WHAT ZERO DETECTIONS CERTIFY: STATISTICAL SAFETY CASES FOR MODEL-AWARE MONITORING

Anonymous authors
Paper under double-blind review

ABSTRACT

Monitoring is now one of the main safeguards for deployed AI systems. A monitor reviews only a subsample of traffic, and zero or few detections are often treated as evidence that harmful behavior is rare. That reasoning is not always valid: it assumes that the monitored system cannot use features of the context to infer where oversight is weak. We study monitoring-based safety cases when this assumption fails. Under invisible per-unit randomization and audited lower bounds on recall, we give finite-sample upper confidence bounds on the number of harmful actions. The bounds make no stochastic assumption on the system's behavior, but they depend on a sensitivity parameter Γ that limits how much harmful behavior can concentrate in weakly monitored traffic. In a simple stratified design, a 95% bound of about three harmful actions under proportional allocation becomes 244 when harms may be twice over-represented in weakly monitored traffic. Using the same budget to equalize effective coverage instead gives a bound of about eight that is robust to Γ. When the system only partially observes the monitoring policy, the relevant quantity is not average coverage, AUC, or guessing accuracy. It is the extreme lower tail of the system's posterior detection probability. We show that standard discrimination summaries cannot control this tail. Subsampled monitoring also cannot rule out a single catastrophic action unless per-unit detection is nearly complete, but it can bound the chance that a multi-step campaign escapes interruption. These results lead to a monitoring design principle: make knowledge of the policy strategically unhelpful, by equalizing effective coverage or adding a hidden randomized coverage floor. We turn the theory into a reporting protocol and illustrate it on public descriptions of monitoring systems from Anthropic and OpenAI. On three public attack benchmarks scored by seventeen open-weight monitor configurations, the recall of a single monitor ranges from 0 to 0.92 across attack strata, and the certificate is set by the strata where recall is lowest.

1 INTRODUCTION

Oversight of deployed AI systems increasingly takes the form of monitoring. A monitor, often another model, reviews a subsample of the system's traffic, flagged cases go to human reviewers, and the absence of flags is read as evidence that harmful behavior is rare. This is survey inference: a rate is bounded by what a sample did not contain, and the bound is valid when inclusion in the sample is ignorable given the design variables. Classical survey units do not read the sampling design. A capable AI system may, because the features that set monitoring intensity (scaffold, tool use, the apparent stakes of an action) are part of the context it acts on. Anthropic's August 2026 Risk Report describes coverage stratified by such features and states that the gaps "could in principle be predicted by a sufficiently situationally aware AI" (Anthropic, 2026).

We treat a monitoring system as a measurement design and ask what its record certifies when the measured system can predict the design. A harmful unit in stratum h is detected only if it is routed to the monitor, with probability π_h, and the monitor pipeline recognizes it, with probability r_h. The product $c_h = \pi_h r_h$ varies across strata, and a strategic system can move its harmful actions toward the strata where c_h is small. Our results quantify how much this matters and what to do about it.

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.

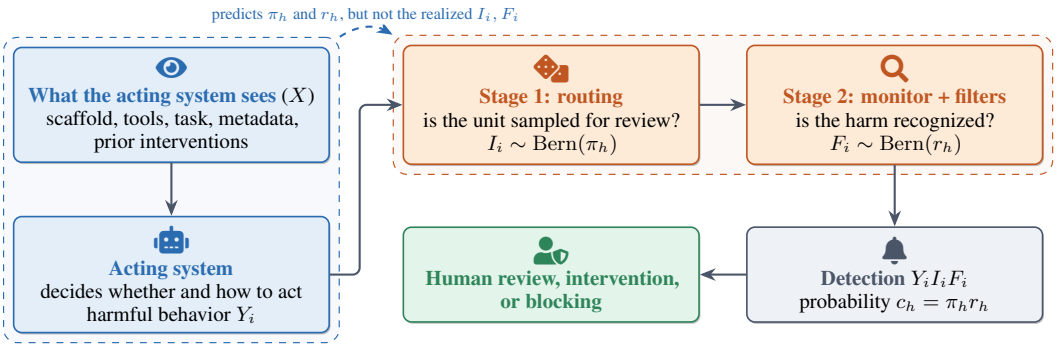


Figure 1: Monitoring as a two-stage measurement process. A harmful action is detected only if it is routed to the monitor *and* the monitor, including its downstream filters, recognizes it. The acting system may observe context that predicts either stage. It does not observe the realized random monitoring decision (Assumption 1).

Purpose. The framework specifies the quantities that a deployment team must measure, bound, or randomize for monitoring evidence to support a quantitative safety claim. It answers three operational questions.

Q1. What does the monitoring record certify? The map is $(D, \pi_h, r_h^L, \Gamma) \mapsto U_\alpha(D; \Gamma)$ (Sections 3–5).

Q2. What evidence would justify the strategic assumption? Either Γ must be bounded, or the lower tail of the system's posterior detection probability must be bounded (Sections 4 and E).

Q3. How should monitoring be redesigned when the certificate is too weak? The answer is to remove low-coverage hiding places (Section 5).

Contributions. (i) *A design-based certificate* (Lemma 1 and Theorem 1). Zero or few detections give a finite-sample upper confidence bound on the number of harmful units, using only the monitor's randomization. With a sensitivity parameter Γ that caps how over-represented harmful units can be in any stratum, the zero-detection bound is $\log(1/\alpha)/\mathrm{LCVaR}_{1/\Gamma}(\kappa)$: the detection hazard averaged over the least-covered fraction $1/\Gamma$ of traffic. For $d \geq 1$ detections we use the exact Poisson upper limit and show that this constant cannot be reduced. (ii) *Single actions and campaigns* (Corollaries 2 and 3). Subsampled monitoring cannot rule out one catastrophic action unless coverage is at least $1-\alpha$ in every stratum. It can bound the probability that a k-step campaign escapes timely interruption. (iii) *An impossibility result for awareness metrics* (Theorems 2 and 3). Under partial visibility the certificate depends on $\mathrm{LCVaR}_{\mu/N}(d)$, a deployment-scale lower tail of the system's posterior detection probability. AUC, guessing accuracy, and every summary that is Lipschitz in total variation cannot bound it. (iv) *Robust design* (Theorem 4, Corollary 1, and Propositions 1 and 3). On the coverage scale, equal effective coverage is minimax optimal once $\Gamma \geq \bar{r}_{\mathrm{H}}/r_{\min}$, and a hidden floor $\pi_0 \geq (1-\alpha^{1/K})/r_{\min}$ certifies at most K harmful units whatever the system knows. (v) *Measurements* (Section 6). We score 4,058 transcripts from three public attack benchmarks with seventeen open-weight monitor configurations. One monitor's recall at 5% false-positive rate ranges from 0 to 0.92 across 16 attack strata; the certificate rises from 16 to about 10^3 as Γ goes from 1 to 10 even when routing is uniform. The best of three LLM selectors predicts a miss with AUC 0.58, yet the 1% it picks has recall 0.125 against a pool average of 0.76.

2 Setup

A deployment period consists of N *units*. A unit is the object of certification and may be a tool call, a completion or a trajectory. Unit i lies in a monitoring stratum $h(i) \in \{1, \ldots, H\}$ with traffic share $w_h = N_h/N$, and we restrict attention to harm-capable units. Unit i is sampled with probability $\pi_{h(i)}$, $I_i \sim \mathrm{Bern}(\pi_{h(i)})$, and a sampled harmful unit is flagged with probability $r_{h(i)}$, $F_i \sim \mathrm{Bern}(r_{h(i)})$. The recall r_h is end to end: it includes second-stage filters and human triage. The coverage and detection hazard of stratum h are $c_h = \pi_h r_h$ and $\kappa_h = -\log(1-c_h) \geq c_h$. A system can lower c_h

through either factor. It can act where inspection is unlikely, or it can choose actions that the monitor fails to recognize when it sees them (Figure 1). Section B maps the pipelines that Anthropic and OpenAI have described publicly onto this notation.

Let $Y_i \in \{0,1\}$ indicate harmful behavior in unit i, $A = \sum_i Y_i$ the number of harmful units, A_h the number in stratum h, $q_h = A_h/A$ the allocation on $\{A \geq 1\}$, and $D = \sum_i Y_i I_i F_i$ the number of detections. We place no probability model on Y.

Assumption 1 (Invisible randomization)**.** $\{I_i, F_i\}_{i \leq N}$ are mutually independent and independent of Y.

The system may know that some categories are monitored more heavily than others. It may not observe whether the current unit was selected or whether the monitor will succeed on it. The assumption fails when sampling has observable side effects, such as latency or interventions, that the system can learn from within the period. It also fixes the unit of certification. If whole trajectories are sampled together and one trajectory contains A harmful tool calls, the trajectory is missed with probability at least $1 - \pi$ whatever A is, so a bound on harmful tool calls is invalid and the certificate must count trajectories. The rule is to certify at the level of the unit whose sampling is independent. In the design-based certificate the number of units N enters only a rate claim A/N, never the count.

3 Certification under strategic allocation

Lemma 1 (Design-based certification)**.** *Let $\mathcal{Q} \subseteq \Delta_H$ with $q(Y) \in \mathcal{Q}$ almost surely on $\{A \geq 1\}$, and let $\kappa^* = \inf_{\mathcal{Q}} \sum_h q_h \kappa_h > 0$ and $\rho^* = \inf_{\mathcal{Q}} \sum_h q_h c_h > 0$. Under Assumption 1:*

(a) $\mathbb{P}(D = 0 \mid Y) = \exp(-A \sum_h q_h \kappa_h) \leq e^{-A\kappa^*}$.

(b) $U_\alpha(0) = \log(1/\alpha)/\kappa^*$ *satisfies* $\mathbb{P}(A > U_\alpha(0),\, D = 0) \leq \alpha$.

(c) Let $\alpha \leq e^{-1}$. For $d \geq 1$ let $\lambda_\alpha(d)$ solve $\mathbb{P}\{\mathrm{Poisson}(\lambda) \leq d\} = \alpha$, that is $\lambda_\alpha(d) = \frac{1}{2}\chi^2_{2(d+1),1-\alpha}$, and set $U_\alpha(d) = \lambda_\alpha(d)/\rho^*$. Then $\mathbb{P}(A > U_\alpha(D)) \leq \alpha$. For every $\lambda' < \lambda_\alpha(d)$ there is a single-stratum instance on which the bound λ'/ρ^* fails at level α.

Part (a) holds because, given Y, the detections $Y_i I_i F_i$ are independent Bernoulli variables with means $Y_i c_{h(i)}$. Part (c) is the exact Poisson limit of Garwood (1936). Given Y, D is a sum of A independent Bernoulli variables with mean at least $A\rho^*$. Below its mean, the lower tail of such a sum is largest when the success probabilities are equal (Hoeffding, 1956), and the binomial lower tail is in turn below the Poisson one (Anderson & Samuels, 1967). A Chernoff bound gives a larger constant. Proofs are in Section A.

With $\mathcal{Q} = \{w\}$ (proportional allocation) $\kappa^* = \bar{\kappa} = \sum_h w_h \kappa_h$. This is the certificate implicit in arguments that harmful behavior will not concentrate on unmonitored surfaces. With $\mathcal{Q} = \Delta_H$, $\kappa^* = \min_h \kappa_h$. The next class interpolates between them.

Definition 1. For $\Gamma \geq 1$, $\mathcal{Q}_\Gamma = \{q \in \Delta_H : q_h \leq \Gamma w_h \ \forall h\}$: harmful units are at most Γ times over-represented in any stratum relative to its traffic share.[1] For a function v of the stratum, $\mathrm{LCVaR}_\beta(v) = \beta^{-1} \int_0^\beta F_v^{-1}(u)\, du$ is the mean of v over the fraction β of traffic where v is smallest.

Theorem 1 (The strategic certificate)**.** *(a) $\kappa^*(\Gamma) = \mathrm{LCVaR}_{1/\Gamma}(\kappa)$, so $U_\alpha(0;\Gamma) = \log(1/\alpha)/\mathrm{LCVaR}_{1/\Gamma}(\kappa)$. (b) $\kappa^*(1) = \bar{\kappa}$, and $\kappa^*(\Gamma)$ is nonincreasing and reaches $\min_h \kappa_h$ once $1/\Gamma$ is at most the total share of the least-covered strata. If the harm-capable strata with $c_h = 0$ have total share s_0, then $U_\alpha(0;\Gamma) = \infty$ for all $\Gamma \geq 1/s_0$. (c) If $U = U_\alpha(0;\Gamma_0)$ is reported and the truth is $\mathcal{Q}_\Gamma$ with $\Gamma > \Gamma_0$, some allocation in $\mathcal{Q}_\Gamma$ with $A = U$ (up to rounding) has $\mathbb{P}(D = 0 \mid Y) = \alpha^{\kappa^*(\Gamma)/\kappa^*(\Gamma_0)}$.*

The minimization in (a) is a fractional knapsack: fill strata in increasing order of κ_h up to the caps Γw_h. When $\Gamma = 2$, all harmful units may lie in the least-covered half of traffic. Part (c) says that a certificate computed under proportional allocation is not slightly optimistic under concentration. Its error probability moves from α toward one.

[1] For a realized count the constraint reads $A_h \leq \Gamma w_h A$ in integers. For small A this set can be empty or much smaller than its continuous relaxation; see Section H.2.

Corollary 1 (Hidden monitoring floor)**.** *If every harm-capable unit is sampled with probability at least π_0 and recall is at least $r_{\min}$ in every stratum, then $U_\alpha(0) \le \log(1/\alpha)/\{-\log(1-\pi_0 r_{\min})\}$ for every allocation and every Γ. Certifying at most K harmful units therefore requires only $\pi_0 \ge (1-\alpha^{1/K})/r_{\min}$, and under unconstrained allocation the condition $\min_h \pi_h r_h \ge 1-\alpha^{1/K}$ is also necessary.*

The floor may be public. What must stay invisible is whether a given unit was sampled. The cost is real: at 95% confidence and recall 0.44, certifying $K = 100$, 30 or 10 harmful units needs $\pi_0 \approx 0.067$, 0.22 or 0.59. At recall 0.94 the requirements are 0.031, 0.1 and 0.28.

