How Many More Labels Are Needed? Minimax Evaluation under Selective Labels

Anonymous authors
Paper under double-blind review

Abstract

Good observed accuracy need not establish acceptable prediction risk when difficult cases lack labels. How much can an additional audit establish? For a frozen cohort with outcome-dependent missingness, we characterize the exact fixed-query asymptotic minimax power over adaptive audits using the full observed record. A standard residual rule attains the truthful frontier with finite-sample false-alarm control. A second result identifies which risk guarantees preserve information in existing observations: count-tail and cumulative predictable-risk guarantees have the same truthful frontier but different noisy critical experiments. The latter reduce auditing to query-only testing, even with independent heterogeneous errors; the former can retain native information. Above criticality, we derive the exact robust frontier with uncertain conditional label reliability, including its blind region, and invert it to quantify the query cost of label-quality uncertainty. This separates three resources: query count, knowledge of label quality, and cohort size. Under count-tail risk bounds and truthful labels, finite certificates show that at 10% versus 20% risk and 5% false alarms, 57 queries suffice and are necessary for 95% power among 1,000 predictions with 500 pending labels. At 12% risk, even complete labels require at least 2,645 predictions for that power. These are scoped design examples, not a general finite-budget formula or validation of deployment assumptions.

1 Introduction

A model can look accurate because its easiest cases receive labels first. How many unresolved cases must an evaluator check to detect excessive risk while controlling false alarms? Acquiring labels is a familiar remedy; knowing what that budget establishes is a different problem. We consider a frozen model and cohort whose missing outcomes already exist and can be adjudicated. This is narrower than general selective labels (Lakkaraju et al., 2017): counterfactual or future outcomes cannot necessarily be obtained by requesting a label.

The obstacle is not only that errors may be hidden. An acceptable-risk model can also have hidden errors, so finding an error does not by itself establish a violation. Both explanations can be compatible with the same observed record: 500 observed correct predictions could accompany 100 or 200 errors among 500 pending cases. The observed record alone does not distinguish these explanations; a private random audit samples their different hidden compositions. Our benchmark asks how well even the best adaptive audit can distinguish them.

A reliability claim and the ability to detect its violation are different. Lifecycle reliability connects pre-deployment calibration with run-time assessment (Simeone et al., 2025); the CONTRACT research programme pursues this agenda for wireless AI (European Commission, 2025). Pre-deployment vulnerability diagnosis motivates scrutiny; it does not establish how much selectively observed feedback can prove. Probability sampling, active testing and residual audits supply established interventions (Horvitz and Thompson, 1952; Kossen et al., 2021; Kato and Nakagawa, 2026). Our contribution is their worst-case detection frontier, not the introduction of monitoring or a new calibration method.

An outcome-aware mechanism withholds exactly b of n labels and commits the truthful remainder before an auditor privately queries at most k pending items. We match achievable and impossible power levels as the cohort grows for every fixed k. Uniform nonadaptive queries attain the truthful

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.

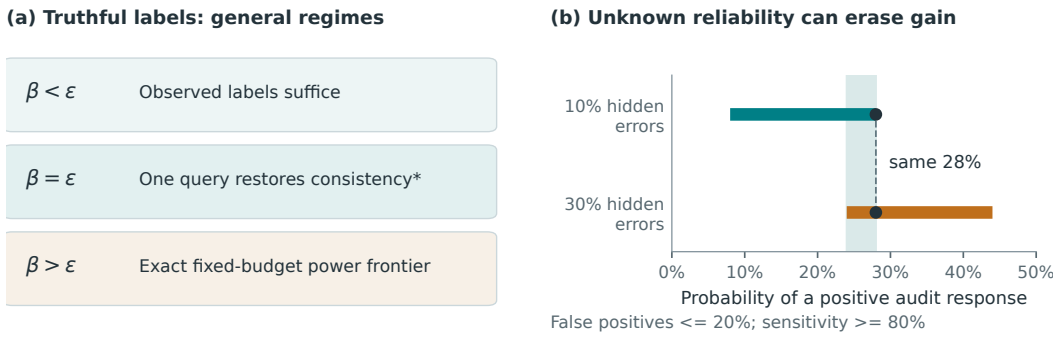


Figure 1: More labels and better information about labels are different resources. (a) The truthful frontier depends on the missing fraction β relative to the risk gap $\epsilon = p - a$; the critical passive comparison additionally requires a root-n offset (Corollary 1). (b) At $a = .1, p = .2, \beta = .5$, a common observed error count $.05n$ leaves hidden error fractions .1 and .3. With false positives at most .2 and sensitivity at least .8, the allowable response probabilities overlap. Bars are response ranges, not confidence intervals. Two legal channels produce the same .28 positive rate. Corollary 4 turns this mechanism into an all-policy fixed-k statement.

frontier without knowing alternative risk, with finite-sample false-alarm control. Above criticality, the upper bound matches *full* native observations under legal stochastic populations, ruling out gains from positions or adaptive identities that a count-only argument would miss.

A second finding explains when existing labels can help an audit. Count-tail and cumulative predictable-risk guarantees yield the same truthful frontier, but with imperfect audit labels their critical experiments differ. The latter can eliminate all native contribution to minimax power even with independent heterogeneous errors; the former can retain it. Mean risk alone can prevent consistency even with full labels. Thus a design benchmark must specify not just a query budget, but the risk guarantee and label reliability under which existing and newly acquired observations are combined. Indeed, with sufficiently many missing labels, uncertainty about otherwise informative audit responses can remove all fixed-budget asymptotic gain. Figure 1 distinguishes this identification obstacle from a shortage of truthful queries.

Contributions. (i) An exact fixed-query frontier over full-transcript adaptive policies, a p-agnostic truthful attainer, and query-budget consequences (Sections 3–4). (ii) Noisy experiments and risk-contract comparisons explaining when existing observations help, exact noisy budget frontiers, and when uncertainty removes the audit's asymptotic gain (Sections 5–6). (iii) Certified finite design decisions separating query requirements, procedural losses, and minimum complete-information cohort size (Sections 3, 7). Noisy optima and finite design-point rules may know a planning alternative; their scope differs from the truthful attainer's.

2 Experiment and Guarantees

Condition on a frozen predictor, calibration objects, and preassigned cohort indices, not on all realized case features or labels. The errors retain the conditional law specified below. Let $V_i \in \{0, 1\}$ indicate an error, with $K = \sum_i V_i$. The null class $\mathcal{H}_0(a)$ comprises laws satisfying

$$\mathbb{E}[V_i \mid V_1, \ldots, V_{i-1}] \leq a \quad \text{for each } i. \tag{1}$$

It contains iid Bernoulli(a) but also dependent losses. It is not implied by the marginal guarantee of a conformal predictor. A rejection concerns (1), not a contradiction of the marginal conformal theorem. The baseline alternative is iid Bernoulli(p), $p > a$. Section 6 extends both latent-law classes explicitly.

