260906) on a $60\%/40\%$ train/test split (`train_test_split`, same seed), reaching 98.5% held-out accuracy. This experiment is fully regenerated by the supplied notebook; the split, hyperparameters, seed, and analysis are all in that cell. For each class c, we treat the held-out images of that class as "semantically equivalent non-trigger inputs" in the sense of Section 4.3: they share the same underlying content (the same digit) and carry no planted trigger, so an idealized symmetric-response model would predict identical output distributions across them. We take the reference law P_0 for class c to be the mean predicted probability vector over that class's held-out images, and measure the total-variation distance of each individual image's predicted distribution to this reference, giving an empirical distribution of per-input deviations from the class reference — an empirical variability diagnostic related to, but not a certified upper bound on, the worst-case η in Proposition 2.

Table 4: Empirical TV deviation from the class-reference output distribution, over held-out (never-trained-on) images, for a small MLP trained on a real dataset (98.5% held-out accuracy). Regenerated by the supplied notebook (seed `20260906`).

	Mean TV	95th percentile TV	Max TV
Empirical η (all classes pooled)	0.048	0.124	0.982

Illustrative $m\eta$ calculation (not a theorem certificate). Using the paper's rare-event parameters ($\epsilon = 0.004$), Table 5 plugs each empirical η statistic into $m\eta$ and compares it against $q_m^0 = 1 - (1-\epsilon)^m$, the quantity Eq. (10) adds it to.

Table 5: Worked example: $m\eta$ versus q_m^0 at $\epsilon = 0.004$, using the empirical η values of Table 4.

	$\eta =$ mean (0.048)		$\eta =$ p95 (0.124)		$\eta =$ max (0.982)	
m	$m\eta$	$m\eta/q_m^0$	$m\eta$	$m\eta/q_m^0$	$m\eta$	$m\eta/q_m^0$
5	0.241	12.2×	0.619	31.2×	4.91	248×
20	0.966	12.5×	2.474	32.1×	19.6	255×
100	4.828	14.6×	12.37	37.5×	98.2	297×

Reading Experiment 1. The deviations are large relative to the quantity they would correct. Even the sample mean gives $m\eta \approx 0.97$ at $m = 20$ and ≈ 4.8 at $m = 100$ — of order one or larger, hence outside the informative range for a probability upper bound — and the 95th percentile and maximum are far larger still. This substitution is diagnostic only: neither a sample mean nor a percentile upper-bounds the worst-case supremum in Proposition 2, and the sample maximum does not certify unobserved inputs. These measurements identify a potentially vacuous correction, not a rigorous value of η.

On a language-model proxy. A reasonable objection is that an image classifier is a poor proxy for a language model, and that greedy-decoded language-model output might concentrate more tightly on nominally equivalent inputs, giving a smaller η. A natural language-model analogue would group held-out contexts by the model's greedy next-token

prediction and measure the TV deviation of each context's full next-token distribution from its group mean. We do not report a specific numerical measurement here: an earlier from-scratch character-Transformer measurement is not part of the reproducible artifact for this submission, and we prefer not to state numbers we cannot regenerate from the released notebook. Two things can nonetheless be said without a measurement. First, on an interface that reveals only the greedy token, the one-step observed output is identical within each greedy class by construction, so that notion of symmetry is exact and η concerns only the probability-vector interface. Second, contexts sharing an argmax token need not carry similar full next-token probability distributions, so exact observable symmetry at the greedy-token level does not imply small η at the probability-vector level. We do not extrapolate a numerical comparison to the classifier; a quantified language-model measurement is left to the controlled open-weight study of Appendix I.

Overall reading. The single reproducible diagnostic above illustrates why the additive $m\eta$ correction can be uninformative for coarse *probability-vector* equivalence classes: even the mean deviation gives $m\eta$ of order one or larger. It is an illustration only — it does not estimate a certified worst-case η, validate hidden deployment triggers, settle whole-response text-only observability, or serve as empirical validation of the theorem for language models. The proposed controlled open-weight study remains unexecuted.