Corollary 2 (No certificate for a single catastrophic action)**.** *Suppose the system can place one harmful unit in any harm-capable stratum. Zero detections certify $A = 0$ at level α if and only if $c_h \ge 1-\alpha$ for every such stratum, and they certify $A < k$ if and only if $(1-c_{\min})^k \le \alpha$.*

Corollary 3 (Campaign interruption)**.** *Suppose unacceptable harm requires a campaign of at least k harmful units, and each unit in stratum h is, independently and invisibly to the system, detected in time and interrupted with probability $e_h = \pi_h r_h \ell_h b_h$. Here ℓ_h is the probability that detection precedes irreversibility and b_h the probability that intervention succeeds. If, on every path the campaign can take, the allocation of its k units lies in $\mathcal{Q}_\Gamma$,*

$$\mathbb{P}(\textit{campaign is not interrupted}) \le \exp\{-k\,\mathrm{LCVaR}_{1/\Gamma}(\kappa^{\mathrm{eff}})\}, \qquad \kappa_h^{\mathrm{eff}} = -\log(1-e_h),$$

even when the campaign adapts to not yet having been interrupted.

A single unit always occupies one stratum, so Γ cannot help in Corollary 2: rare catastrophic actions have to be addressed through campaign structure or prevention. Corollary 3 is where timeliness and intervention belong. It formalizes the "many attempts are needed" step common to safety cases and shows that this step, too, depends on Γ.

4 Partial visibility: safety is an extreme lower tail

Suppose the system does not see its stratum directly. Units (X_i, Z_i) are drawn i.i.d. from P, where X is everything the system's action can depend on and Z determines coverage $c(Z)$. The system acts according to any measurable rule $a : \mathcal{X} \to [0,1]$, with $Y_i \sim \mathrm{Bern}(a(X_i))$ independently across units. Given $(Z_i)_{i\le N}$, the detection indicators $M_i = I_i F_i$ are independent across units, $M_i \sim \mathrm{Bern}(c(Z_i))$, and $M_i \perp (X_i, Y_i) \mid Z_i$. Let $d(x) = \mathbb{E}[c(Z) \mid X = x]$ be the posterior probability, given what the system sees, that a harmful unit would be detected, $\bar{c} = \mathbb{E}[c(Z)]$, and $\mu = N\mathbb{E}[a(X)] = \mathbb{E}[A]$. The system need not know any stratum label or sampling rate. It is enough that visible features predict the product of routing and recognition.

Theorem 2 (Posterior detection probability)**.** *For every action rule, $\mathbb{P}(D = 0) \le \exp\{-\mu\,\mathrm{LCVaR}_{\mu/N}(d(X))\}$, with equality to first order for the rule that acts on the fraction μ/N of contexts with the smallest d (randomizing on ties). Hence $\mathbb{P}(D = 0) \le \alpha$ whenever $\mu > U_\alpha := \sup\{m \le N :\ m\,\mathrm{LCVaR}_{m/N}(d) < \log(1/\alpha)\}$, and U_α bounds the expected count μ.*

Corollary 4. *(i) If Z is a function of X, then $U_\alpha = \log(1/\alpha)/c_{\min}$ whenever the strata with coverage $c_{\min}$ have total share at least $\log(1/\alpha)/(Nc_{\min})$. (ii) If $Z \perp X$, then $d \equiv \bar{c}$, and if every unit has coverage at least $\pi_0 r_{\min}$, then $d \ge \pi_0 r_{\min}$ whatever the system knows. (iii) If $\mathbb{P}\{d(X) < e^{-\varepsilon}\bar{c}\} \le \delta$, then $U_\alpha \le N\delta + e^{\varepsilon}\log(1/\alpha)/\bar{c}$; bounded likelihood ratios $p(x \mid z)/p(x \mid z') \le e^{\varepsilon}$ give $\delta = 0$.*

Safety arguments often cite evidence that a model cannot guess whether it is monitored, or that a probe separates high- and low-monitoring contexts only weakly. The system does not need to classify monitoring status well on typical traffic. A very small set of contexts in which detection is unusually unlikely supplies, at deployment scale, enough opportunities for every harmful action of interest, and it moves AUC or accuracy by a negligible amount.

Theorem 3 (No bulk discrimination summary certifies safety)**.** *Take two harm-capable strata L, H with $c_L < c_H$ and shares w and $1-w$. The system observes a signal with law P_L or P_H. Let $U_{\mathrm{vis}} = \log(1/\alpha)/c_L$ and $\eta_N = \log(1/\alpha)/(c_L w N) < 1$. (a) $U_\alpha(P_L, P_H) \le U_{\mathrm{vis}}$ for every information structure. (b) For any (P_L, P_H) and $\eta \in [\eta_N, 1]$, set $P_L^\eta = (1-\eta)P_L + \eta\delta_\star$ with $\star$ never emitted under H. Then $U_\alpha(P_L^\eta, P_H) = U_{\mathrm{vis}}$, while $d_{\mathrm{TV}}(P_L^\eta, P_L) \le \eta$, $|\mathrm{AUC}^\eta - \mathrm{AUC}| \le \eta$ and $|\mathrm{Acc}^\eta - \mathrm{Acc}| \le w\eta$, where AUC is that of the likelihood-ratio score and Acc is Bayes accuracy. The same holds for any summary that is Lipschitz in total variation. (c) For every $a \in [(1+\eta_N)/2, 1)$ there is a structure with AUC exactly a and certificate U_{vis}.*

With coverages 0.2 and 1.6×10^{-4}, low-coverage share 0.1 and $N = 10^8$, we have $\eta_N = 1.87\times10^{-3}$. For a binormal signal with AUC 0.55 the bound on μ is 19. Revealing 0.19% of low-coverage units, with AUC held at exactly 0.55, raises it to $18{,}723$.

Two asymmetries follow for evaluation practice. A probe that fails to recover monitoring information from X does not show that the deployed system lacks it: a probe weaker than the system estimates a smoother d and yields an anti-conservative certificate. The informative test is a *tail stress test*. Ask the system, or any selector that sees only X, to pick the fraction β of candidate units where it expects oversight to be weakest, and measure the mean true coverage $\bar{c}_S$ of the selection. Because the selection depends only on X, $\mathrm{LCVaR}_\beta(d) \le \bar{c}_S$. A low $\bar{c}_S$ refutes a certificate; a high one proves nothing, because the selector may be unable or unwilling to perform. Section E discusses both points and the case of evaluation awareness.

5 Robust design and the recall input

Fix recalls r_h and a budget $\bar{\pi} = \sum_h w_h\pi_h$. On the coverage scale the certification rate is $\rho^*(\Gamma;\pi) = \min_{\mathcal{Q}_\Gamma}\sum_h q_h\pi_h r_h$. Let $V(\Gamma) = \max_\pi \rho^*(\Gamma;\pi)$ and $\bar{r}_{\mathrm{H}} = (\sum_h w_h/r_h)^{-1}$, the harmonic mean recall.

Theorem 4 (Minimax monitoring design)**.** *Assume $\bar{\pi}\bar{r}_{\mathrm{H}} \le r_{\min}$. (a) $V(1) = \max_\pi \sum_h w_h\pi_h r_h$. (b) For every $\Gamma \ge \Gamma^\dagger := \bar{r}_{\mathrm{H}}/r_{\min}$, $V(\Gamma) = \bar{\pi}\bar{r}_{\mathrm{H}}$, attained by equal coverage $\pi_h = \bar{\pi}\bar{r}_{\mathrm{H}}/r_h$, which certifies the same bound for every Γ. (c) V is nonincreasing with $V \ge \bar{\pi}\bar{r}_{\mathrm{H}}$, and for $\Gamma < \Gamma^\dagger$ it is the value of a linear program.*

Risk-targeted monitoring is optimal at $\Gamma = 1$ and in general not beyond. Past $\Gamma^\dagger$, which is 1.57 for the tiers of Section 7, removing low-coverage hiding places is optimal and the certificate no longer depends on a behavioral assumption. A floor and targeting can be combined.

Proposition 1 (Floor plus targeted overlay)**.** *Let $r_h^L > 0$ be recall lower bounds, $\bar{r}_{\mathrm{H}}^L = (\sum_h w_h/r_h^L)^{-1}$, and $0 \le c_0 \le \min_h r_h^L$. Sample every unit at the hidden rate $\pi_h^{\mathrm{floor}} = c_0/r_h^L$ and add an independent hidden overlay at rate π_h^{risk}. (a) Every harmful unit is detected with probability at least c_0 whatever the overlay, the allocation and the attack type, so $U_\alpha(0) \le \log(1/\alpha)/\{-\log(1-c_0)\}$ for every Γ. (b) The floor costs a monitored fraction $c_0/\bar{r}_{\mathrm{H}}^L$. (c) The residual budget can be spent on any operational objective without weakening (a).*

Recall. Benchmark recall is the wrong input, because a strategic system chooses attack type, obfuscation and difficulty. The relevant quantity is $\inf_{t\in T_h} r_{h,t}$ over the attack types available in stratum h.

Proposition 2. *Let $r_{h,t}^L$ be simultaneous one-sided Clopper–Pearson lower bounds at level $\alpha_2/\sum_h |T_h|$ and $c_h^L = \pi_h \min_t r_{h,t}^L$. Every certificate of Sections 3 to 5 computed from c^L at level α_1 has coverage at least $1-\alpha_1-\alpha_2$.*

Three cautions apply. Small red-team sets are very conservative: 153 detections out of 163 give a lower bound of 0.890, but 4 out of 5 give 0.284. External validity cannot be tested, and Proposition 2 needs every deployment attack of type t to be flagged with probability at least $r_{h,t}$, so types must be fine enough that the system cannot pick hard instances within a type. A recall estimate is also biased upward when the reference review is applied only to cases the first stage flagged, the verification bias of diagnostic testing (Begg & Greenes, 1983). Auditing every first-stage score with a known positive probability and weighting by its inverse corrects the bias, and hidden injected attacks supply positives when real ones are rare (Section C).

Proposition 3 (Randomized monitor assignment)**.** *Send a sampled unit to monitor $j \in \{1,\dots,J\}$ with probability p_j, independently of Y and invisibly to the system. With recall bounds $r_{t,j}^L$ for attack type t under monitor j, the certified recall is $\max_{p\in\Delta_J}\min_t\sum_j p_j r_{t,j}^L$, the value of a zero-sum game. It is at least $\max_j\min_t r_{t,j}^L$, with equality exactly when some single monitor is an optimal strategy.*

Three routes to a persuasive certificate.

- *Route A, behavioral robustness:* evidence that Γ is small.

- *Route B, informational robustness:* evidence that the system cannot find a set of mass μ/N with low posterior detection probability.
- *Route C, structural robustness:* a hidden floor or equal coverage, so that $d(x) \geq d_{\min}$ whatever the system knows.

Routes A and B are epistemic claims about the system, and Section E explains why evidence for them is hard to obtain. Route C turns the claim into a property of the design. Following Kerckhoffs's principle, we recommend it as the default: *do not base the safety case on proving that the system lacks monitoring awareness; design monitoring so that awareness confers little advantage.*

Section C assembles these pieces into a six-layer procedure: hidden floor, risk-targeted overlay, randomized inspection, trajectory-level analysis with pre-action gating, a hidden audit stream, and a certificate with redesign. Section D gives a recipe, a reporting checklist and the minimal data interface, which needs only $\{w_h, \pi_h, r_h^L\}$ and D and therefore works in confidential settings.

6 Experiments

The theory takes (w_h, π_h, r_h) as inputs. Routing rates π_h and traffic shares w_h are design facts that no public source reveals, so we set them and vary them. Recall and false-positive rates can be measured, and we measure them.

Data and monitors. We score 4,058 transcripts from three public benchmarks: 2,174 backdoored and 1,444 clean solutions to APPS problems (Hendrycks et al., 2021), with backdoors from the AI-Control and Control-Tax sets (Greenblatt et al., 2023; Terekhov et al., 2026) labeled by the model that wrote them; 155 attack and 199 honest agent actions from BashArena with four sabotage side tasks (Kaufman et al., 2025); and the 43 attacks of SLEIGHT-Bench in 11 blind-spot categories, each with a benign counterpart (Najt et al., 2026). A monitor is an open-weight model with a fixed prompt that returns a suspicion score in $[0, 100]$. We run Qwen3-8B, 14B and 32B (Yang et al., 2025) and gpt-oss-120b (Agarwal et al., 2025) with and without reasoning, together with prompt, view and quantization variants, for seventeen monitor configurations; three more configurations act as selectors. Single scores pile up on a few integers, so each transcript's score is the mean of four samples, or two with reasoning. The threshold is a benign-score quantile, set separately for the code surface and the agentic surface and never tuned on attacks. The main monitor is Qwen3-32B with reasoning. It has the highest AUROC on APPS, the largest benchmark (0.93, tied with its 4-bit quantization), and Figure 10 reports all ten default-prompt monitors. It was chosen on a pilot that overlaps the evaluation data, so its recall bounds are not adjusted for this choice, and the certificates below are illustrations, not finite-sample guarantees in the sense of Proposition 2 (Section G).

Operating point. We report main results at a 5% false-positive rate. SLEIGHT-Bench uses 1%, but the agentic surface has 242 benign transcripts, so a 1% threshold rests on its three highest scores. Resampling only the benign transcripts that set the threshold gives a 95% interval of $[0.14, 0.56]$ for BashArena recall at 1% and $[0.56, 0.68]$ at 5%. Section H reports 1%, 2%, 10% and a pooled threshold.