All n errors are realized before a single prespecified deadline. There are no earlier outcome messages. A possibly randomized kernel κ inspects the entire latent vector, chooses exactly b pending positions, and publishes $S \in \{\perp, 0, 1\}^n$. Every nonpending entry is truthful. Let $\mathcal{K}_b$ be this unrestricted

Table 1: What must be supplied before using a result. Risk classes are defined in Section 6; $\mathcal{C}$ and $\mathcal{P}$ are not nested. None is certified by the observed cohort itself.

Supplied justification	Available guarantee	Not implied
Count-tail risk bounds $\mathcal{C}$; truthful audits	General frontier and certified finite query/cohort designs	Same finite budget under heterogeneous average risks
Pathwise predictable-risk budgets $\mathcal{P}$	Truthful frontier; critical query-only noisy limit	The 57-query finite certificate
Mean risk alone $\mathcal{M}$	Sharp, possibly low power ceiling	Consistency from complete labels
Conditional reliability envelope plus $\mathcal{C}$ or $\mathcal{P}$	Robust channel guarantees and identification criterion	Known channel or guaranteed transport from average accuracy

cohort-end kernel class and $J = \sum_i \mathbf{1}\{S_i = 1\}$. Pending labels may be released at the next tick, after our decision; this does not make them available at the deadline.

After the kernel irrevocably commits S, an audit policy π chooses at most $k \le b$ distinct pending positions. Each requested true label is obtained. The policy may use S, previous query identities and results, and an independent random seed. There are no new native releases or extra outcome proxies during this phase. The kernel cannot anticipate the seed or modify the pending set after seeing a query. These are protocol assumptions, not missing-at-random assumptions about natural feedback. The full transcript T includes S, query identities, their order, their results, and retained provenance.

For fixed a, p, n, b, k, define the maximin power

$$V_{n,k}(\delta) = \sup_{\pi,\, 0\le\phi\le 1} \inf_{\kappa\in\mathcal{K}_b} \mathbb{E}_{p,\kappa,\pi}\phi(T), \quad \text{subject to} \quad \sup_{Q\in\mathcal{H}_0(a),\, \kappa\in\mathcal{K}_b} \mathbb{E}_{Q,\kappa,\pi}\phi(T) \le \delta. \tag{2}$$

For example, freeze a classifier and independently recruit cases whose true outcomes already exist in an accessible reference archive. Prespecify a deadline and a pathwise pending-count bound; commit the partial record, then privately sample pending cases for gold adjudication. Under identical recruitment, errors are iid at the frozen classifier's risk. Guarantees average over recruitment and auditing, not conditional on the realized feature list. A supplied bound $B \le b$ can replace exactly b by padding (Appendix F.4); waiting until a favorable B is observed cannot. This is an illustrative protocol, not deployment evidence. Commitment and private sampling are enforceable design requirements, not assumptions of random natural feedback. Recruitment laws and transported label quality require external justification; pre-deployment calibration alone does not supply them (Table 1).

A finite-population count certificate concerns the realized K. Here size and power also average over its generating law: an acceptable-risk law can realize an unusually error-rich cohort. Even full labels with $K/n > a$ need not justify rejection at level δ; the target is a risk-generating law. The count cap pays for this distinction before residual sampling. The absence of features is not an advantage over feature-assisted testing.

Test randomization is allowed. The null and alternative kernels need not coincide. The supremum can know p; our attaining rule does not need it. This is a single-deadline guarantee, not repeated spending of δ or ARL.

3 Exact intervention frontier

Let $R_k(u, v; \delta)$ denote the randomized Neyman–Pearson (NP) power at size δ for $\mathrm{Bin}(k, u)$ versus $\mathrm{Bin}(k, v)$.

Theorem 1 (Fixed trusted-query budget)**.** *Fix $0 < a < p < 1$, $0 < \delta < 1$, and an integer $k \ge 1$. Suppose $b_n/n \to \beta$ with $0 < \beta < 1$, and write $\epsilon = p - a$. If $\beta > \epsilon$, define*

$$d = \epsilon/\beta, \qquad I_\beta = \left[\max\left(0, \frac{a+\beta-1}{\beta}\right), \min\left(\frac{a}{\beta}, 1-d\right)\right]. \tag{3}$$

Then the optimum over all adaptive policies in (2) *satisfies*

$$\lim_{n\to\infty} V_{n,k}(\delta) = \min_{u\in I_\beta} R_k(u, u+d; \delta). \tag{4}$$

If $0 < \beta \le \epsilon$, the limit is one. A single sequence of tests using uniform nonadaptive queries attains these limits pointwise in $p > a$ without knowing p, and has finite-n size at most δ under (1)*.*

What determines the nuisance interval? If observed errors per cohort member equal $y = J/n$ in both worlds, feasibility requires $\max(0, p - \beta) \le y \le \min(a, 1 - \beta)$. Hidden miss proportions are then $u = (a - y)/\beta$ under the null and $u + d = (p - y)/\beta$ under the alternative. The interval in (3) is exactly this feasible set, not a freely chosen worst-case parameter range. Large amounts of native data do not remove this ambiguity because the adversary can make their laws asymptotically identical in total variation.

Why the full-transcript bound is necessary. A count-only impossibility claim could discard useful native hit patterns or audit identities. Our upper construction chooses uniform within-type native releases. Conditional on its common native miss count, the native ledger is uniform with the same law in both worlds. The pending labels are exchangeable, so any adaptive distinct-index policy sees the same without-replacement label law, irrespective of which pending index it chooses. Its entire transcript is a common postprocessing of that sequence. In the fixed-k limit only the binomial experiment remains. No symmetry of the larger ordered null class is assumed. Appendix A supplies the complete proof.

Attaining procedure. Choose H_n so that every null has $\mathbb{P}(K > H_n) \le \eta_n$, where $\eta_n \to 0$, $\eta_n < \delta$, and $H_n/n \to a$. A conditional Hoeffding bound or a binomial upper quantile suffices. Reject directly if $J > H_n$. Otherwise, given $J = j$, use a randomized upper-tail test of the audit miss count with reference law

$$\mathrm{HG}\big(b_n, \min(b_n, H_n - j), k\big), \qquad \text{size } s_n = (\delta - \eta_n)/(1 - \eta_n). \tag{5}$$

On every whole latent population with $K \le H_n$, the true pending miss count is no larger than this reference. Conditional on that population and the committed mask, uniform auditing is honest finite-population sampling. This establishes size $\eta_n + (1 - \eta_n)s_n \le \delta$ without assuming that the pending labels are iid conditional on the native observations.