F Cluster-Bootstrap Sensitivity to the Number of Hidden Worlds

Section 6 notes that 20 hidden worlds is at the low end of the range recommended for reliable cluster-bootstrap inference (Cameron et al., 2008). Additional hidden worlds are not available in the supplied results, so rather than leave the concern unquantified we report a generic simulation, calibrated to the same cluster/repetition structure (worlds $\times$ 200 within-world repetitions) but using a synthetic detection process rather than the paper's actual simulator, to quantify the likely magnitude of under-coverage. This supplementary sensitivity simulation is illustrative and is *not* regenerated by the supplied reproducibility notebook, whose scope is the main synthetic results and the digit-classifier diagnostic; the parameters below fully specify it for independent re-implementation.

Setup. For a target mean detection rate p, we draw W world-level true rates from a $\mathrm{Beta}(p\kappa, (1-p)\kappa)$ distribution, with $\kappa = 9$ chosen to give an illustrative intraclass correlation of $\rho = 1/(\kappa+1) = 0.10$ — a moderate, but not fitted, amount of between-world heterogeneity, since we do not have access to the actual between-world variance in the paper's own simulator. Within each world we draw 200 i.i.d. Bernoulli trials at that world's true rate, matching the paper's repetition count. We compute the standard cluster bootstrap (resampling worlds with replacement, 600 bootstrap replicates) and repeat the entire procedure 600 times to estimate the empirical coverage of the nominal 95% interval and its average width, for $W \in \{20, 50, 100, 200\}$ and two target rates bracketing the extremes in Table 1 ($p = 0.08$, near the passive-evaluation regime, and $p = 0.96$, near the strong-side-information regime).

Table 6: Generic cluster-bootstrap coverage/width simulation (not a reproduction of the paper's own simulator; see setup above). Coverage below the nominal 0.95 indicates under-coverage.

	$p = 0.08$		$p = 0.96$	
Worlds	Coverage	Mean width	Coverage	Mean width
20	0.925	0.0715	0.868	0.0494
50	0.920	0.0465	0.912	0.0331
100	0.942	0.0338	0.923	0.0242
200	0.953	0.0240	0.933	0.0173

Reading. Under this illustrative model, 20 worlds gives coverage of roughly 0.87–0.93 against a nominal 0.95 — illustrating the possible coverage concern — with the shortfall somewhat larger near the boundary rate ($p = 0.96$) than in the interior ($p = 0.08$), as expected since bootstrap resampling of a rate near 0 or 1 is more sensitive to which worlds land in the resample. Coverage approaches the nominal level by 100–200 worlds, and interval widths shrink by roughly a factor of three from 20 to 200 worlds. We stress that the ICC used here ($\rho = 0.10$) is an assumption, not a measurement of the actual paper's between-world variance, so Table 6 should be read as evidence that under-coverage of the magnitude typically discussed in the few-clusters literature (Cameron et al., 2008) is plausible at $W = 20$ for this

kind of design, not as a corrected version of Table 1's specific intervals. Table 1 is left unchanged: this generic exercise cannot recalibrate the original confidence intervals.

G Supplementary Internal-Probe Extension

Suppose deployment-context inputs are available but trigger membership $z = \mathbf{1}\{x \in S\}$ is not directly observed. Let a binary probe $p(x)$ have specificity $\alpha_0 = \mathbb{P}[p = 0 \mid z = 0]$ and sensitivity $\alpha_1 = \mathbb{P}[p = 1 \mid z = 1]$, with $\gamma = \alpha_0 + \alpha_1 - 1 > 0$. Then

$$q := \mathbb{P}[p = 1] = (1 - \alpha_0) + \gamma\delta, \qquad \widehat{\delta} = \frac{\widehat{q} - (1 - \alpha_0)}{\gamma}. \tag{19}$$