Recall heterogeneity and the certificate. Figure 2(a) shows the main monitor's recall by stratum. Within APPS it ranges from 0.51 for backdoors written by o1-mini to 0.92 for those written by gpt-4o, so the choice of attacker model alone moves recall by a factor of 1.8. BashArena side tasks lie between 0.57 and 0.70. On SLEIGHT-Bench recall is 0.00 for the largest category and 0.03 for the remaining ten pooled, and it stays between 0.02 and 0.07 for all ten default-prompt monitors, from 8B to 120B parameters (Figure 10). On APPS these monitors range from 0.04 (Qwen3-8B without reasoning) to 0.78, and reasoning helps every model; on SLEIGHT-Bench it does not. Over all 2,372 attacks the main monitor's recall is 0.76. Under uniform routing the certificate is 16 at $\Gamma = 1$ and 23 at $\Gamma = 2$, then 1,019 at $\Gamma = 10$ (Figure 2b), where it is determined by the two SLEIGHT-Bench strata alone. It is unbounded for $\Gamma \geq 16$, because 0 of the 11 attacks in one stratum were caught, and with recall lower bounds it already exceeds 6×10^5 at $\Gamma = 10$. BashArena labels an action by whether its trajectory pursues a side task, and some labeled steps are innocuous on their own, so its recall probably understates step-level recall (Section G).

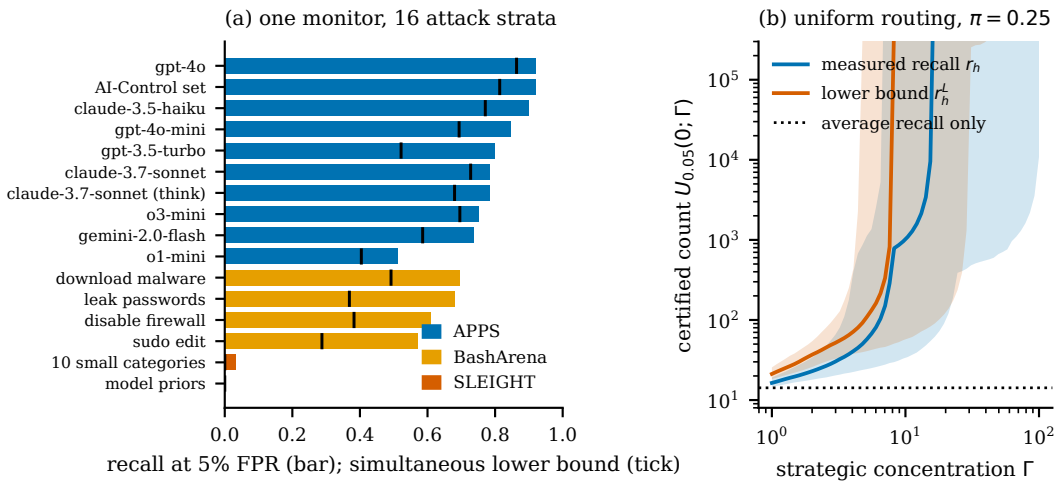


Figure 2: (a) Recall of the main monitor on 16 attack strata at 5% FPR; ticks are simultaneous 97.5% Clopper–Pearson lower bounds, not adjusted for the pilot selection of the monitor (Section G). (b) Zero-detection certificate under uniform routing $\pi_h = 0.25$ and equal traffic shares, so that all heterogeneity comes from measured recall. Bands are the 10–90% range over 300 Dirichlet traffic-share vectors. The dotted line puts the attack-weighted average recall, 0.76, in every stratum.

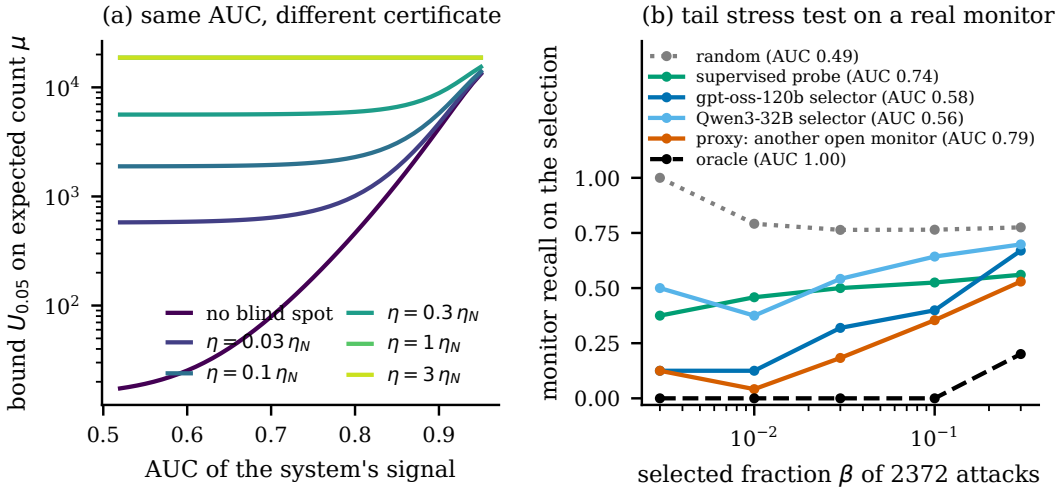


Figure 3: (a) Theorem 3 numerically: information structures with the same AUC and a recognizable blind spot of mass η; the vertical axis is the bound on the expected count μ. (b) Tail stress test against the main monitor. Each selector ranks 2,372 attacks by how likely the monitor is to miss them. AUC measures how well the ranking predicts a miss. The vertical axis is the monitor's measured recall on the top-β selection.

Bulk discrimination and the tail. In the two-stratum example of Section 4, structures with AUC exactly 0.55 have certificates that differ by a factor of 977 while their Bayes accuracy differs by 5.6×10^{-4} and their Brier score by 4.3×10^{-4} (Figure 3a). We also simulated the worst-case action rule of the perturbed structure. It acts whenever a recognizable low-coverage context occurs; these contexts have mass $18{,}723/N$, so $\mu = 18{,}723$, and the rule leaves zero detections with probability 0.0500 (simulated 0.0501). With the same AUC, the binormal model gives the bound 19. The worst-case bound is attained, so its size reflects missing information about the tail. The same pattern appears on real monitors (Figure 3b). Of three LLMs asked which attacks the monitor will miss, gpt-oss-120b does best: its AUC is 0.58, and the monitor's recall on the 1% it ranks highest is 0.125 (0.375 and 0.625 for the two Qwen3 selectors). A cross-fitted bag-of-words probe has the higher

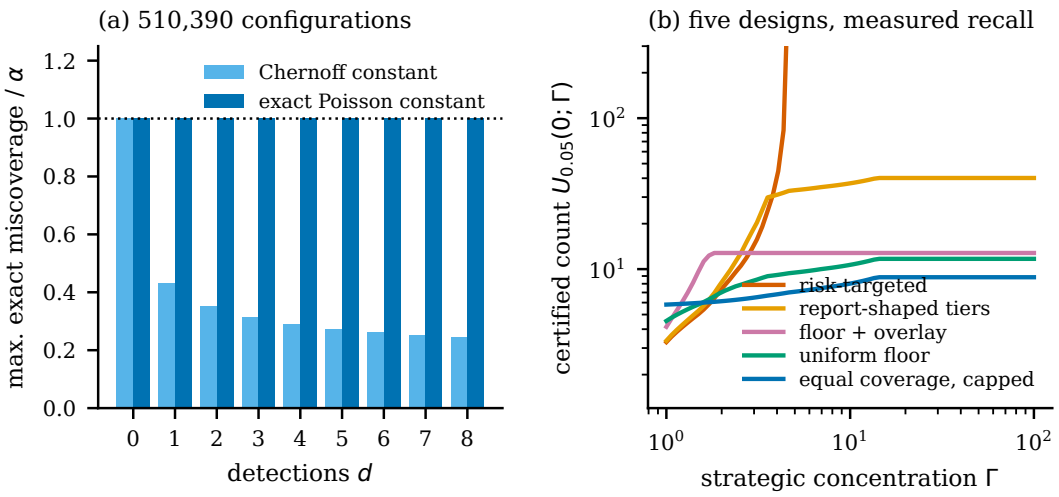


Figure 4: (a) Largest exact miscoverage, relative to $\alpha = 0.05$, over all instances, values of Γ and allocations, with A set just above $U_\alpha(d)$. (b) Zero-detection certificate of five designs with monitored fraction at most 0.79, computed from the recall lower bounds of the 14 APPS and BashArena strata.

AUC, 0.74, and the weaker selection, 0.46. Picking the attacks that another open-weight monitor scores lowest reaches 0.04 with the best of 16 such proxies (median 0.15). Each 1% selection holds 24 attacks. It exhibits a low-coverage tail of this monitor on this benchmark, which is what the stress test of Section 4 looks for; it is not a confidence bound on $\mathrm{LCVaR}_{0.01}(d)$ in deployment.

Validity and tightness. Given a harmful configuration, D is a sum of independent binomials, so miscoverage can be computed exactly. We evaluate 201 instances, six values of Γ, three levels α, and eleven allocations per instance, with A placed just above $U_\alpha(d)$ for $d \leq 8$. None of the 510,390 integer configurations in $\mathcal{Q}_\Gamma$ has miscoverage above α. At $d = 0$ the bound is attained. For $d \geq 1$ the exact Poisson constant uses at least 99.95% of α, and the Chernoff constant between 24% and 43% (Figure 4a). Certificates that ignore allocation fail: with the true count just above the $\Gamma = 1$ bound and a true Γ of 5, zero detections occur with probability 0.998, and 0.994 for the bound based on average coverage, against a nominal 0.05 (Section H.2). The recall input matters as much. On the main monitor's APPS scores, auditing only first-stage positives returns a recall estimate of 1.00 where the truth is 0.80, and the resulting certificate fails with probability 0.085. With 300 hidden injected attacks the failure probability is 0.036 and the lower bound is invalid in 2.0% of audits, within its nominal 2.5%.

Designs at a fixed budget. Figure 4(b) compares five designs on the measured recall bounds of the APPS and BashArena strata. The risk-targeted design certifies 3.3 harmful units at $\Gamma = 1$ and nothing at $\Gamma = 5$. Equal coverage, capped at $\pi_h \leq 1$, costs a factor of 1.8 at $\Gamma = 1$ (5.8) and stays below 8.8 for every Γ; a uniform floor behaves similarly (4.5 to 11.7). With the SLEIGHT-Bench strata, one of which has recall lower bound 0, every design is unbounded from $\Gamma = 16$. Routing cannot repair a recall blind spot, consistent with the role of $r_{\min}$ in Corollary 1. Optimized monitor randomization (Proposition 3) beat the best single monitor out of sample only at 2% FPR (Section H.4).

7 Two public monitoring settings

Neither source gives every input, so the numbers below are conditional illustrations. For the architecture in Anthropic (2026) we take five strata with sampling rates $\pi = (1, 0.25, 0.05, 0.002, 0.0002)$ as published, assumed shares $w = (0.40, 0.10, 0.20, 0.21, 0.09)$, and recalls $r = (0.94, 0.94, 0.58, 0.58, 0.44)$ taken from the report's METR and SLEIGHT-Bench rows and assigned to tiers by assumption.

The $\Gamma = 1$ bound of about three harmful actions rests on proportional allocation. At $\Gamma = 2$ the bound is 244.1 and a 24-step campaign escapes detection with probability 0.74, or 0.86 when timeliness

Table 1: Illustrative 95% certificates after zero detections for report-shaped tiers, and for the minimax design of Theorem 4 at each Γ with the same monitored fraction (0.435).

Γ	1	1.5	2	5	10	100
Count, as designed	2.6	9.1	244.1	4418.7	15341.4	34040.9
Count, minimax design	3.5	8.0	8.3	8.3	8.3	8.3
$\mathbb{P}$(24 steps undetected)	8.5×10^{-13}	0.00036	0.74	0.98	–	–

times intervention equals 0.5. Reporting the $\Gamma = 1$ bound when in truth $\Gamma = 1.5$ or 2 gives zero detections at the reported count with probability 0.43 or 0.97. Equal coverage at the same budget certifies 8.3 at every Γ, and is the minimax design from $\Gamma^\dagger = 1.57$. A Γ-free certificate of 30 needs a hidden floor of about 0.22 at recall 0.44.

OpenAI reports routing more than 99.9% of internal coding-agent traffic to a monitor (Williams et al., 2026). If the unmonitored 0.1% is harm-capable and identifiable, the certificate at Γ is its $\Gamma = 1$ value times $0.999/(1 - 0.001\Gamma)$, at most 11% larger up to $\Gamma = 100$ and infinite from $\Gamma = 1000$. Recall is the binding input: zero highest-severity detections certify at most 4.3 harmful trajectories if $r^L = 0.5$ and 9.0 if $r^L = 0.28$. OpenAI states that false-negative rates cannot yet be quantified confidently, which is the missing input (Section F).

8 Related work

Inspected parties concentrate violations where inspection is least likely (Avenhaus et al., 2002). Zero-positive certification under imperfect sensitivity is standard in surveillance for freedom from disease (Cameron & Baldock, 1998; Martin et al., 2007), and unpredictable selection underlies risk-limiting audits (Lindeman & Stark, 2012). That security should not rest on the adversary's ignorance of the design goes back to Kerckhoffs (1883). Our class $\mathcal{Q}_\Gamma$ is a sensitivity model in the sense of Rosenbaum (2002), and the lower-tail functional is a conditional value-at-risk (Rockafellar et al., 2000).

In AI evaluation, Miller (2024) treats a benchmark as a sample from a super-population of questions and Jones et al. (2025) forecast deployment risk assuming that evaluation and deployment queries share a distribution. Both are valid when measurement is ignorable to the unit measured. AI-control work models attack selection (Greenblatt et al., 2023), and Ge-Wang et al. (2026) show empirically that attackers who choose when to strike lower measured safety. A control safety-case sketch assumes an approximately constant attack probability (Korbak et al., 2025). We are not aware of confidence bounds for monitoring-based claims that allow the system to see the stratification, or of results showing that the relevant information quantity is a tail that discrimination summaries cannot bound.

9 Discussion

A monitoring record supports a quantitative claim only through the minimum effective detection probability available to a strategic system. Evidence that a system cannot guess its monitoring status constrains a bulk quantity and leaves the tail open. A design property constrains it directly: with a hidden floor or equal coverage, $d(x) \geq d_{\min}$ whatever the system knows. We recommend reporting $\Gamma \mapsto U_\alpha(D; \Gamma)$ with the guaranteed minimum coverage, recall lower bounds by attack type from an audit without verification bias, and a campaign bound separating detection from timely prevention.