Established baseline, distinct target. Ordinary residual sampling supplies the attaining rule, not a new sampling primitive. In several one-query regimes a familiar Horvitz–Thompson (HT) test already attains the limit (Appendix S.6). The theorem instead establishes the exact full-record stochastic-risk benchmark for every fixed query budget.

3.1 A CONCRETE QUERY-BUDGET CONSEQUENCE

For $a = .1, p = .2, \beta = .5, \delta = .05$, the minimum fixed budget for 95% *limiting* power is **56**. An exact rational interval certificate bounds the continuous nuisance minimum, rather than assuming its minimizer lies on a grid. This is a planning statement about the alternative risk, not a finite-cohort 95% guarantee (Appendices D, E).

A finite decision, not just an asymptotic rate. For $n = 1000, b = 500$ at the same risks and level, the minimum hard budget for 95% worst-case power is **exactly 57**, also under the count-tail classes $\mathcal{C}_0(.1), \mathcal{C}_1(.2)$ in Section 6. A legal iid two-world construction bounds every full-transcript adaptive 56-query policy below .94901. A supplied monotone residual rule with 57 queries guarantees at least .95243, with exact size at most .05. Ignoring extra queries excludes every smaller budget. The upper bound retains all rare totals. The lower rule calibrates against the complete worst-mask null expectation, replacing a conservative separate cap-tail allowance, and selects its cap for the planning alternative. Appendix E gives proofs and matched calibration comparisons. This does not identify the exact finite optimal power or a general finite-budget formula.

4 WHEN ONE QUERY CHANGES CONSISTENCY

The exact passive benchmark is the ROC hull of native-count thresholds (Proposition 8, stated in Appendix S). Comparing it with the intervention frontier isolates a qualified consistency transition.

Corollary 1 (Qualified zero-versus-one-query transition)**.** *Under Theorem 1's fixed parameters, suppose* $(b_n - n\epsilon)/\sqrt{n} \to s \in \mathbb{R}$. *Set* $c = \sqrt{a(1-a)/[p(1-p)]}$ *and* $g_s(x) = \Phi(c\Phi^{-1}(x) - s/\sqrt{p(1-p)})$, *with endpoints zero and one. Then* $V_{n,0}(\delta)$ *converges to the least concave majorant of* g_s *at* δ, *strictly below one, whereas* $V_{n,k}(\delta) \to 1$ *for every fixed* $k \ge 1$. *In particular,* $b_n = \lfloor n\epsilon \rfloor$ *has* $s = 0$.

For $a = .1, p = .2, \delta = .05$ and this rounded critical budget, the zero-query limit is approximately .1086685, versus one with a query. The former is a normal-probability calculation from (43), not a simulation. The cohort and its native labels grow without bound: we are not asserting that one label in isolation can identify a risk parameter.

The mechanism is concrete. Under the alternative, either the native miss count already exceeds the null's total-count cap, or almost the entire missing pool must consist of misses to hide the excess risk. Under any null population within that cap, the same native count leaves only a vanishing miss proportion in the pending pool. A fresh pending query distinguishes these two explanations.

This is not broad high power at arbitrary missing fractions: one-query power falls to .10 at $\beta = .20$ (Appendix Figure 6). Nor can the second-order qualification be dropped: $b_n = \lfloor n\epsilon \pm n^{.9} \rfloor$ have the same limiting fraction but passive limits δ and one, respectively. Appendix B proves the distinction.

5 What changes when audit labels are imperfect?

Above criticality, the common-native-ledger construction can erase native information asymptotically. In the $\sqrt{n}$ critical window, native information can survive. Imperfect queries can make it consequential, although some one-sided channels still yield consistency. Keep the same experiment, but replace truthful audit responses by a known binary channel with $q_0 = \mathbb{P}(W = 1 \mid V = 0)$ and $q_1 = \mathbb{P}(W = 1 \mid V = 1)$, $0 \le q_0 \le q_1 \le 1$. Each query uses fresh channel randomness conditional on the latent vector, past transcript, and current identity. Native labels remain truthful. Conditioning on the *complete* adaptive identity path need not preserve independence of past responses. Denote the new minimax power by $V^q_{n,k}(\delta)$ and the ROC for $\mathrm{Bin}(k, q_0)$ versus $\mathrm{Bin}(k, q_1)$ by R^q_k.

Theorem 2 (Joint critical information)**.** *Fix $0 < a < p < 1$, $0 < \delta < 1$, and $k \ge 1$. Suppose $(b_n - n(p-a))/\sqrt{n} \to s \in \mathbb{R}$. Let $X_0 \sim N(0, a(1-a))$ and $X_1 \sim N(-s, p(1-p))$. Then*

$$\lim_n V^q_{n,k}(\delta) = \sup_{\substack{0 \le h \le 1,\ h \text{ nondecreasing} \\ \mathbb{E}h(X_0) \le \delta}} \mathbb{E}R^q_k(h(X_1)). \tag{6}$$

The upper bound permits every adaptive distinct-query policy using the full ledger. Uniform queries and finite-valid, native-count-dependent tests attain the limit pointwise; their construction may depend on p, s, q_0, q_1.

Here h allocates test size to the audit after observing the native error count. Its monotonicity is forced by masks that can lower the native count under the null. It is not a demand for size δ separately at every observed count. The ordinary ROC of a raw Gaussian observation plus noisy queries can use nonmonotone evidence that the adversary removes. Conversely, using only the query ROC can discard native information that survives. Appendix P proves the full-record reduction and exact conditional-null calibration before taking limits.

Uninformative queries recover the passive boundary; perfect queries recover consistency. A one-sided channel with $q_1 = 1$ is consistent iff $q_0^k \le \delta$, whereas $q_1 < 1$ and $q_0 < q_1$ preclude fixed-k consistency (Appendix P). These results concern an increasing native cohort, not a lone label in isolation. The conditional-error extension below removes known homogeneity, not the need for justified reliability.

The functional is ordinary product testing after a concave-envelope repair of the native null law (Appendix P.4); testing-curve products are established (Dong et al., 2019). Our result identifies the native experiment that survives legal selection and its finite-valid transfer, not a new ROC-composition operation.

6 Which assumptions carry information?

All extensions retain committed selection, truthful native labels, a hard query budget, and one deadline. They alter the latent-risk or audit-response classes, with the infimum in (2) now also over alternative laws.

Independence is not necessary. Write $\preceq_{\mathrm{st}}$ for stochastic order and define $\mathcal{C}_0(a) = \{Q : K \preceq_{\mathrm{st}} \mathrm{Bin}(n, a)\}$ and $\mathcal{C}_1(p) = \{Q : K \succeq_{\mathrm{st}} \mathrm{Bin}(n, p)\}$. Pointwise conditional risks at most a or at least p suffice, but these classes also contain nonexchangeable dependent vectors.