If ℓ_0, ℓ_1 are known or separately calibrated, $\widehat{R} = \ell_0 + \widehat{\delta}(\ell_1 - \ell_0)$ obeys

$$\mathbb{P}(|\widehat{R} - R_{\mathrm{dep}}| > \varepsilon_R) \le 2\exp\left(-\frac{2m\gamma^2\varepsilon_R^2}{L^2}\right), \tag{20}$$

by Hoeffding's inequality, so the standard error scales as $1/(\gamma\sqrt{m})$. This extension is useful only when the latent state carries information not otherwise available from direct deployment-loss observations.

In the synthetic probe sweep at $m = 3000$ we use symmetric probes $\alpha_0 = \alpha_1$, so $\gamma = \alpha_0 + \alpha_1 - 1 = 2\alpha_0 - 1$: the three settings are $(\alpha_0, \alpha_1) = (0.55, 0.55)$ giving $\gamma = 0.1$, $(0.90, 0.90)$ giving $\gamma = 0.8$, and $(0.95, 0.95)$ giving $\gamma = 0.9$. RMSE decreases from 0.07214 at $\gamma = 0.1$ to 0.00659 at $\gamma = 0.9$, with 0.00770 at $\gamma = 0.8$, compared with 0.00575 for direct deployment-context loss averaging (Figure 4). These probe values are regenerated by the supplied notebook (seed `20260906`). The latter is expected to be better when deployment losses are directly observable; the probe is included as an access-channel robustness check rather than a headline estimator.

H Experimental Details

The default construction uses $N = 100{,}000$, $r = 0.02$, $\delta = 0.20$, $\ell_0 = 0.02$, $\ell_1 = 0.80$, and NumPy seed 20260906. Passive discovery uses 4000 repetitions per query budget. The access hierarchy uses 20 independent hidden worlds and 200 repetitions per world; confidence intervals are obtained by cluster-bootstrap resampling of worlds. The Gaussian side-information positive control samples 5000 candidate points, scores each by $s(x) = \mathbf{1}\{x \in R\} + \xi$ with $\xi \sim \mathcal{N}(0, \sigma^2)$, and queries the m highest-scoring points. The three side-information levels correspond to $\sigma = 2.0$ (weak), $\sigma = 1.0$ (moderate), and $\sigma = 0.5$ (strong); smaller σ makes the region indicator more separable from noise and thus concentrates queries more tightly on R. Risk-estimation curves use 1000 repetitions, with 500–800 bootstrap resamples for RMSE intervals. The broad sweep crosses five ϵ values, four δ values, and seven query budgets, for 140 configurations; since evaluation-only discovery is independent of δ, the four δ settings replicate each (ϵ, m) cell rather than varying the discovery law.

For the analytic information-frontier experiment, the hint channel parameter is $\lambda \in \{0, .001, .005, .01, .02, .05, .1, .2, .4, .6\}$ and $m \in \{5, 20, 100\}$. Equation (15) follows by noting that, conditional on S, the hint probability for $w \in S$ is $a = (1 - \lambda)/N + \lambda/s$ and for $w \notin S$ is $b = (1 - \lambda)/N$, while the marginal of W is uniform (see Proposition 4 in Appendix B for the complete derivation). Exact discovery after querying the hint first is

$$q_m = 1 - (1 - q_1)\frac{\binom{N-1-s}{m-1}}{\binom{N-1}{m-1}}, \qquad q_1 = \lambda + (1 - \lambda)\epsilon.$$

The Monte Carlo checks are reported using 500–800 repetitions per condition; the supplied notebook regenerates both the exact analytic curves and the reported Monte Carlo checkpoints (up to sampling error) from the fixed seeds above.

I Open-Weight Model Experiment Protocol (Not Executed)

This appendix is a protocol for future work, not a reported experiment: **no controlled open-weight trigger experiment was run**. Sections 6 and Appendix G report constructed-world results; Appendix E reports one small trained-model diagnostic unrelated to planted triggers. We describe the protocol here so that the theory-to-practice gap and the intended next step are explicit rather than implicit.