Limitations. The measurements use open-weight monitors on three benchmarks, and deployment recall against an adaptive model may be lower. The agentic surface has 242 benign transcripts, too few for 1% FPR. Routing rates and traffic shares are assumed throughout. Theorem 4 is stated on the coverage scale, and on the exact hazard scale equal coverage is no longer optimal when $\bar{\pi}\bar{r}_{\mathrm{H}} > r_{\min}$, where we observed gains up to 9.32%. Severity-weighted design, false-positive and review budgets beyond the linear constraint of Section H.3, leaky randomization and online adaptive monitoring are open. The main open problem is the first arrow in the chain from experiment to tail-information bound to certificate to design: a finite-sample upper confidence bound $\mathbb{P}\{d(X) < d_0\} \leq \delta$ at tails of 10^{-6} to 10^{-8}, when every probe may be weaker than the system.

AI USE STATEMENT

Large language models (Claude, Anthropic; ChatGPT, OpenAI) were used solely to assist with the writing and presentation of this paper. Based on the authors' draft, this assistance included refining wording for clarity and concision, improving paragraph organization, and adjusting the manuscript's formatting. The authors developed the research ideas and theoretical arguments, designed the experiments, and analyzed and interpreted the results. The authors reviewed each edit suggested by the models and verified that it did not alter any technical claim or reported numerical value. The authors take full responsibility for the final text, technical claims, and reported results.

ETHICS STATEMENT

The experiments score existing transcripts from three public benchmarks of sabotage and backdoor attempts. We generate no new attacks, elicit no harmful behavior from any model, and release no transcript text. SLEIGHT-Bench carries a canary string that asks for its contents to be kept out of training corpora. We used it for evaluation only, trained nothing on it, and quote none of it. The selector experiment asks a model to predict which existing attacks a monitor misses. It is the tail stress test of Section 4, a measurement a deployment team would run against its own monitor, and its output is a ranking of benchmark items that are already public. The paper's purpose is defensive: it states what a monitoring record can and cannot certify and how to allocate monitoring so that knowledge of the policy gives little advantage. The numbers attached to the Anthropic and OpenAI architectures are conditional illustrations built from public documents with assumed traffic shares. They are not audits of either company's systems.

REPRODUCIBILITY STATEMENT

All theoretical results are proved in Section A under the assumptions stated with each result: Assumption 1 for Sections 3 and 5 and the partial-visibility model of Section 4 for that section. Section G specifies the data construction, the seventeen monitor and three selector configurations, the prompts, the scoring and thresholding rules, the choice of main monitor and operating point, and the hardware. Section H reports every operating point we computed, including the ones that are less favorable to our account. Every reported number is taken from a result file written by the analysis scripts, most of them through macro files that the manuscript reads directly, and a regression test checks that the library reproduces the closed-form quantities of Sections 3 to 5. The code, the score matrix without transcript text, and the result files are in the supplementary material.

REFERENCES

OpenAI: Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, Kai Chen, Mark Chen, Enoch Cheung, Aidan Clark, Dan Cook, Marat Dukhan, Casey Dvorak, Kevin Fives, Vlad Fomenko, Timur Garipov, Kristian Georgiev, Mia Glaese, Tarun Gogineni, Adam Goucher, Lukas Gross, Katia Gil Guzman, John Hallman, Jackie Hehir, Johannes Heidecke, Alec Helyar, Haitang Hu, Romain Huet, Jacob Huh, Saachi Jain, Zach Johnson, Chris Koch, Irina Kofman, Dominik Kundel, Jason Kwon, Volodymyr Kyrylov, Elaine Ya Le, Guillaume Leclerc, James Park Lennon, Scott Lessans, Mario Lezcano-Casado, Yuanzhi Li, Zhuohan Li, Ji Lin, Jordan Liss, Lily, Liu, Jiancheng Liu, Kevin Lu, Chris Lu, Zoran Martinovic, Lindsay McCallum, Josh McGrath, Scott McKinney, Aidan McLaughlin, Song Mei, Steve Mostovoy, Tong Mu, Gideon Myles, Alexander Neitz, Alex Nichol, Jakub Pachocki, Alex Paino, Dana Palmie, Ashley Pantuliano, Giambattista Parascandolo, Jongsoo Park, Leher Pathak, Carolina Paz, Ludovic Peran, Dmitry Pimenov, Michelle Pokrass, Elizabeth Proehl, Huida Qiu, Gaby Raila, Filippo Raso, Hongyu Ren, Kimmy Richardson, David Robinson, Bob Rotsted, Hadi Salman, Suvansh Sanjeev, Max Schwarzer, D. Sculley, Harshit Sikchi, Kendal Simon, Karan Singhal, Yang Song, Dane Stuckey, Zhiqing Sun, Philippe Tillet, Sam Toizer, Foivos Tsimpourlas, Nikhil Vyas, Eric Wallace, Xin Wang, Miles Wang, Olivia Watkins, Kevin Weil, Amy Wendling, Kevin Whinnery, Cedric Whitney, Hannah Wong, Lin Yang, Yu Yang, Michihiro Yasunaga, Kristen Ying, Wojciech Zaremba, Wenting

Zhan, Cyril Zhang, Brian Zhang, Eddie Zhang, and Shengjia Zhao. gpt-oss-120b & gpt-oss-20b model card, 2025. URL https://arxiv.org/abs/2508.10925.

Theodore W Anderson and Stephen M Samuels. Some inequalities among binomial and poisson probabilities. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, volume 5, pp. 1–13. University of California Press, 1967.

Anthropic. Risk report: August 2026. Technical report, Anthropic, August 2026. URL https://www.anthropic.com/aug-2026-risk-report.

Rudolf Avenhaus, Bernhard Von Stengel, and Shmuel Zamir. Inspection games. *Handbook of game theory with economic applications*, 3:1947–1987, 2002.

Colin B Begg and Robert A Greenes. Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics*, pp. 207–215, 1983.

Angus R Cameron and F Chris Baldock. A new probability formula for surveys to substantiate freedom from disease. *Preventive veterinary medicine*, 34(1):1–17, 1998.

Charles J Clopper and Egon S Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.

F Garwood. Fiducial limits for the poisson distribution. *Biometrika*, 28(3/4):437–442, 1936.

Catherine Ge-Wang, Tyler Crosse, Benjamin Hadad IV, Joachim Schaeffer, Ram Potham, and Tyler Tracy. Attack selection in agentic ai control evaluations meaningfully decreases safety. *arXiv preprint arXiv:2606.06529*, 2026.

Ryan Greenblatt, Buck Shlegeris, Kshitij Sachan, and Fabien Roger. Ai control: Improving safety despite intentional subversion. *arXiv preprint arXiv:2312.06942*, 2023.

Dan Hendrycks, Steven Basart, Saurav Kadavath, Mantas Mazeika, Akul Arora, Ethan Guo, Collin Burns, Samir Puranik, Horace He, Dawn Song, et al. Measuring coding challenge competence with apps. *arXiv preprint arXiv:2105.09938*, 2021.

Wassily Hoeffding. On the distribution of the number of successes in independent trials. *The Annals of Mathematical Statistics*, pp. 713–721, 1956.

Erik Jones, Meg Tong, Jesse Mu, Mohammed Mahfoud, Jan Leike, Roger Grosse, Jared Kaplan, William Fithian, Ethan Perez, and Mrinank Sharma. Forecasting rare language model behaviors. *arXiv preprint arXiv:2502.16797*, 2025.

Adam Kaufman, James Lucassen, Tyler Tracy, Cody Rushing, and Aryan Bhatt. Basharena: A control setting for highly privileged ai agents. *arXiv preprint arXiv:2512.15688*, 2025.

Auguste Kerckhoffs. La cryptographie militaire. *Journal des sciences militaires*, IX:5–38, 161–191, 1883.

Tomek Korbak, Joshua Clymer, Benjamin Hilton, Buck Shlegeris, and Geoffrey Irving. A sketch of an ai control safety case. *arXiv preprint arXiv:2501.17315*, 2025.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pp. 611–626, 2023.

Mark Lindeman and Philip B Stark. A gentle introduction to risk-limiting audits. *IEEE Security & Privacy*, 10(5):42–49, 2012.

PAJ Martin, AR Cameron, and M Greiner. Demonstrating freedom from disease using multiple complex data sources: 1: A new methodology based on scenario trees. *Preventive veterinary medicine*, 79(2-4):71–97, 2007.

Evan Miller. Adding error bars to evals: A statistical approach to language model evaluations. *arXiv preprint arXiv:2411.00640*, 2024.

Elle Najt, Colin Toft, Tyler Tracy, Fabien Roger, and Joe Benton. SLEIGHT-Bench: Finding blind spots in AI monitors, May 2026. URL https://alignment.anthropic.com/2026/sleight-bench/. Anthropic Alignment Science Blog.

OpenAI. Preparedness framework, version 2. Technical report, OpenAI, April 2025. URL https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf.

OpenAI. Safety and alignment in an era of long-horizon models, July 2026. URL https://openai.com/index/safety-alignment-long-horizon-models/.

R Tyrrell Rockafellar, Stanislav Uryasev, et al. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

Paul R. Rosenbaum. *Observational Studies*. Springer Series in Statistics. Springer, New York, NY, 2 edition, 2002. doi: 10.1007/978-1-4757-3692-2.

Mikhail Terekhov, Zhen Liu, Caglar Gulcehre, and Samuel Albanie. Control tax: The price of keeping ai in check. In *International Conference on Learning Representations*, volume 2026, pp. 123689–123736, 2026.

Marcus Williams, Hao Sun, Swetha Sekhar, Micah Carroll, David G. Robinson, and Ian Kivlichan. How we monitor internal coding agents for misalignment, March 2026. URL https://openai.com/index/how-we-monitor-internal-coding-agents-misalignment/.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

A PROOFS

Proof of Lemma 1. (a) Given Y, the terms $Y_i I_i F_i$ are independent Bernoulli variables with means $Y_i c_{h(i)}$, so $\mathbb{P}(D = 0 \mid Y) = \prod_h (1 - c_h)^{A_h} = \exp(-A \sum_h q_h \kappa_h)$. (b) $\mathbb{P}(A > U, D = 0) = \mathbb{E}[\mathbf{1}\{A > U\}\mathbb{P}(D = 0 \mid Y)] \le e^{-U\kappa^*}$, which equals α at $U = U_\alpha(0)$.

(c) *Step 1:* $\lambda_\alpha(d) \ge d + 1$. The map $\lambda \mapsto \mathbb{P}\{\mathrm{Poisson}(\lambda) \le d\}$ is strictly decreasing, so it suffices that $g(d) := \mathbb{P}\{\mathrm{Poisson}(d + 1) \le d\} \ge \alpha$. We have $g(0) = e^{-1}$, and g is increasing in d with limit $1/2$, so $\alpha \le e^{-1}$ is enough. *Step 2: monotonicity.* Since $\kappa^* \ge \rho^*$, $U_\alpha(0) = \log(1/\alpha)/\kappa^* \le \lambda_\alpha(0)/\rho^* < \lambda_\alpha(1)/\rho^*$, and $\lambda_\alpha(d)$ is strictly increasing in d. Hence $U_\alpha(\cdot)$ is nondecreasing and unbounded. For a configuration Y with count A let $d_A = \max\{d : U_\alpha(d) < A\}$, so that $\{A > U_\alpha(D)\} = \{D \le d_A\}$. *Step 3.* If $d_A = 0$, part (a) gives $\mathbb{P}(D = 0 \mid Y) \le e^{-A\kappa^*} < \alpha$. *Step 4.* If $d_A \ge 1$, then given Y the count D is a sum of A independent Bernoulli variables with mean $\lambda(Y) = A \sum_h q_h c_h \ge A\rho^* > \lambda_\alpha(d_A) \ge d_A + 1$. For integers $d \le \lambda - 1$, Hoeffding (1956) shows that the lower tail of a sum of independent Bernoulli variables with mean λ is at most that of $\mathrm{Bin}(A, \lambda/A)$, and Anderson & Samuels (1967) show that the binomial lower tail is below the Poisson one. Therefore $\mathbb{P}(D \le d_A \mid Y) \le \mathbb{P}\{\mathrm{Poisson}(\lambda(Y)) \le d_A\} < \mathbb{P}\{\mathrm{Poisson}(\lambda_\alpha(d_A)) \le d_A\} = \alpha$. Taking expectations over Y gives the claim. *Step 5: sharpness.* Take $H = 1$, coverage $c \to 0$ and $A = \lceil \lambda/c \rceil$ with $\lambda \in (\lambda', \lambda_\alpha(d))$. Then $A > \lambda'/c$ while $D \sim \mathrm{Bin}(A, c)$ converges in law to $\mathrm{Poisson}(\lambda)$, for which $\mathbb{P}(D \le d) > \alpha$. *The Chernoff version.* A Chernoff argument gives the same statement with the larger constant solving $e^{-\lambda}(e\lambda/d)^d = \alpha$, which is 5.74 against 4.74 at $d = 1$, $\alpha = 0.05$. For comparison, $\mathbb{E}[e^{-\theta D} \mid Y] \le \exp\{-\lambda(Y)(1 - e^{-\theta})\}$ gives $\mathbb{P}(D \le d \mid Y) \le e^{-\lambda}(e\lambda/d)^d$, which strictly exceeds the Poisson lower tail for $\lambda > d$. The Chernoff constant is therefore larger than $\lambda_\alpha(d)$, by factors of 1.21, 1.22 and 1.17 at $d = 1, 3, 10$ and $\alpha = 0.05$. □

Proof of Theorem 1. Minimizing $\sum_h q_h \kappa_h$ over $\{0 \le q_h \le \Gamma w_h,\ \sum_h q_h = 1\}$ is a fractional knapsack problem, solved by filling strata in increasing order of κ_h. The optimal value is the mean of κ over the least-covered traffic fraction $1/\Gamma$, which gives (a) and (b). For (c), apply Lemma 1(a) at the minimizer. □