Corollary 2 (Count-envelope transfer)**.** *Under $\mathcal{C}_0(a), \mathcal{C}_1(p)$, Theorems 1 and 2 retain their limiting frontiers; Proposition 8 retains its exact finite passive value. Monotone finite lower certificates remain valid.*

For a table increasing in native and queried error counts, the maximum and minimum over masks are both increasing in the total K. Tail ordering therefore transfers its size and power bounds exactly; the iid hard subproblem still bounds all policies. This is not sufficiency of counts for each specified dependent alternative.

A reliability bound can replace a known audit channel. Let $\mathcal{G}_t$ contain the entire latent vector, committed ledger, past transcript, and current queried identity, but no unexposed private audit seed. Allow any causal response process satisfying

$$\mathbb{P}(W_t = 1 \mid \mathcal{G}_t) \le u \text{ if } V_{I_t} = 0, \qquad \mathbb{P}(W_t = 1 \mid \mathcal{G}_t) \ge v \text{ if } V_{I_t} = 1. \tag{7}$$

Robust size and power range over all such causal response kernels; their extremizers may differ between hypotheses. Write $V^{\mathcal{C},[u,v]}$ or $V^{\mathcal{P},[u,v]}$ for the value under the indicated latent classes. Errors may vary with identity and history. The bounds are supplied, not ordinary marginal-accuracy estimates.

Corollary 3 (Robust reliability)**.** *For $0 \le u \le v \le 1$ under the count-envelope classes, the robust critical limit equals* (6) *with $(q_0, q_1) = (u, v)$. At finite n, monotone rules' exact worst-case size and power use channels $(u, 1)$ and $(0, v)$, respectively. For any $0 \le v \le u \le 1$, the robust value equals the passive value at every n, k.*

The overlap statement follows because a truth-independent channel is legal. A sensitivity bound alone ($u = 1$) can thus yield no guaranteed detection gain, even if every pending item is queried. This contrasts with noisy missed-mass *upper certification*, for which false positives can be conservative (Anthony and Salehzadeh Nobari, 2026, Section 9). Nor can marginal noise rates replace (7): a common Bernoulli(u) response across k true zeros makes the all-positive event have probability u, not u^k.

Average predictable risk preserves intervention, not native information. Set $\mu_i = \mathbb{E}[V_i \mid V_{<i}]$, and define the alternative contract

$$\mathcal{P}_0(a) : \sum_i \mu_i \le na \text{ a.s.}, \qquad \mathcal{P}_1(p) : \sum_i \mu_i \ge np \text{ a.s.} \tag{8}$$

This allows individual risks to exceed the null threshold. The $\mathcal{C}$ and $\mathcal{P}$ classes are not nested; both broaden the pointwise conditional benchmark in different ways.

Theorem 3 (Critical query-only experiment)**.** *Under* (8)*, Theorem 1's truthful first-order frontier still holds, using its Hoeffding-cap rule. Under* (7)*, for fixed $0 \le u \le v \le 1$, $k \ge 1$, and $b_n = n(p - a) + o(\sqrt{n})$, the robust value satisfies*

$$\lim_n V_{n,k}^{\mathcal{P},[u,v]}(\delta) = R_k(u, v; \delta). \tag{9}$$

A finitely valid uniform-query rule attains this without knowing p or v. The passive limit is δ.

The upper bound uses independent heterogeneous errors and a fixed mask: pending labels are all zero or all one, while the *full* native laws approach each other in total variation. Temporal dependence and adaptive selection are not needed for this loss. The lower rule gates extreme native counts and calibrates a query test using martingale concentration. Average conditional risk is already an established monitoring estimand (Howard et al., 2021; Podkopaev and Ramdas, 2022); the conclusion here is the exact selected-feedback experiment it leaves, not a new concentration bound.

An exact budget frontier under reliability uncertainty. For $\beta > p - a$, put $d = (p - a)/\beta$, and let I_β be (3). Define $R_k^+(r, s; \delta) = \delta$ when $s \le r$, and the binomial NP power $R_k(r, s; \delta)$ otherwise.

Table 2: Numerical minimum budgets for 95% limiting power at $a = .1, p = .2, \beta = .5, \delta = .05$. A known common channel $(e, 1-e)$ and conditional bounds $(u, v) = (e, 1-e)$ supply different information. The final entry is an exact blind-region conclusion; finite entries are numerical nuisance minimizations, not rational finite certificates.

e	0	.05	.10	.15	.20
Known channel	56	72	94	127	176
Conditional bounds	56	117	367	5,981	∞

Theorem 4 (Robust noisy frontier)**.** *Suppose $b_n/n \to \beta \in (p-a, 1)$. For fixed $k \geq 1$ and $0 \leq u < v \leq 1$, under either $\mathcal{C}$ or $\mathcal{P}$,*

$$\lim_n V_{n,k}^{[u,v]}(\delta) = \min_{x \in I_\beta} R_k^+\big(u + (1-u)x,\, v(x+d); \delta\big). \tag{10}$$

This also holds for the original risk classes and full adaptive transcripts.

The null's largest response mean is $u + (1-u)x$; the alternative's smallest is $v(x+d)$. Overlap erases information. Outside overlap, a finitely valid uniform-query test attains the displayed limit; a level-δ coin attains it in the blind region. Appendix H gives the matching full-record proof and a conservative finite power certificate. For a known common channel, replace both means by the same affine map $t \mapsto q_0 + (q_1 - q_0)t$.

Corollary 4 (Reliability blind region)**.** *Under the preceding theorem, for $(u, v) = (e, 1-e)$, $0 < e < 1/2$, every fixed k has limiting power δ exactly when $\beta \geq (p-a)(1-e)/e$.*

Otherwise every target limiting power below one is achievable with a finite fixed budget. Inverting (10) gives Table 2: knowing a 10% symmetric channel requires about 94 queries, while only knowing the same conditional error bounds requires about 367. This quantifies the value of reliability information beyond merely detecting a blind region. These are fixed-k asymptotic inversions, not finite-cohort certificates or a growing-budget theorem. The cost and transport of reliability validation are additional resources.

Mean risk alone cannot support the same conclusion. Let $\mathcal{M}_0(a) = \{Q : \mathbb{E}K/n \leq a\}$ and $\mathcal{M}_1(p) = \{Q : \mathbb{E}K/n \geq p\}$.