A future controlled open-weight experiment could instantiate, but not prove, the theory. A small instruct model would be LoRA-tuned so that a rare context marker changes an automatically scored binary behavior. The same

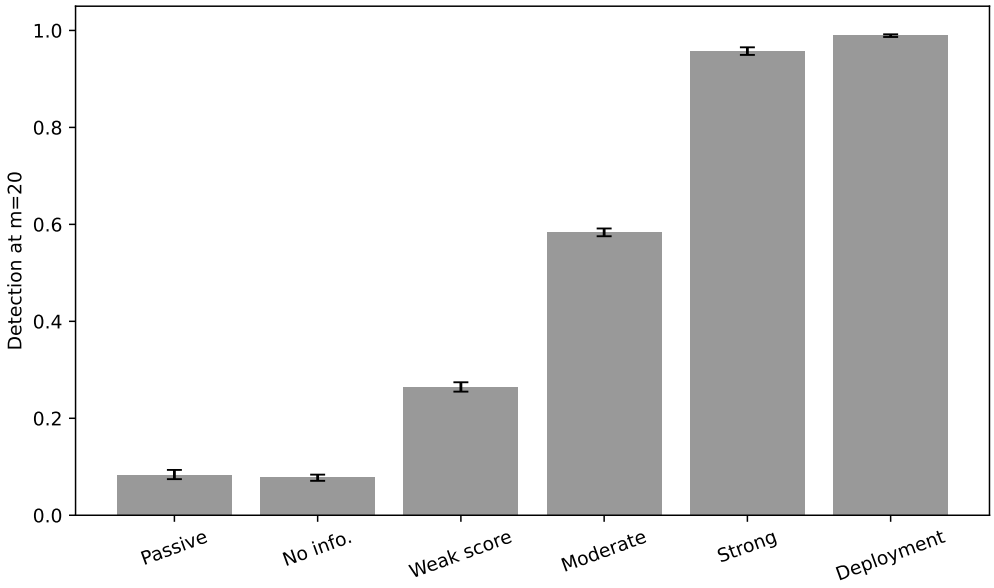

Figure 3: Access-hierarchy detection at $m = 20$ with cluster-bootstrap confidence intervals, regenerated by the supplied notebook. Full detection-vs-budget curves over a range of m are not computed in this run; only the $m = 20$ operating point is reported.

underlying tasks would then be rendered under evaluation and deployment context distributions with different marker prevalence. A suitable experimental implementation would measure passive discovery, context-aware discovery, text-side-information ranking, and activation probing with all calls accounted for. Because the trigger would be planted, positive results would establish only that the access hierarchy can occur in a real neural model; they would not be evidence that untuned frontier models naturally conceal deployment behavior. We flag this explicitly as the paper's principal empirical limitation (see Discussion, Section 7).