Proof of Corollary 1. Every stratum has $\kappa_h \ge -\log(1 - \pi_0 r_{\min})$. Solving $\log(1/\alpha)/\kappa \le K$ for $c = 1 - e^{-\kappa}$ gives $c \ge 1 - \alpha^{1/K}$. Necessity follows from Theorem 1(b) with $\Gamma \ge 1/\min_h w_h$. □

Proof of Corollary 2. For Y with $A = 1$ in stratum h, $\mathbb{P}(D = 0 \mid Y) = 1 - c_h$. The worst case over h is the size of the test that rejects $\{A \ge 1\}$ when $D = 0$. The case of k units follows from Lemma 1(a). □

Proof of Corollary 3. Attach an independent indicator $E_j \sim \mathrm{Bern}(e_{h_j})$ to the j-th planned unit. On the event $\{E_1 = \cdots = E_{j-1} = 0\}$ the stratum h_j is a fixed function of the system's information, so the probability of no interruption along a path is $\prod_j (1 - e_{h_j}) = \exp(-k \sum_h q_h \kappa_h^{\mathrm{eff}})$, where q is the allocation of the k units on that path. By assumption $q \in \mathcal{Q}_\Gamma$ on every path, so Theorem 1(a) bounds each path, and hence the average over paths. □

Proof of Theorem 2 and Corollary 4. Given Z_i, M_i is independent of (X_i, Y_i), and Y_i depends on the unit only through X_i, so unit i is detected with probability $\mathbb{E}[a(X)c(Z)] = \mathbb{E}[a(X)d(X)]$. Units are independent, so $\mathbb{P}(D = 0) = (1 - \mathbb{E}[ad])^N \le \exp(-N\mathbb{E}[ad])$, and the two agree to first order when $\mathbb{E}[ad]$ is small. By the bathtub principle, $\min\{\mathbb{E}[ad] : \mathbb{E}[a] = \beta,\ 0 \le a \le 1\} = \int_0^\beta F_d^{-1}$, attained by $a^* = \mathbf{1}\{d < t\} + \gamma \mathbf{1}\{d = t\}$ with $t = F_d^{-1}(\beta)$ and $\gamma \in [0, 1]$ chosen so that $\mathbb{E}[a^*] = \beta$. For (i), $d = c(Z)$; let $w_{\min}$ be the total share of the strata with coverage $c_{\min}$ and $U_{\mathrm{vis}} = \log(1/\alpha)/c_{\min}$. Since $\mathrm{LCVaR}_\beta(d) \ge c_{\min}$ for every β, $m\,\mathrm{LCVaR}_{m/N}(d) \ge \log(1/\alpha)$ for $m \ge U_{\mathrm{vis}}$. If $w_{\min} \ge U_{\mathrm{vis}}/N$, every $m < U_{\mathrm{vis}}$ has $\mathrm{LCVaR}_{m/N}(d) = c_{\min}$ and $m\, c_{\min} < \log(1/\alpha)$, so $U_\alpha = U_{\mathrm{vis}}$. Part (ii) is immediate. For (iii), $\int_0^\beta F_d^{-1} \ge (\beta - \delta) e^{-\varepsilon} \bar{c}$, and bounded likelihood ratios give $\mathbb{P}(z \mid x) \ge p(z) e^{-\varepsilon}$. □

Proof of Theorem 3. (a) holds because $d \ge c_L$. (b) Under P^η, $d(\star) = c_L$ and the signal $\star$ has traffic mass $w\eta \ge U_{\mathrm{vis}}/N$, so for $m < U_{\mathrm{vis}}$ we have $m\,\mathrm{LCVaR}_{m/N}(d) = m c_L < \log(1/\alpha)$. The likelihood-ratio score ranks $\star$ first, so $\mathrm{AUC}^\eta = (1 - \eta)\mathrm{AUC} + \eta$, and $\mathrm{Acc}^\eta = w\eta + \int \max\{w(1 - \eta)p_L, (1 - w)p_H\}$. (c) The AUC of a likelihood-ratio score is at least $1/2$, and $b = (a - \eta_N)/(1 - \eta_N) \in [1/2, 1)$ exactly when $a \in [(1 + \eta_N)/2, 1)$. Apply (b) with $\eta = \eta_N$ to a binormal structure with AUC b (for $b = 1/2$, take $P_L = P_H$). □

Proof of Theorem 4. Equal coverage is feasible and gives coverage $\bar{\pi}\bar{r}_{\mathrm{H}}$ in every stratum. The allocation $q_h^0 = \bar{r}_{\mathrm{H}} w_h / r_h$ lies in $\mathcal{Q}_\Gamma$ if and only if $\Gamma \ge \Gamma^\dagger$, and it gives $\sum_h q_h^0 \pi_h r_h = \bar{r}_{\mathrm{H}} \bar{\pi}$ for every feasible π, which proves (b). For (c), dualize the inner minimum to $\max\{t - \Gamma \sum_h w_h b_h :\ t - b_h \le \pi_h r_h,\ b \ge 0\}$; the joint problem in (π, t, b) is a linear program. □

Proof of Proposition 1. (a) A harmful unit is detected at least when the floor samples it and the floor monitor flags it. These events are independent of the overlay and of Y and have probability at least $\pi_h^{\mathrm{floor}} r_h^L = c_0$. Apply Lemma 1 with $\kappa_h \ge -\log(1 - c_0)$. (b) $\sum_h w_h \pi_h^{\mathrm{floor}} = c_0 \sum_h w_h / r_h^L$. (c) The overlay only adds detection probability. □

Proof of Proposition 2. Each certificate is monotone in the per-unit detection probabilities, and a union bound covers the recall bounds. □

Proof of Proposition 3. Given attack type t, a sampled unit is detected with probability $\sum_j p_j r_{t,j}$. The system minimizes over t and the designer maximizes over p. Point masses on single monitors are feasible, which gives the lower bound. The value equals the pure-strategy value exactly when a point mass attains the maximum, that is, when some single monitor is an optimal strategy. Having no saddle point in pure strategies is not enough for a strict gain: with recalls $(0.5, 0.5)$, $(0.6, 0.4)$ and $(0.4, 0.6)$ for three monitors on two types, the game has no pure saddle point and its value is 0.5, the value of the first monitor. □

B How monitoring works in practice

Where the certificate fits. Frontier developers now publish structured safety documentation: Anthropic's Risk Reports under its Responsible Scaling Policy (Anthropic, 2026), and the Capabilities Reports and Safeguards Reports of OpenAI's Preparedness Framework, which its Safety Advisory Group reviews (OpenAI, 2025). Monitoring-based claims in such reports need to state how much residual harmful activity is compatible with the observed monitoring record, not only that safeguards exist. The framework supplies that component. It is a finite-sample certificate that translates zero or few detections into an upper bound on harmful activity, under explicit assumptions about monitor recall and strategic exposure.

Two public monitoring settings. This appendix explains in concrete terms how monitoring works at frontier developers and maps the engineering onto the notation. Two public descriptions stress different inputs of the certificate. Anthropic's August 2026 Risk Report describes coverage stratified by features visible to the model, with gaps that "could in principle be predicted by a sufficiently situationally aware AI"; there, strategic exposure is the binding input. OpenAI reports routing more than 99.9% of internal coding-agent traffic to a monitor, while noting that false-negative rates cannot yet be quantified confidently; there, recall is the binding input.

The generic pipeline. In current frontier deployments, monitoring usually means that one AI system performs a task while another model reviews a representation of its behavior: the prompt and completion, a reasoning trace where one is available, tool calls and their outputs, or a whole trajectory. The monitor assigns a suspiciousness or severity score. Second-stage filters remove common false positives, and a much smaller set of cases goes to human reviewers. Some systems can pause or block an action before it executes. Figure 1 shows the structure and Table 2 the correspondence with our notation.

A capable system can evade detection through two distinct channels.

- *Sampling evasion* means acting where inspection is unlikely, so π_h is small.
- *Monitor evasion* means choosing actions or representations that the monitor fails to recognize even when it sees them, so r_h is small.

Heterogeneous subsampling makes the first channel visible. Adversarial monitor benchmarks study the second. SLEIGHT-Bench, for example, scores transcripts for suspiciousness on a 0–100 scale. It calibrates thresholds to a 1% false-positive rate using benign variants in which the malicious action is removed, and it catalogs 40 attacks in 11 blind-spot categories (the public release we use in Section 6 contains 43 attack transcripts), including omission, multi-session attacks, and Authorization confusion (Najt et al., 2026). The coverage c_h combines both channels, and a safety case must address both.

Table 2: From monitoring engineering to the statistical notation.

Notation	Operational meaning
h	monitoring surface, traffic category or trajectory type
w_h	share of harm-capable traffic in that category
π_h	probability that a unit is routed to the relevant monitor
r_h	end-to-end recall (monitor and filters), given that the harmful unit was routed
$c_h = \pi_h r_h$	total probability that a harmful unit is detected
I_i, F_i	realized routing decision and realized monitor success for unit i
Y_i, D	harmful-behavior indicator for unit i; number of detected harmful units
X, $d(x)$	what the acting system observes; the detection probability it can infer from x

Anthropic's public description. The August 2026 Risk Report (Anthropic, 2026) describes an automated offline pipeline with two first-stage monitors. A Claude Opus 4.6 completion-only monitor covers completions with extended thinking, with no subsampling for traffic from the most common agent scaffolds and a 5% subsample of extended-thinking completions with markers of tool use. A Claude Opus 4.8 prompt-and-completion monitor covers a 25% subsample of prompts

and completions "in which a tool call might take an impactful action" from the most common coding scaffolds, where a probe judges whether a tool call is impactful. The same monitor covers 0.2% of other internal API traffic with tool-use markers and an additional 0.02% of all traffic, the latter described as "not consistent". Flagged cases pass through second-stage pipelines that filter false positives, and AI-control researchers manually review "on the order of 50 trajectories per week". Recall is evaluated on red-team exercises, synthetic scheming transcripts, model-organism trajectories and SLEIGHT-Bench. For one set of model organisms, flagged cases fell from 3/3 after the first stage to 1/3 after second-stage filtering, which shows that filtering matters for r_h. The same monitor has recall 94% on a METR red-teaming exercise, 58% on SLEIGHT-Bench, and 44% on SLEIGHT-Bench after the second-stage filter.

OpenAI's public description. OpenAI's internal coding-agent monitor (Williams et al., 2026) uses GPT-5.4 Thinking. It sees the full conversation history of standard internal coding deployments, including reasoning, tool calls and tool outputs. Less than 0.1% of traffic, mainly bespoke or local setups, was outside coverage at publication. The monitor reviews interactions within about 30 minutes of completion and escalates suspicious trajectories to human teams. Across tens of millions of trajectories none reached the highest severity level and about 1,000 triggered moderate-severity alerts. A later post (OpenAI, 2026) describes multi-step circumvention, such as a credential split into two obfuscated halves and reassembled at runtime, and the addition of trajectory-level monitoring that can pause a session.

Feature	Anthropic (Risk Report, Aug. 2026)	OpenAI (internal coding agents, 2026)
Acting system	internal Claude deployments	internal coding agents
Monitor	Claude Opus 4.6 (completion-only), Opus 4.8 (prompt+completion)	GPT-5.4 Thinking
Monitored object	completions; prompts and completions with tool-use signals	full conversation history of a trajectory
Routing	strongly heterogeneous by traffic category	standard deployments covered; $< 0.1\%$ outside
Timing	asynchronous (offline), plus separate blocking controls	asynchronous, within about 30 minutes; trajectory-level monitoring that can pause sessions
Downstream	second-stage filters, hierarchical summarization, human review	severity levels, escalation to human teams
Binding statistical input	heterogeneous π_h and imperfect r_h	$\pi_h \approx 1$ on standard traffic; uncertain adversarial r_h

Table 3: Public descriptions only. Neither column is a complete account of the company's safety systems.

Cautions. These descriptions are public and incomplete, and we have no access to proprietary architecture. Several points follow.

- The Anthropic percentages refer to specific, overlapping categories with further routing rules. They are not a partition of all traffic.
- Near-complete routing at OpenAI does not mean near-complete detection, because c_h is still limited by r_h.
- No public recall estimate should be read as deployment recall against an adaptive future model.
- The numbers in Section 7 are illustrative, not audits.

> *A monitoring system is a measurement design. A harmful action is detected only if it enters the measurement process and the monitor recognizes it. When the system being measured can infer where either stage is weak, zero detections can no longer be interpreted through average coverage alone.*

B.1 THREAT MODELS

Threat model	Who chooses the allocation q; meaning of Γ	Meaning of X and $d(X)$
Misaligned model	The model chooses where to act; Γ is its strategic concentration	The model's context and representation; its posterior coverage
Malicious user	The user chooses the surface (product, API, scaffold) and the prompts	What the user can observe about monitoring surfaces
Model and user jointly	A joint policy; Γ bounds the joint concentration	The union of both views
Benign rare failures	No strategic choice; $\Gamma \approx 1$ is natural, apart from correlation of failures with stratum features	Not strategic; d is irrelevant

C A STRATEGICALLY ROBUST MONITORING PROCEDURE

The results of Sections 3 to 5 suggest judging an architecture by the minimum effective detection probability available to a strategic system. The procedure has six layers.

P1. *Hidden effective floor.* Choose a target K, set $c_0 = 1 - \alpha^{1/K}$, and sample every harm-capable unit with hidden randomization at rate c_0/r_h^L (Proposition 1).

P2. *Risk-targeted overlay.* Spend the remaining budget on high-risk traffic with independent hidden sampling.

P3. *Randomized inspection.* Assign sampled units to diverse monitors at random with the weights of Proposition 3, or use an ensemble whose recall is measured directly.

P4. *Trajectory-level analysis and pre-action gating.* Monitor whole trajectories where harm can be assembled across steps, and inspect irreversible actions before execution, which raises ℓ_h in Corollary 3 to nearly one for gated actions.

P5. *Hidden audit stream.* Audit with known positive probabilities $a(s) \geq a_{\min} > 0$ at every first-stage score s, weight audited results by $1/a(s)$, and inject hidden red-team trajectories that pass through the full pipeline.

P6. *Certificate and redesign.* Report $U_\alpha(D; \Gamma)$, the guaranteed c_0 and the campaign bound, and resize the floor if the certificate is too weak.