Proposition 1 (Sharp mean-risk ceiling)**.** *For $0 < \delta \leq a < p < 1$, $1 \leq k \leq b \leq n$, with truthful queries, the exact finite minimax value under $\mathcal{M}_0(a), \mathcal{M}_1(p)$ is $\delta p/a$, even if $k = b$. One uniform query attains it.*

The standard mixture null $(a/p)\,\mathrm{Bern}(p)^{\otimes n} + (1-a/p)\delta_{0^n}$ forces the upper bound even with full labels; the HT test $(\delta/a)\widehat{r}_{\mathrm{HT}}$ attains it. At the same $.1/.2/.05$ parameters the ceiling is .10 for every cohort size: recovering labels cannot replace restrictions on the risk-generating law. Appendix Q supplies proofs, non-nesting counterexamples, and a failure example when only the alternative is weakened.

The same truthful limit can hide different information. Table 3 isolates the distinction: $\mathcal{C}$ and $\mathcal{P}$ give the same truthful one-query limit but different noisy limits. Under $\mathcal{P}$, independent heterogeneous risks can shift the excess entirely into the pending set and make the full native laws match. In this comparison, count-tail restrictions under $\mathcal{C}$ preserve critical fluctuations that a noisy query can exploit. The difference is in the permissible population laws, not a better query policy or a universal ordering of the two nonnested classes.

Appendix Figure 5 extends this comparison across budgets and symmetric bounds $(u, v) = (e, 1-e)$. Native information need not give a large gain: at $e = .1, k = 4$, the exhibited improvement over .994 query-only power is only .000186. This is not an upper bound on its value.

Independent heterogeneous errors with average probability at most a justify $\mathcal{P}_0$, not necessarily $\mathcal{C}_0$: rates $(.02, .18)$ have average .1 but $\mathbb{P}(K \geq 1) = .1964 > .19$, the $\mathrm{Bin}(2, .1)$ tail. Thus averaging risks cannot justify binomial-tail calibration.

Table 3: One protocol, different risk guarantees. Limiting minimax powers at $a = .1, p = .2, \delta = .05, b = \lfloor .1n \rfloor$. Noisy queries obey conditional $(u, v) = (.1, .9)$. The count-tail lower bound $B_{\mathcal{C}} \approx .486355$ comes from an explicit rule, not an optimized or rationally certified decimal (Appendix Q). The classes are not nested.

Risk guarantee	No query	One truthful query	One noisy query
Count-tail bound $\mathcal{C}$	$\approx .10867$	1	$\geq B_{\mathcal{C}} \approx .48636$
Predictable-risk budget $\mathcal{P}$	.05	1	.45

These classes are hypotheses, not post-data tuning options. Per-step conditional risk bounds imply the corresponding $\mathcal{C}$ classes; heterogeneous predictable budgets motivate $\mathcal{P}$, and an expectation alone supplies $\mathcal{M}$. No single cohort certifies its own class.

Supplying a reliability bound costs information and resources. Gold controls can supply bounds only under a response model transported to pending items; they estimate parameters, not transport. A bound-failure budget η_0 reduces the stage level to $(\delta - \eta_0)/(1 - \eta_0)$; sensitivity uncertainty affects power (Appendix R). Controls cost gold labels that might instead resolve pending items. Joint calibration/query allocation remains open; average accuracy alone does not supply the conditional envelope.

7 Verification and finite-sample strength

All numerical evidence is deterministic; rational certificates and numerical optimizations are identified separately. The asymptotic attainer need not be efficient at a finite cohort size. At $n = 100, b = 10, k = 1$, its binomial-cap implementation has power .02005, while rational primal/dual witnesses place the finite optimum near .38130. At $n = 1000, b = 100$, approximately .66051 is achievable, but no full-record one-query policy can guarantee .67. Appendix I retains the complete comparison, including known and bounded noise. These certificates apply to $\mathcal{C}$. A separate finite $\mathcal{P}$ construction certifies power above .9553 with 1200 bounded-noise queries from a cohort of 10^6; it is conservative, not a minimum budget (Appendix H).

Planned and actual risk can differ. The $b = 500, k = 57$ truthful rule designed at $p = .2$ has only .01155 worst-case power at .12. A private mixture improves this to .08661 while retaining .95017 at .2, but remains weak (.01695 at .11). These negative results are retained: a design-point guarantee is not uniform power (Appendices F and J).

A larger cohort may be necessary. At $n = 1000, a = .1, p = .12, \delta = .05$, even complete truthful labels give optimal power only .6531752. Exact inversion places the smallest complete-label cohort attaining 95% power at **2,645**. This is a necessary cohort size for any partial audit, not a sufficient partial-label budget (Appendix F.3).

Before inspecting labels, fix the cohort, a, δ, a planning alternative, desired power, and independently justified risk and reliability assumptions. Use finite certificates when available and inspect off-design performance. The 57-query certificate uses truthful $\mathcal{C}$ classes; a pathwise pending bound cannot be replaced by conditioning on an observed random count. Nonrejection certifies neither acceptable risk nor the realized cohort total.

8 Intellectual lineage and scope of the contribution

Selective labels and active evaluation. Selective-label problems can involve unrecoverable outcomes (Lakkaraju et al., 2017); ours permits post-selection measurement, not counterfactual identification. Recovery sampling already supports testing MAR versus MNAR (Noonan et al., 2022), a different target from excess prediction risk. Frozen models and adaptive evaluation are established (Kossen et al., 2021), as are proxy-assisted and robust sampling (Ashouritaklimi et al., 2026; Zrnic and Candès, 2024; Li et al., 2025). We characterize worst-case detection without promised feature

advantage. Kuboyama (2026) certifies an AUGRC model-choice winner using finite-pool label budgets; our target is population-risk testing after adversarial outcome-dependent withholding.

Residual-population auditing. Kato and Nakagawa (2026, Appendix B.4) subtract observed errors from total-count thresholds before auditing the residual population, directly preceding our attainer. Our stochastic total-count correction supplies validity, not a new sampling method. Anthony and Salehzadeh Nobari (2026, Theorem 13, Remark 14) give excluded-pool hypergeometric certification and an adaptive inspection lower bound for zero misses, not an all-outcome adversarial power frontier. Their planted-set coupling with exact avoidance already gives our $u = 0$ endpoint. Our distinct target is (2): endogenous native composition and a matching full-transcript experiment.

Robust testing. Sample-aware deletion retains positions in robust testing (Vallinayagam et al., 2026). For deterministic totals, standard symmetry and NP arguments already yield residual hypergeometric ROCs. Our step handles stochastic totals while matching full observed records under legal laws; conditioning the ordered null on a total changes its class and cannot replace that step. Noisy certification and active hypothesis testing also precede our experiment (Anthony and Salehzadeh Nobari, 2026; Naghshvar and Javidi, 2013).