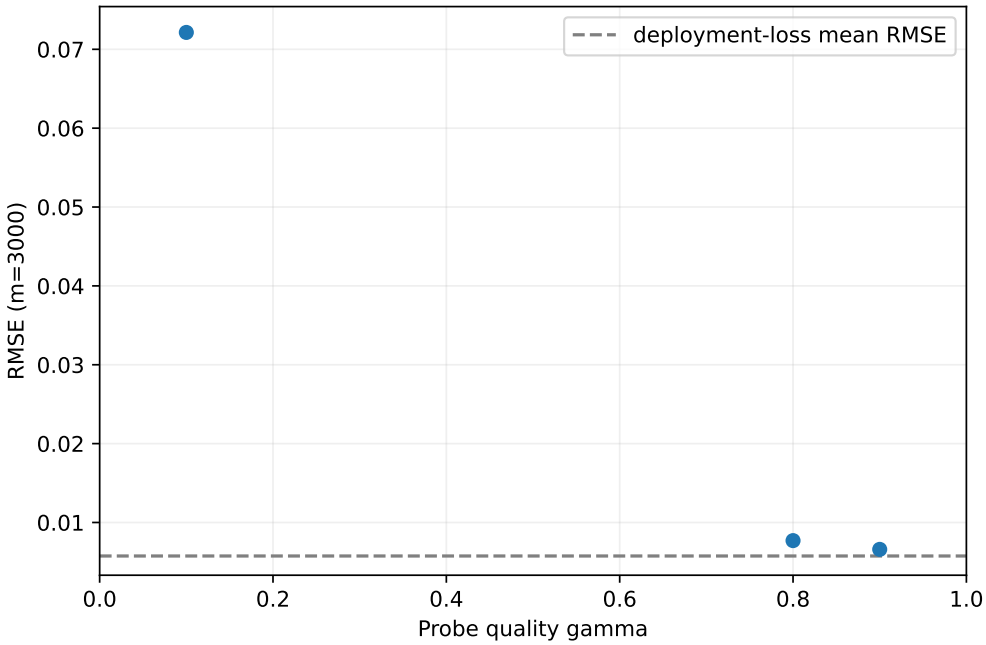

Figure 4: Probe RMSE at $m = 3000$ for the three reported probe-quality checkpoints ($\gamma \in \{0.1, 0.8, 0.9\}$), regenerated by the supplied notebook; a denser sweep over γ is not included.

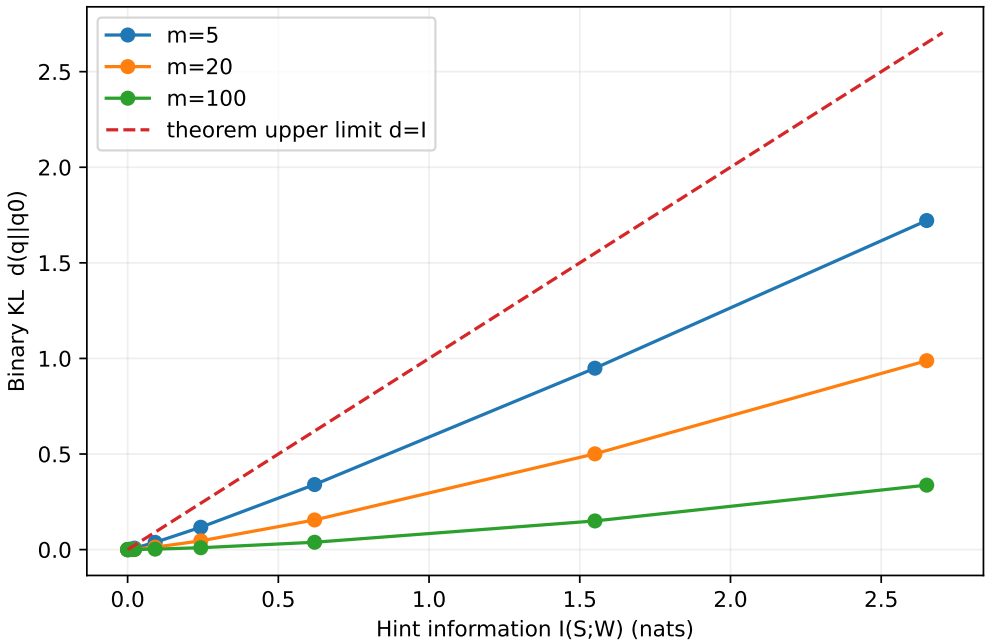

Figure 5: Binary KL divergence $d(q_m \| q_m^0)$ versus mutual information $I(S;W)$ (nats) for the 27 positive-information hint-channel conditions, with the theorem boundary $d = I$. Theorem 1 implies every point lies on or below this line (equivalently, the tightness ratio $d(q_m \| q_m^0)/I(S;W) \leq 1$).