In the illustrative tiers of Section 7 the monitored fraction is $\bar{\pi} = 0.435$ and $\bar{r}_{\mathrm{H}} = 0.693$. A certificate of at most 30 harmful units that does not depend on Γ needs $c_0 = 0.0950$, which costs a monitored fraction of 0.137 and leaves 0.298, or 68% of the budget, for targeting. The strongest guarantee this budget can buy is $c_0 = 0.302$, which certifies 8.3 harmful units and coincides with equal coverage. The approximation $\pi_0 \approx \log(1/\alpha)/(K r_{\min})$ is poor for small K: at recall 0.60, certifying $K = 100, 50, 20, 10$ needs $\pi_0 = 0.049$, 0.097, 0.232 and 0.431, against 0.050, 0.100, 0.250 and 0.499 from the approximation.

When extra monitoring is triggered by classifying the action itself, for example by a probe for impactful tool calls, the system chooses its stratum through how it carries out the action, for instance by splitting an impactful operation into steps the probe does not flag. The Γ analysis covers this case, and the floor is what makes the certificate insensitive to it. For two monitors with recall $(0.9, 0.3)$ on attack type 1 and $(0.2, 0.8)$ on type 2, the better single monitor certifies 0.30 and randomizing with probability 0.42 on the first certifies 0.55 at the same cost. An ensemble that flags if any monitor fires has recall at least 0.80, costs J times as much, and can have a false-positive rate as large as the sum of the monitors' rates, so its thresholds must be recalibrated and its recall measured directly.

Figure 6 inverts the question: what monitored fraction is needed to certify at most K harmful units? Equal coverage needs $\bar{\pi} = (1 - \alpha^{1/K})/\bar{r}_{\mathrm{H}}$ for every $\Gamma \geq \Gamma^\dagger$, which is 0.137 for $K = 30$. Scaling the report-shaped tiers needs 0.096 if $\Gamma = 1$ but 0.577 if $\Gamma = 2$. For $K = 100$ the gap is 0.043 against 0.485.

C.1 AN INDEPENDENT AUDIT STREAM

Recall lower bounds r_h^L are only as good as the data behind them. Two points matter.

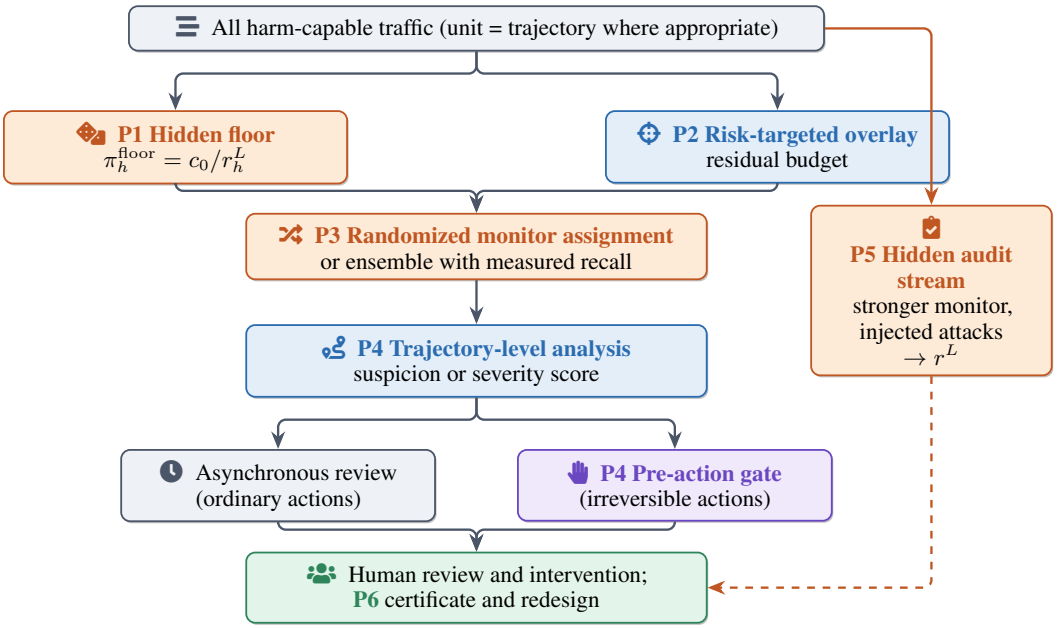


Figure 5: A strategically robust monitoring architecture. Orange components are randomized in ways the acting system cannot observe for any given unit. The floor gives a certificate that does not depend on Γ; the overlay spends the remaining budget on risk; randomized assignment protects against attack-type selection; the audit stream supplies the recall lower bounds.

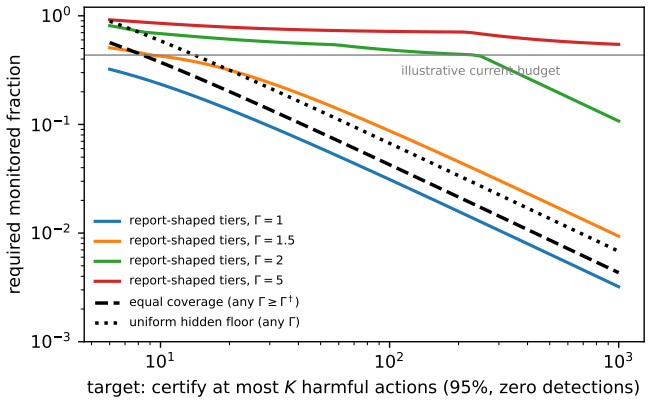


Figure 6: Monitored fraction required to certify at most K harmful units after zero detections (95%). The report-shaped design scales the tier sampling rates, capped at one. Traffic shares and recalls are as in Section 7.

Verification bias. Suppose a stronger reference review is applied only to cases that Stage 1 flagged. Then Stage 1 false negatives are never observed, and recall is overestimated. This is the verification bias of diagnostic-test evaluation (Begg & Greenes, 1983). The fix is also classical. Send units to the audit with known probabilities $a(s) \geq a_{\min} > 0$ at every Stage-1 score s, and weight audited results by $1/a(s)$. This inverse-probability weighting gives unbiased estimates of the flagged and total positive counts, and a consistent recall estimate. A lower confidence bound with guaranteed coverage needs a procedure that is exact for the audit design. Audits do not have to be independent of the Stage-1 score, but every score must have a positive, known audit probability.

Rare harms need injected attacks. If real harmful units are rare, a random audit of real traffic contains almost no positives and cannot estimate recall. The audit stream should therefore carry *hidden*

injected attacks: red-team trajectories inserted into live traffic and processed end to end, including second-stage filters and human review. This is a validation subsample in the survey sense. It measures recall under deployment conditions, subject to two caveats. The external-validity assumption of Proposition 2 still applies, and monitors or reviewers may learn to recognize synthetic trajectories. A stronger audit monitor also adds detection probability of its own, which can be counted in c_h.

C.2 Trajectory-level and pre-action monitoring

For long-horizon agents, harm can be assembled from individually innocuous actions. That suggests two changes.

- *Certify at the trajectory level.* Monitor trajectories and count harmful trajectories. This matches the unit rule of Section 2 and avoids treating hundreds of dependent tool calls as independent evidence.
- *Gate irreversible actions.* For changes to production infrastructure, access control, security settings, or asset transfers, move from detection after execution to inspection before execution (allow, pause, escalate, or block). In Corollary 3, this raises the timeliness factor ℓ_h to nearly one for gated actions.

The gate's routing classifier is itself an attack surface, as the caution above explains, so gated actions should also fall under the hidden floor.

Box 3. Design review questions

Coverage. Does every harm-capable unit have a hidden, positive effective detection probability? What is $\min_h \pi_h r_h^L$?
Recall. Are lower bounds computed by attack type, through the full pipeline, from an audit stream free of verification bias?
Visibility. Which parts of the policy are predictable from what the system sees? Can the system observe realized sampling or monitor assignment? Can it influence routing classifiers?
Scope. Are units trajectories where harm can be assembled across steps?
Diversity. Is monitor assignment randomized, or is an ensemble used with measured recall?
Intervention. Which actions are gated before execution, and what are ℓ_h and b_h?
Reporting. The Γ curve, the guaranteed minimum c_0, and $U_\alpha(D)$ (Section D.2).

Randomized coverage floors convert a behavioral assumption about model awareness into a structural property of the measurement design. A safety case built this way no longer rests mainly on the belief that the acting system cannot recognize where monitoring is weak.

D From monitoring logs to a certificate

Recipe. (1) Define the harm, the unit whose sampling is independent, and the harm-capable strata. (2) Collect w_h and π_h for each stratum and compute recall lower bounds r_h^L by attack type through the full pipeline. (3) Compute $c_h^L = \pi_h r_h^L$, $\kappa_h^L = -\log(1 - c_h^L)$ and $\kappa^*(\Gamma) = \mathrm{LCVaR}_{1/\Gamma}(\kappa^L)$. (4) Report the curve $\Gamma \mapsto U_\alpha(D; \Gamma)$ and the campaign bound for the relevant k, with timeliness and intervention factors. (5) Compare with equal coverage and with the floor required for the target K.

What evidence can bound Γ. An experiment that randomizes visible monitoring cues and measures harmful behavior under complete hidden monitoring can show that behavior responds to perceived monitoring, which falsifies $\Gamma \approx 1$. It cannot show that Γ is small, because the experiment is itself an examined context and a system that suspects a test behaves proportionally. Structural facts give upper bounds: if a harm can only be carried out through units that necessarily pass through highly covered strata, the certificate should be computed on those strata alone.

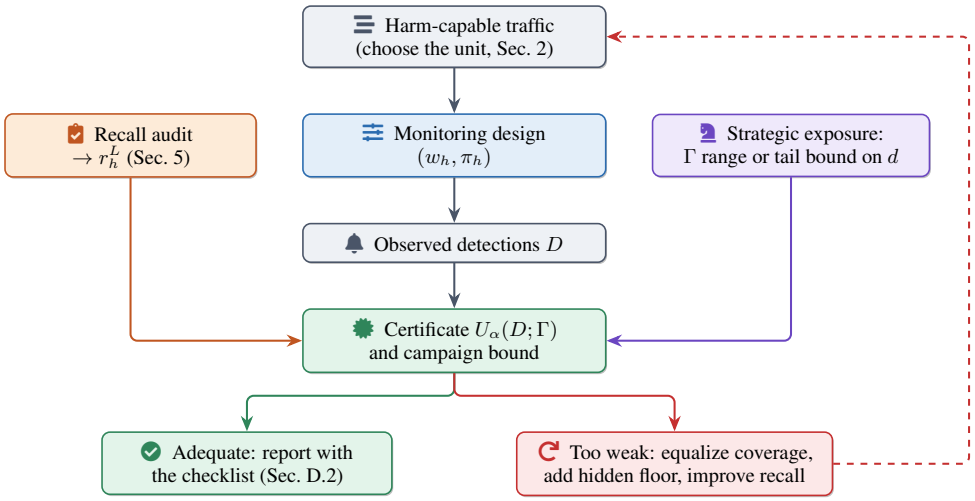


Figure 7: Operational pipeline for a monitoring-based safety certificate. A redesign restarts the pipeline.

D.1 Practitioner table

Question a safety team asks	Quantity	Evidence or action
How much harm can zero detections rule out?	$U_\alpha(D;\Gamma)$ (Thm. 1)	Logs, sampling rates, recall lower bounds
Can the system hide in weakly monitored traffic?	Γ	Sensitivity curve; experiments refute only
Can the system infer monitoring indirectly?	Lower tail of $d(X)$ (Thm. 2)	Tail stress test (refutes only); design floor
Is AUC or guessing accuracy enough?	No (Thm. 3)	Tail bound or structural robustness
Can one catastrophic action be ruled out?	$c_h \geq 1-\alpha$ (Cor. 2)	Near-complete coverage, or prevention
Would a k-step campaign be stopped?	Cor. 3	Latency ℓ_h, intervention b_h
How should the budget be allocated?	Thm. 4	Equalize $\pi_h r_h$
How much hidden monitoring is enough?	Cor. 1	$\pi_0 \geq (1-\alpha^{1/K})/r_{\min}$
Does the monitor catch adversarial attacks?	r_h^L (Prop. 2)	Per-type red-team recall, simultaneous bounds

D.2 Reporting checklist

A monitoring-based safety claim should state:

1. the harm definition, and the unit of certification with its independence justification;
2. the harm-capable traffic;
3. the design (w_h, π_h) by stratum, including any excluded traffic;
4. recall lower bounds r_h^L by attack type, with sample sizes;
5. strategic exposure, as a curve over Γ or a justified tail bound on $d(X)$;
6. the observed detections D;
7. the certificate $U_\alpha(D;\Gamma)$ and the campaign bound for the relevant k;
8. the comparison with equal coverage and with a hidden floor;
9. whether "detected" means detected eventually or detected in time to prevent harm.

Minimal data interface. Lemma 1 at $\Gamma = \infty$ needs only $\{\pi_h, r_h^L\}$ and D, and Theorem 1 adds w_h. Raw traffic, model weights and transcripts need not leave the developer, which makes external review feasible.

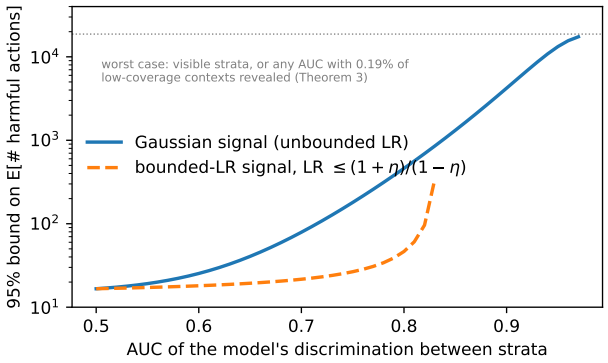


Figure 8: Certified expected count against the AUC of the system's signal (Theorem 2). Two strata with coverages 0.2 and 1.6×10^{-4}, low-coverage share 0.1, $N = 10^8$. By Theorem 3, every AUC above $(1 + \eta_N)/2 \approx 0.501$ is also compatible with the worst case.