Unknown label noise and identification. Unrestricted noise mechanisms can make clean distributions unidentifiable (Scott et al., 2013; Liu et al., 2023). Our corollary specializes that established principle: it locates the missingness–reliability boundary for the full selective-audit transcript and gives the exact power frontier throughout the nonblind region, with a finite-valid attainer. Repeated labels or informative proxies can alter identification; they are not supplied in this experiment.

Additional lineage for power-floor restrictions and operational motivation is given in Appendix G.1.

9 Conclusion and limitations

A reliability claim requires a separate account of its auditability. Our full-record frontier and noisy extensions distinguish three shortages: too few queries, insufficiently known label quality, and too small a cohort. The response can be more queries, stronger reliability information, or a larger cohort. These are conditional design guarantees, not validation of deployment assumptions; nonrejection never certifies safety.

Observation and external validity. We exclude covariates, uncommitted selection, concurrent releases, and unrecoverable outcomes. Repeated audits need separate global error control. Marginal conformal coverage implies neither $\mathcal{C}$ nor $\mathcal{P}$; neither need bound every coordinate's marginal risk. Results are composite worst-case guarantees, not count sufficiency for every dependent law. Reliability transport and optimal calibration/query allocation remain unresolved.

Asymptotic and design scope. Gap, positive level, and query count are fixed; vanishing-gap, growing-k, and vanishing-level extensions are unproved. The predictable-budget critical noisy result covers only zero root-n offset. Finite rules need not be uniformly optimal; noisy design can know p, s and be inefficient. Integer certificates validate supplied calculations, not protocol assumptions, arbitrary implementations, or post-hoc cohort selection.

Reproducibility statement

Complete proofs are in the PDF appendices. The anonymous numerical supplement replays the included tests, exact primal/dual witnesses, calibration checks, and the separately documented query-budget calculation. Rational certificates and floating-point diagnostics are distinguished; optimization supplies proposals, not validity. The combined supplement replays the finite-budget and native-information calculations, rebuilds their figures, and independently verifies the integer decision tables by exact complete-population sums. It also replays the two-target mixture, its private branch sampler, the complete-information cohort threshold, and exact checks of the reliability overlap construction. It also evaluates the robust noisy frontier and checks finite-response stochastic inequalities with rational arithmetic. Numerical checks support, but do not replace, the proofs. It does not rebuild the manuscript PDF or all historical research commands.

ETHICS STATEMENT

The reported analyses use mathematical populations and deterministic calculations, not a new human-subject study. A deployed audit would additionally need a justified labeling standard, access controls, and an assessment of heterogeneous errors; worst-case global risk bounds do not certify fairness across unmodeled subgroups.

AI USE STATEMENT

The original hypotheses and experiments were developed and conducted by the author. Generative AI assisted literature search, subsequent mathematical claims and proofs, computational design and implementation, result interpretation, and manuscript revision. The author is responsible for the final work.

REFERENCES

Martin Anthony and Kaveh Salehzadeh Nobari. Finite-Sample Coverage Audits for High-Recall Candidate Generation: Certification and Learning-Theoretic Design. arXiv:2607.21480v1, 2026. https://arxiv.org/abs/2607.21480v1.

Kianoosh Ashouritaklimi, Valentin Kilian, Daolang Huang, Tom Rainforth, and François Caron. Prediction-Powered Active Testing. arXiv:2607.08347v1, 2026. https://arxiv.org/abs/2607.08347v1.

Attila Dobi, Aravindh Manickavasagam, Benjamin Thompson, Xiaohan Yang, and Faisal Farooq. Measuring the Prevalence of Policy-Violating Content with ML-Assisted Sampling and LLM Labeling. arXiv:2602.18518v1, 2026. https://arxiv.org/abs/2602.18518v1.

Jinshuo Dong, Aaron Roth, and Weijie J. Su. Gaussian Differential Privacy. arXiv:1905.02383v3, 2019, Section 3.1. https://arxiv.org/abs/1905.02383v3.

European Commission, CORDIS. CONTRACT: Conformal Calibration for Reliable AI-Based Wireless Communications. Project 101198347, fact sheet, 2025. Accessed September 16, 2026. https://cordis.europa.eu/project/id/101198347.

Vineet Goyal and Karl Sigman. On Simulating a Class of Bernstein Polynomials. 2012. Author manuscript, Proposition 2.2 and Theorem 2.3. https://www.columbia.edu/~vg2277/simulate_bernstein.pdf.

D. G. Horvitz and D. J. Thompson. A Generalization of Sampling Without Replacement from a Finite Universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952. https://doi.org/10.1080/01621459.1952.10483446.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, non-parametric, nonasymptotic confidence sequences. *Annals of Statistics*, 49(2):1055–1080, 2021. https://arxiv.org/abs/1810.08240.

Masahiro Kato and Kei Nakagawa. Sequential Audit Sampling for Finite Populations with Exact and Simulation-based Guarantee. arXiv:2604.06116v2, 2026. https://arxiv.org/abs/2604.06116v2.

Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Tom Rainforth. Active Testing: Sample-Efficient Model Evaluation. In *ICML*, PMLR 139:5753–5763, 2021. https://proceedings.mlr.press/v139/kossen21a.html.

Tetsuji Kuboyama. How Many Labels Does Model Choice Need? Certificates and Budgets for Selective Prediction. arXiv:2609.18622v1, 2026. https://arxiv.org/abs/2609.18622v1.

Himabindu Lakkaraju, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables. In *KDD*, 2017. https://doi.org/10.1145/3097983.3098066.

E. L. Lehmann and Henry Scheffé. Completeness, Similar Regions, and Unbiased Estimation—Part I. *Sankhyā*, 1950. Section 4; original-paper facsimile. https://link.springer.com/content/pdf/10.1007/978-1-4614-1412-4_23.pdf.

Puheng Li, Tijana Zrnic, and Emmanuel J. Candès. Robust Sampling for Active Statistical Inference. arXiv:2511.08991v1, 2025. https://arxiv.org/abs/2511.08991v1.

Yang Liu, Hao Cheng, and Kun Zhang. Identifiability of Label Noise Transition Matrix. In *ICML*, PMLR 202:21475–21496, 2023. https://proceedings.mlr.press/v202/liu23g.html.

Mohammad Naghshvar and Tara Javidi. Active Sequential Hypothesis Testing. *Annals of Statistics*, 41(6):2703–2738, 2013. https://arxiv.org/abs/1203.4626.

NIST TREC Total Recall Track. Total Recall Guidelines, May 31, 2016. Task Operation. https://trec.nist.gov/data/total-recall/2016/Total%20Recall%20Guidelines%202016.html.

Jack Noonan, Adetola Adedamola Adediran, Robin Mitra, and Stefanie Biedermann. An integrated approach to test for missing not at random. arXiv:2208.07813v3, 2022. https://arxiv.org/abs/2208.07813v3.