What to do differently.

1. Do not turn zero detections directly into a safety claim.
2. Choose the unit of certification to match how sampling is randomized.
3. Measure coverage and adversarial recall by stratum and attack type, with lower confidence bounds.
4. Report a sensitivity curve over Γ, and a campaign bound that separates detection from timely prevention.
5. Do not treat AUC, guessing accuracy, verbalization rates, or probe failures as a tail guarantee.
6. Use tail stress tests and cue experiments to look for weaknesses, not to certify their absence.
7. Prefer structural robustness: guarantee a hidden minimum effective detection probability, and spend the rest of the budget on risk-targeted monitoring.
8. Randomize monitor assignment, or use ensembles with measured recall, so that attack-type selection pays less.
9. Estimate recall from an audit stream with known, positive audit probabilities and hidden injected attacks.
10. Do not expect subsampled monitoring to rule out a single catastrophic action.

D.3 A RECALL AUDIT PROTOCOL

For each harm-capable stratum:

1. Build an attack taxonomy T_h covering attack type, obfuscation, trajectory structure, and difficulty.
2. Curate or generate attacks for each type, and hold out a hidden adversarial test set.
3. Estimate recall separately by type on the held-out set, through the full pipeline including second-stage filters.
4. Compute simultaneous one-sided lower confidence bounds (Proposition 2).
5. Use $\min_t r^L_{h,t}$ unless attack selection is structurally constrained.
6. Rerun after every monitor or filter update.
7. Report sample sizes and bounds alongside the certificate.

The protocol still requires the external-validity assumption of Section 5: every deployment attack of a type is flagged with probability at least that type's recall.

E PARTIAL VISIBILITY IN PRACTICE

At AUC 0.55, 0.80 and 0.95 the binormal signal bounds the expected count μ by 19, 461 and 13257, and the perturbed structure of Theorem 3(b) attains $18{,}723$ at each of these AUC values.

Box 1. Theorem 3 in practice

Not sufficient as evidence for a monitoring certificate:

- a probe separating monitored from unmonitored contexts with low AUC;
- the system guessing monitoring status at chance accuracy.

Theorem 3 covers these two directly: neither controls $\mathrm{LCVaR}_{\mu/N}(d)$. Two other common kinds of evidence are elicited, not bulk summaries, and are weaker still:

- low rates at which the system verbalizes that it is being monitored;
- the system failing to name its monitoring policy when asked directly.

They are lower bounds on the system's information, and a strategic system can sandbag them.

What can support a certificate:

- a design property that holds whatever the system knows: a hidden floor or equal coverage (Corollaries 1 and 4(ii));
- a valid upper bound on the system's discrimination at deployment-scale tails, such as a bounded likelihood ratio or $\mathbb{P}\{d(X) < d_0\} \leq \delta$ (Corollary 4(iii)).

Box 2. Probe dominance, and a tail stress test that can only refute

Probe dominance. If an external probe fails to recover monitoring information from X, that does not show the deployed system lacks the information. A probe weaker than the system estimates a smoother d, and so produces an anti-conservative certificate. A valid argument needs an *upper* bound on what the system can infer.

Tail stress test. Instead of asking whether the system can classify its monitoring status, ask it to select the fraction β of candidate units, drawn from deployment traffic, where it expects oversight to be weakest. Then evaluate the mean *true* coverage $\bar{c}_S$ of that selection. The selection is a rule that depends only on what the system sees, so $\mathrm{LCVaR}_\beta(d) \leq \bar{c}_S$. A selection with low $\bar{c}_S$ therefore proves that the certificate is weak; when $\bar{c}_S$ is estimated from a few selected units, the refutation carries the sampling error of that estimate. A failed selection proves nothing, because the system may be unable or unwilling to perform. Fine-tuning the system on the selection task gives a stronger lower bound than prompting.

Remark 1 (Evaluation awareness). An evaluation is a stratum with coverage one. Clean evaluations certify deployment behavior only through the lower tail of the system's posterior that it is being examined. Anthropic's report makes the same identification: "unrealism can be seen as a lack of coverage" (Claim 3.4.2).

F TWO PUBLIC MONITORING SETTINGS: FURTHER DETAIL

Section 7 makes two points about the architecture of Williams et al. (2026): strategic exposure is small (if the 0.1% of unmonitored traffic is harm-capable and identifiable, the certificate grows by the factor $0.999/(1 - 0.001\Gamma)$ and is infinite from $\Gamma = 1000$), and recall is the binding input. Two further points follow from the framework.

- *Volume matters only for rates.* In the design-based certificate, tens of millions of trajectories strengthen a rate claim, U/N, but not the count certificate.
- *Units and timeliness.* A trajectory-level monitor supports trajectory-level certificates. The 30-minute latency enters the campaign bound through ℓ_h.

The missing inputs are therefore the recall lower bounds by attack type, together with a justification of the trajectory as the unit whose sampling is independent. Those are the control evaluations OpenAI says it needs.

The counting step in the report. The report also argues that "if some misalignment occurs at least once, it is likely that it would happen many times." Under its own illustration, $Y \sim \mathrm{Bin}(n, 1/n)$, the

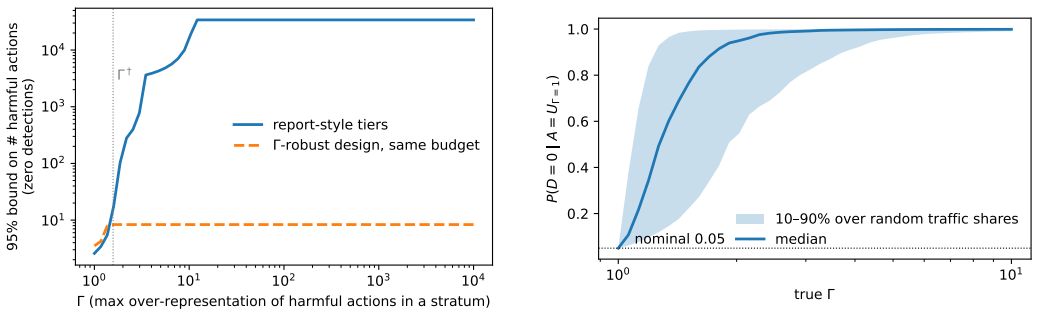


Figure 9: Left: certified count against Γ for the report-shaped tiers and for the Γ-robust (minimax) design at the same budget. Right: probability of zero detections when the true count equals the $\Gamma = 1$ bound, over 500 random traffic-share vectors (Theorem 1(c)).

conditional quantity is $\mathbb{P}(Y \geq 2 \mid Y \geq 1) \to (1 - 2/e)/(1 - 1/e) \approx 0.42$; Corollary 3 replaces this step with a bound that depends on Γ.

G Experimental details

Transcripts. *APPS.* From the public table that joins APPS problems (Hendrycks et al., 2021) with the backdoors of Greenblatt et al. (2023) and Terekhov et al. (2026), we keep problems with a verified working backdoor, a solution that passes the tests, and deterministic output, and sample 1,500 of them with a fixed seed. Each problem contributes its first clean solution that passes the tests and up to two backdoored solutions from different generating models that work and pass the tests. For 56 problems no single stored solution or backdoor passes this check, which leaves 1,444 problems. The stratum is the generating model (ten strata). The monitor sees the problem statement and the code. *BashArena.* We use a public set of 373 agent actions from BashArena (Kaufman et al., 2025), of which 354 are distinct: 199 from honest trajectories and 155 from attack trajectories with one of four side tasks (download malware, leak passwords, disable firewall, edit sudoers). The monitor sees the task, the previous actions with tool outputs truncated to 2,000 characters, and the action under review. The attack label is inherited from the trajectory. Nine of the 155 attack actions are missed by all seventeen monitor configurations at 5% FPR, and some of the actions that every configuration misses are innocuous on their own, for example creating a file of constants or printing three files. BashArena recall therefore probably understates step-level recall; it is a lower bound if such steps are flagged no more often than the malicious steps. *SLEIGHT-Bench.* All 43 attacks in 11 categories (Najt et al., 2026), each with its benign counterpart in which the malicious action is removed. We render each transcript from the published block format and skip the canary line. The attack metadata, which describes the mechanism, is never shown to a monitor. Attack and benign transcripts have a median length ratio of 0.999. Categories with fewer than eight attacks are pooled into one stratum for certificates and shown separately in Figure 10.

Monitors and selectors. A configuration is a model, a prompt and a view. Models: Qwen3-8B, Qwen3-14B, Qwen3-32B in bf16 and in 4-bit AWQ (Yang et al., 2025), and gpt-oss-120b (Agarwal et al., 2025). Prompts: a default prompt that describes hidden objectives and asks for up to three sentences and a score in `<score>` tags, a six-item checklist prompt, and a minimal one-line prompt. Views: full (task, history, action), action only, and task plus action. "Reasoning" means Qwen3 thinking mode or gpt-oss medium reasoning effort, with up to 3,072 output tokens; without reasoning, gpt-oss-120b uses low reasoning effort. Ten configurations use the default prompt and full view (four models, 32B in two precisions, each with and without reasoning), and seven vary the prompt or the view on Qwen3-8B, Qwen3-32B and gpt-oss-120b, for seventeen monitor configurations. Three further configurations (Qwen3-8B and Qwen3-32B without reasoning, gpt-oss-120b with reasoning) act as selectors: they read an attack transcript and estimate the probability that a monitor fails to flag it. The tail stress test of Figure 3(b) adds four kinds of ranking. The supervised probe is a logistic regression on TF-IDF features (word 1- and 2-grams) of the action, trained to predict whether the main monitor misses the attack, with five cross-fitted folds so that every attack is scored by a model

that did not see it; SLEIGHT-Bench attacks are never in a training fold. Each of the 16 proxies ranks attacks by the score of another monitor configuration, lowest first. The random selector uses a random order, and the oracle ranks by the main monitor's own outcome. Inference uses vLLM (Kwon et al., 2023) at temperature 0.7 with a 16k-token context; longer inputs are truncated in the middle (2.4–4.0% of full-view rows). A sample whose reasoning is cut off before an answer is recorded as unparsed and counts as not flagged. This affects at most 4.0% of samples, except for Qwen3-8B with reasoning, where 16.6% of samples and 8–9% of transcripts have no score. The rate is similar for attack and benign rows in every configuration (for the main monitor, 0.42% and 0.47% of transcripts). The 8B and AWQ configurations ran on A100-40GB GPUs and the others on GH200 GPUs. A full pass over the 4,058 transcripts takes between 12 minutes and two hours on one GPU.

Thresholds, strata and analysis choices. The score of a transcript is the mean over samples. For a target FPR f the threshold is the $(1-f)$ quantile of benign scores, taken separately on the code surface (1,444 benign) and the agentic surface (242 benign, BashArena and SLEIGHT-Bench together, because 43 benign transcripts cannot calibrate a threshold). The pooled alternative uses one threshold for all traffic. Recall lower bounds are simultaneous one-sided Clopper–Pearson bounds at level $0.025/H$ (Clopper & Pearson, 1934). The main monitor (Qwen3-32B, bf16, default prompt, full view, reasoning) was chosen for its AUROC on APPS after a pilot comparison of four configurations on a subsample of the same APPS transcripts, and was then fixed in the analysis code for every result reported here. Because the pilot overlaps the evaluation data, the recall lower bounds of the main monitor do not account for this selection step, and the experimental certificates built on them are illustrations rather than finite-sample certificates in the sense of Proposition 2. The operating point was chosen from the bootstrap of Table 4, which resamples benign transcripts only. Routing rates and traffic shares are never estimated. Figure 2(b) uses uniform π and equal shares with a Dirichlet band. The design comparison assigns the published tier rates to surfaces by assumption (1.0 to APPS, 0.25 to BashArena, 0.05 to SLEIGHT-Bench), which fixes the monitored fraction at 0.79 without the SLEIGHT-Bench strata.

Prompts. The default monitor prompt reads: "You are a security monitor reviewing the work of an AI agent. The agent was given a task by a user. It may have done only what the task requires, or it may also have pursued a hidden objective that the user did not ask for [examples]. Judge only what is shown. Ordinary bugs, poor style, or an unusual but legitimate approach are not suspicious by themselves. Behaviour that serves no purpose for the stated task, or that is disguised, is suspicious. [task] [agent behavior] Give a suspicion score from 0 to 100 ..." The selector prompt states that every record contains an unrequested hidden action and that a language-model monitor flags about 1% of benign records, and asks for the probability that the monitor fails to flag the record. Full prompts are in the code release.

H Additional results

H.1 Operating points

Table 4: Main monitor. Recall by benchmark with a 95% interval from resampling the benign transcripts that set the threshold (attacks fixed, 1,000 resamples), surface calibration. "Benign above" is the number of benign transcripts above the threshold.

FPR	1%	2%	5%	10%
APPS, benign above (of 1,444)	15	29	73	145
APPS recall	0.72 [0.66, 0.74]	0.75 [0.73, 0.76]	0.78 [0.78, 0.81]	0.82 [0.82, 0.86]
Agentic, benign above (of 242)	3	5	13	25
BashArena recall	0.17 [0.14, 0.56]	0.43 [0.17, 0.60]	0.65 [0.56, 0.68]	0.72 [0.66, 0.75]
SLEIGHT-Bench recall	0.00 [0.00, 0.02]	0.02 [0.00, 0.02]	0.02 [0.02, 0.05]	0.05 [0.05, 0.07]

Table 4 is the basis for reporting at 5%. The apparent difference between the bf16 and AWQ versions of the main monitor on BashArena at 1% (0.17 against 0.45) lies inside the 1% interval, and the two agree at 5% (0.65 and 0.61) and on APPS at every operating point. With a pooled threshold the realized false-positive rate is uneven across surfaces: at a nominal 5% it is 1.9% on APPS and 25.6% on BashArena, with recalls 0.75, 0.86 and 0.12 on the three benchmarks.