Aleksandr Podkopaev and Aaditya Ramdas. Tracking the risk of a deployed model and detecting harmful distribution shifts. In *ICLR*, 2022. https://arxiv.org/abs/2110.06177.

Clayton Scott, Gilles Blanchard, and Gregory Handy. Classification with Asymmetric Label Noise: Consistency and Maximal Denoising. In *COLT*, PMLR 30:489–511, 2013. https://proceedings.mlr.press/v30/Scott13.html.

Osvaldo Simeone, Sangwoo Park, and Matteo Zecchin. Conformal Calibration: Ensuring the Reliability of Black-Box AI in Wireless Systems. arXiv:2504.09310v3, 2025. https://arxiv.org/abs/2504.09310v3.

Shankar Vallinayagam, Ankit Pensia, and Varun Jog. On the Sample Complexity of Robust Binary Hypothesis Testing. arXiv:2605.24741v1, 2026. https://arxiv.org/abs/2605.24741v1.

Tijana Zrnic and Emmanuel J. Candès. Active Statistical Inference. In *ICML*, PMLR 235, 2024. https://proceedings.mlr.press/v235/zrnic24a.html.

A PROOF OF THE AUDIT FRONTIER

This proof covers the experiment in Section 2, including randomized adaptive policies with a hard cap of k distinct pending queries. There are no additional native arrivals, covariates, or outcome proxies during the audit phase. The adversary freezes the pending set before the independent audit seed. All queried labels are obtained truthfully. Throughout, $0 < a < p < 1$, $0 < \delta < 1$, $k \geq 1$ is fixed, $b_n/n \to \beta \in (0, 1)$, and $\epsilon = p - a$. The alternative is iid, not an arbitrary exchangeable law.

A.1 FINITE VALIDITY AND THE ATTAINING RULE

For $t_n = n^{3/4}$ put $\eta_n = \exp(-2t_n^2/n)$ and $H_n = \min(n, \lfloor na + t_n \rfloor)$. For the finitely many n with $\eta_n \geq \delta$ or $b_n < k$, never reject. Otherwise let

$$s_n = \frac{\delta - \eta_n}{1 - \eta_n}, \qquad M_0(j) = \min(b_n, H_n - j) \quad (j \leq H_n).$$

Reject if the native miss count $j > H_n$. If $j \leq H_n$, sample k pending positions uniformly without replacement and use the nondecreasing randomized upper-tail test $f_{n,j}$ of size s_n under $\mathrm{HG}(b_n, M_0(j), k)$. Here $\mathrm{HG}(b, M, k)$ denotes the number of misses in k draws from a population of size b containing M misses. Reject above a threshold, randomize at the threshold, and accept below. The definition includes degenerate laws: if $M_0 = 0$, reject positive counts and randomize at zero; if $M_0 = b_n$, randomize at k and accept smaller counts. Final test randomization is independent of the data, mask, and audit draws. The rule uses no value of p.

For any event-order conditional-risk null, conditional Hoeffding's bound gives $\Pr(K > H_n) \leq \eta_n$. For completeness, conditional bounded-increment Hoeffding's lemma implies $\mathbb{E}\exp(\lambda(K - na)) \leq \exp(n\lambda^2/8)$ for $\lambda > 0$; optimizing the exponential bound gives $\exp(-2t_n^2/n)$. On a complete latent vector and frozen mask with $K \leq H_n$, direct rejection is impossible and the actual hidden miss count $K - j$ is at most $M_0(j)$. *Condition on that entire vector and mask*, not only on the observed native history. The audit count has law $\mathrm{HG}(b_n, K - j, k)$. Hypergeometric stochastic monotonicity and monotonicity of $f_{n,j}$ bound its conditional rejection probability by s_n. On the complementary event the probability is at most one. Hence size is bounded uniformly by

$$\eta_n + (1 - \eta_n)s_n = \delta. \tag{A1}$$

This argument permits arbitrary outcome-dependent randomized masks. In implementation one may replace H_n by an exact binomial upper quantile: conditional Bernoulli risks at most a imply $K \preceq \mathrm{Bin}(n, a)$ by sequential uniform coupling. Use an exceptional probability at most η_n, or its exact binomial tail, in (A1).

A.2 A CONTINUITY FACT USED AT THE ENDPOINTS

Write $R(Q, P; s) = \sup_{0 \leq f \leq 1,\, Qf \leq s} Pf$ for randomized NP power. At every fixed $s \in (0, 1)$ this value is continuous in the two finite-alphabet probability vectors and in s, including degenerate distributions. Indeed, if Q' is within total variation e of Q, a size-s test has size at most $s + e$ under Q'. Scaling it by $s/(s + e)$ restores size s and loses at most e/s power. Perturbing P changes power by at most its total variation. Reverse the roles for the other direction. Scaling similarly controls nearby positive sizes. No continuity at size zero is asserted.

For fixed k, sampling without replacement can be coupled to sampling with replacement with failure probability at most $k(k - 1)/(2b)$: draw k indices with replacement and keep them when all are distinct. Thus, uniformly in M,

$$d_{\mathrm{TV}}\{\mathrm{HG}(b, M, k), \mathrm{Bin}(k, M/b)\} \leq \frac{k(k-1)}{2b}. \tag{A2}$$

Hypergeometric count laws have nondecreasing likelihood ratio when M increases, including the new upper support points with zero null mass. This follows directly by taking consecutive ratios of the combinatorial probability masses. Consequently $f_{n,j}$ is NP-optimal against every larger hidden count.

A.3 Uniform lower bound above the critical missing fraction

Put $L_n = \lceil np - t_n \rceil$. Under the iid alternative, $A_n = \{K \geq L_n\}$ has probability at least $1 - \eta_n$. For every feasible mask on A_n, either $j > H_n$ and rejection is certain, or

$$\max(0, L_n - b_n) \leq j \leq \min(H_n, n - b_n). \tag{A3}$$

Eventually $L_n > H_n$. In the second case $M_1 = L_n - j \leq b_n$ is strictly larger than $M_0 = H_n - j \geq 0$; in particular the clipping branch $M_0 = b_n$ cannot occur. The actual hidden count is at least M_1. The rule's conditional power is at least the NP power $B_{n,j}$ for $\mathrm{HG}(b_n, M_0, k)$ versus $\mathrm{HG}(b_n, M_1, k)$ at size s_n.