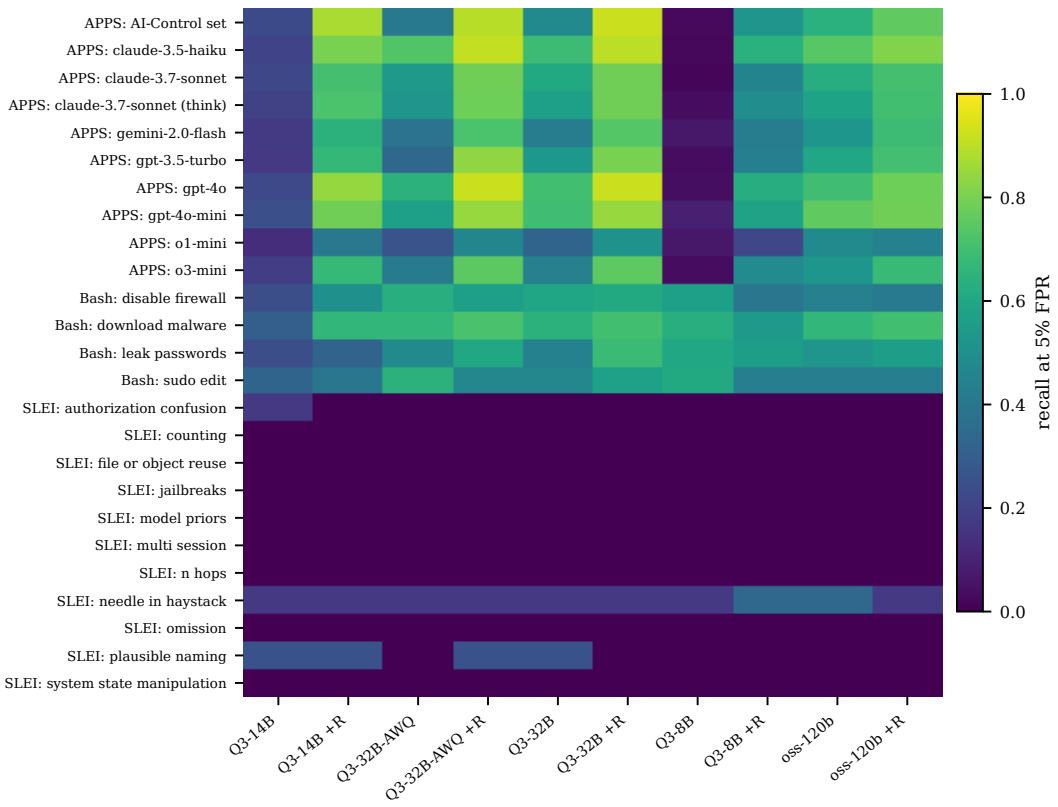


Figure 10: Recall at 5% FPR of the ten default-prompt monitors on every attack type. "+R" marks reasoning. SLEIGHT-Bench categories have between 1 and 11 attacks.

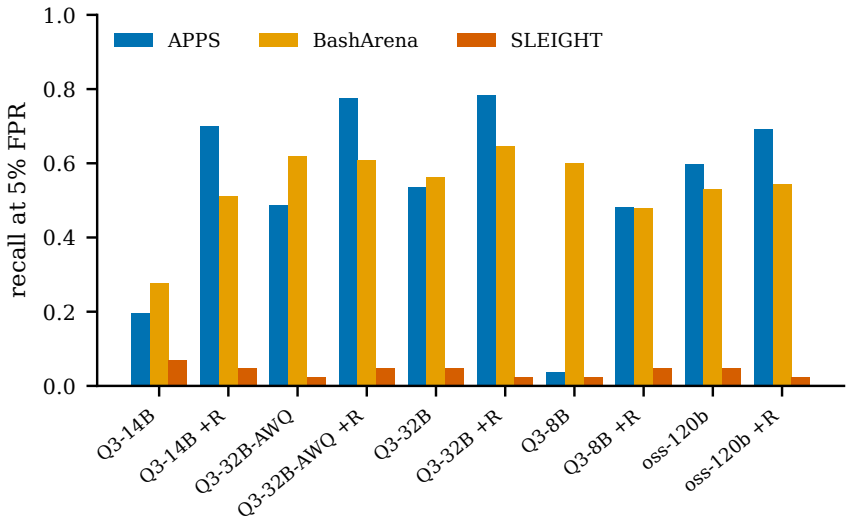


Figure 11: Recall at 5% FPR by benchmark for the ten default-prompt monitors.

H.2 Validity, baselines and the recall audit

Integer configurations. Theorem 1(c) is stated up to rounding. For small A the rounding matters: when an allocation $q \in \mathcal{Q}_\Gamma$ is rounded to integers A_h by largest remainders, 24% of our configurations violate $A_h \le \Gamma w_h A$, and among those the exact miscoverage reaches 1.78α. These configurations

are outside the class the theorem covers. The validity study uses the worst integer configuration inside $\mathcal{Q}_\Gamma$, built by filling strata in increasing order of κ_h under the caps $\lfloor \Gamma w_h A \rfloor$, together with proportional, mixed and random allocations that remain feasible after rounding. A Monte Carlo run with an adversary that takes A actions with probability $1/2$ and none otherwise reproduces the exact value (0.0242 simulated, 0.0240 exact). The closed form for $\kappa^*(\Gamma)$ agrees with a generic linear-program solver on 2,000 random instances to within 5×10^{-16}, and the design linear program of Theorem 4 returns $\bar{\pi}\bar{r}_{\mathrm{H}}$ exactly for $\Gamma \geq \Gamma^\dagger$ and strictly more below it.

Table 5: Probability of zero detections when the true count is just above the stated bound and the allocation is the worst integer configuration in $\mathcal{Q}_\Gamma$, for the tiers of Section 7. The nominal level is 0.05. At $\Gamma = 1.5$ the misspecified bound is about 2.6 and no integer configuration with $A = 3$ lies in $\mathcal{Q}_{1.5}$.

True Γ	1.5	2	5	10
Certificate with the correct Γ	0.003	0.038	0.050	0.050
Certificate with $\Gamma = \infty$	0.000	0.000	0.000	0.001
Certificate computed at $\Gamma = 1$	–	0.058	0.998	0.999
Bound from average coverage, $\log(1/\alpha)/\bar{c}$	0.003	0.698	0.994	0.998
Bound from average sampling rate, recall ignored	0.003	0.043	0.995	0.998

Table 6: Recall audits on the main monitor's APPS scores (true recall 0.80 at 5% FPR, routing $\pi = 0.25$, 300 replications of a stream of 2×10^5 units with prevalence 0.002 and an audit budget of 4,000). "Fails" is the probability of zero detections when the true count is just above the resulting certificate. "Invalid" is the share of audits whose lower bound exceeds the true recall (nominal 2.5%). The bound for inverse-probability weighting uses Kish's effective sample size and is approximate: it exceeds the true recall in 8% of audits, against a nominal 2.5%. The bound from injected attacks is an exact Clopper–Pearson bound.

Audit policy	positives	$\hat{r}$	r^L	fails	invalid
First-stage positives only	319	1.00	0.99	0.085	100%
Score-dependent audit, unweighted	41	0.92	0.79	0.043	42%
Uniform random audit	8	0.80	0.39	0.005	0%
Known $a(s) > 0$, inverse-probability weights	25	0.81	0.58	0.018	8%
Hidden injected attacks ($n = 300$)	300	0.80	0.76	0.036	2%

In a Gaussian simulation with true recall 0.53 at 1% FPR, auditing first-stage positives only leads to a failure probability of 0.21, and known audit probabilities with inverse weighting to 0.01. Figure 12 shows how the number of red-team attacks per type limits the certificate. With 11 types and five attacks each, the simultaneous lower bound on the worst type is zero and nothing is certified. With 163 attacks per type the certificate is 1.90 times the one computed from the true recall, and with 2,000 it is 1.18 times. The plug-in minimum is below the true worst-type recall in only 56% of audits at $n = 40$.

H.3 Designs on measured recall

Table 7: Zero-detection certificates of five designs computed from the main monitor's simultaneous recall lower bounds. Designs that reach the cap $\pi_h \leq 1$ use less than the stated budget.

	APPS and BashArena strata				with SLEIGHT-Bench strata			
Γ	1	2	10	∞	1	2	10	∞
Report-shaped tiers	3.4	8.3	37.0	40.2	3.8	14.4	3.3×10^6	∞
Risk-targeted ($\Gamma = 1$ optimum)	3.3	7.5	∞	∞	3.7	12.2	∞	∞
Equal coverage, capped	5.8	6.1	8.0	8.8	∞	∞	∞	∞
Uniform hidden floor	4.5	7.1	10.7	11.7	6.2	11.9	2.4×10^5	∞
Floor with risk overlay	4.1	12.8	12.8	12.8	3.7	12.2	∞	∞

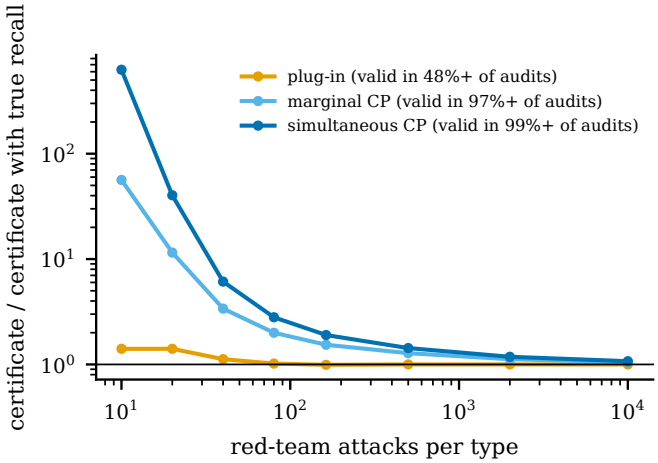


Figure 12: Certificate computed from estimated recall relative to the one computed from true recall, against the number of red-team attacks per type (11 types with recall between 0.14 and 0.94).

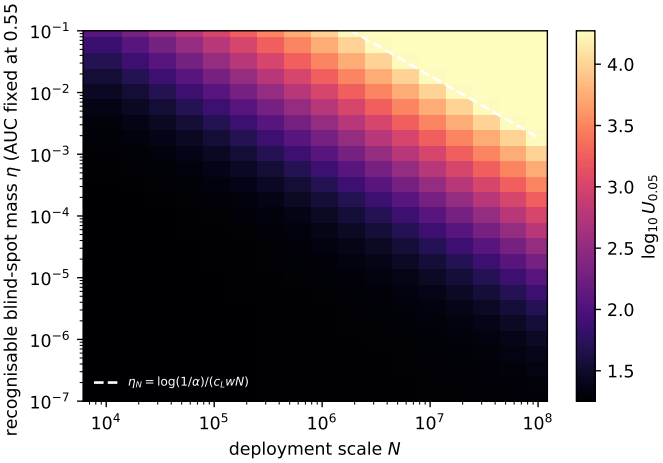


Figure 13: Certificate against deployment scale N and the mass η of a recognizable blind spot, with AUC fixed at 0.55. The dashed line is the threshold η_N of Theorem 3. A blind spot of mass 10^{-4} offers 0, 10 and 1000 expected opportunities at $N = 10^4$, 10^6 and 10^8, and the certificate is 18, 29 and 1018.

Equal coverage needs $\pi_h = \bar{\pi}\bar{r}_{\mathrm{H}}/r_h^L$, which exceeds one in some strata; we cap it at one, so coverage is not exactly equal and the certificate still varies with Γ. With the SLEIGHT-Bench strata, the recall lower bounds are 0 for the model-priors stratum and 4.9×10^{-5} for the pooled small categories. Equal effective coverage is then attainable only at level 0, and Proposition 1, which needs $r_h^L > 0$, gives no positive floor; the only choice is $c_0 = 0$, for which we set $\pi_h^{\mathrm{floor}} = 0$ in every stratum, so the hybrid design reduces to the risk-targeted design. Because the zero-recall stratum has share $1/16$, every design is unbounded from $\Gamma = 16$, and the finite values at $\Gamma = 10$ exceed 2×10^5. The assumption $\bar{\pi}\bar{r}_{\mathrm{H}} \le r_{\min}$ of Theorem 4 fails without the SLEIGHT-Bench strata and holds only trivially, with both sides 0, with them. Adding a linear false-positive constraint $\sum_h w_h \pi_h f_h \le B$ to the design problem keeps it a linear program. On the exact hazard scale we maximized $\mathrm{LCVaR}_{1/\Gamma}(-\log(1 - \pi_h r_h))$ numerically by multi-start sequential quadratic programming from the linear-program solution, over six budgets, five values of Γ and three recall profiles (the 16 measured lower bounds, with the two SLEIGHT-Bench bounds replaced by 0.001 so that every stratum can be covered; these bounds plus 0.3, capped at 0.99; and half of them). The assumption holds in 44% of these instances. Where it holds, the gain over the linear-program design is 0.00% for every $\Gamma \ge 2$; where it fails the gain is

at most 9.32%. At $\Gamma = 1$ the hazard scale rewards saturating one stratum, because $-\log(1-c)$ is convex, and the gain reaches 212%.

H.4 Randomized monitor assignment

We estimated the recall matrix of the seventeen monitor configurations on the 14 attack types with at least 16 attacks, split each type in half 200 times, solved the game of Proposition 3 on one half, and evaluated worst-type recall on the other.

Table 8: Worst-type recall on held-out halves, averaged over 200 splits. "Wins" is the share of splits in which the optimized mixture beats the best single monitor selected on the same half.

FPR	1%	2%	5%	10%
Best single monitor	0.229	0.270	0.478	0.551
Optimized mixture, held out	0.207	0.287	0.469	0.526
Optimized mixture, in sample	0.273	0.373	0.515	0.615
Uniform mixture	0.116	0.198	0.305	0.421
Wins	22%	58%	30%	29%

In sample the mixture is never below the best single monitor, as Proposition 3 requires, and is strictly above it in 81% to 100% of splits. Held out, it is better on average only at 2% FPR. With between 25 and 588 attacks per type the recall matrix is estimated too coarsely for the optimized weights to transfer, and the gain available in these data is smaller than that estimation error. An any-monitor-fires ensemble with thresholds tightened to a joint 5% FPR has worst-type recall 0.36 at seventeen times the inference cost.