To justify minimization over masks, take any sequence of feasible nonrejecting j_n along a subsequence of n. Compactness yields a further subsequence with $j_n/n \to y$. Relation (A3) implies

$$\max(0, p - \beta) \leq y \leq \min(a, 1 - \beta), \qquad u = \frac{a - y}{\beta} \in I_\beta,$$

where I_β is the interval in Theorem 1. Then $M_0/b_n \to u$ and $M_1/b_n \to u + \epsilon/\beta$. By (A2) and positive-size ROC continuity, $B_{n,j_n} \to R_k(u, u + \epsilon/\beta; \delta)$. If the infimum over feasible j fell below the minimum of this continuous function on I_β by a fixed positive amount infinitely often, an approximately minimizing sequence would contradict this subsequence argument. Multiplying by $\Pr(A_n) \to 1$ proves the desired minimax lower bound. It is achieved by one p-agnostic rule pointwise for every fixed $p > a$.

A.4 Matching all-observation upper bound, including adaptive queries

Fix y strictly inside $[\max(0, p - \beta), \min(a, 1 - \beta)]$, an interval of positive length when $\epsilon < \beta < 1$, and put $j_n = \lfloor ny \rfloor$. Choose iid Bernoulli(a) as one admissible null and iid Bernoulli(p) as the alternative. In either world, whenever $\max(0, K - b_n) \leq j_n \leq \min(K, n - b_n)$, reveal j_n uniformly chosen misses and $n - b_n - j_n$ uniformly chosen hits. Otherwise choose any legal mask. Strict interiority gives linear slack in the random feasibility constraints; the exceptional probability tends to zero in both worlds by concentration. Boundary y values are handled only by subsequent interior approximation.

On the feasible branch, conditional on K, every native ternary status configuration with j_n misses and $n - b_n - j_n$ hits is equally likely. This law does not depend on K or on the world. Given K and a native configuration, the pending miss locations are uniform among all $(K - j_n)$-subsets of the b_n pending entries.

Consider any randomized policy that selects an as-yet-unqueried pending position using this native configuration, its independent random seed, and past query identities and labels. After t queries containing z misses, the probability of a miss at *any* chosen unqueried position is $(K - j_n - z)/(b_n - t)$. Therefore simulate its complete transcript by first generating a length-k without-replacement binary sequence, then feeding the next bit to each selected index. If the policy stops, discard the unused bits. This includes the policy seed and identities in the simulated transcript. Conditional on the final total Z of that length-k sequence, its order is uniform among the $\binom{k}{Z}$ orders, independently of K and the world. Thus the policy transcript is a common Markov postprocessing of $Z \sim \mathrm{HG}(b_n, K - j_n, k)$. The postprocessing may depend on the policy and n, but is identical in the two worlds. This step also covers nonsymmetric policies and tests; no permutation invariance of the full conditional-risk null is assumed.

For $\theta \in \{a, p\}$, $K/n \to \theta$ in mean, and the feasible-branch conditioning has probability tending to one. Using (A2), and coupling k Bernoulli variables whose probabilities differ, gives total variation convergence of Z to $\mathrm{Bin}(k, (\theta - y)/\beta)$. The fallback contributes only its vanishing probability. Contraction under the common postprocessing makes this error uniform over the policies. Every valid test has size at most δ in the chosen null world. Its induced binomial test therefore has size at most $\delta + o(1)$, and its power differs by $o(1)$. NP optimality and the preceding continuity fact give

$$\limsup_n V_{n,k}(\delta) \leq R_k((a - y)/\beta, (p - y)/\beta; \delta).$$

Taking the infimum over interior y and then using continuity at the endpoints proves the matching upper bound. This proves Theorem 1 for $\beta > \epsilon$, and proves that adaptive choice of audit indices cannot improve the limiting maximin value in this no-side-information experiment.

A.5 CRITICAL AND SEPARATED REGIMES

If $0 < \beta < \epsilon$, then $L_n - b_n > H_n$ eventually; on A_n every legal mask triggers direct rejection. No audit is required for consistency. If $\beta = \epsilon$, (A3) instead squeezes $j/n \to a$ uniformly over all nonrejecting masks. Consequently $M_0/b_n \to 0$ and $M_1/b_n \to 1$ uniformly. The two hypergeometric laws converge to point masses at zero and at k. Because $k \geq 1$ and $\delta > 0$, their NP power tends to one. This proves consistency uniformly over masks, using the same finite-valid rule. It permits $b_n - n\epsilon = o(n)$ of either sign. It does not by itself establish a gain over zero audits for every such second-order sequence.

B ZERO-AUDIT CRITICAL COMPARISON

Use the finite no-audit hull of Proposition 8, proved in Appendix L. Let $(b_n - n\epsilon)/\sqrt{n} \to s \in \mathbb{R}$, $\sigma_\theta = \sqrt{\theta(1-\theta)}$, and $c = \sigma_a/\sigma_p$. The uniform binomial normal approximation, for $z_n = (r-1-na)/\sqrt{n}$, gives the two hull coordinates

$$x_n(r) = 1 - \Phi(z_n/\sigma_a) + o(1), \qquad y_n(r) = 1 - \Phi((z_n + s)/\sigma_p) + o(1)$$

uniformly over thresholds. Hence $\sup_r |y_n(r) - g_s(x_n(r))| \to 0$, where $g_s(x) = \Phi(c\Phi^{-1}(x) - s/\sigma_p)$, extended by $g_s(0) = 0, g_s(1) = 1$. Uniform continuity at the endpoints justifies this substitution. The allowed normalized threshold range tends to the whole real line and its lattice spacing tends to zero, so these points approximate the entire graph.

Let C_s be its least concave majorant. Concavity bounds every finite hull mixture by $C_s(\delta) + o(1)$. Conversely, approximate a finite convex combination of graph points on the upper hull by lattice thresholds. If the resulting size exceeds δ by $o(1)$, scale the test down by $\delta/(\delta + o(1))$; its power changes by $o(1)$. This proves convergence to $C_s(\delta)$. For any $x_0 \in (\delta, 1)$, at most δ/x_0 of a size-δ mixture's weight is at $x > x_0$. Therefore

$$C_s(\delta) \leq 1 - (1 - \delta/x_0)(1 - g_s(x_0)) < 1. \tag{A4}$$

For $s = 0$, $g_0'(x) = c\exp\{(1-c^2)(\Phi^{-1}(x))^2/2\}$. If $c < 1$, g_0 is concave below $1/2$ and convex above. Its envelope follows g_0 until the unique $t \in (0, 1/2)$ satisfying $(1-t)g_0'(t) = 1 - g_0(t)$, and then the chord to $(1, 1)$. Uniqueness follows since the derivative of the difference between these two sides is $(1-t)g_0''(t) < 0$, with opposite limiting signs at zero and $1/2$. If $c > 1$, the envelope is the chord from $(0, 0)$ to the unique $t \in (1/2, 1)$ satisfying $tg_0'(t) = g_0(t)$, followed by g_0. For $c = 1$ the envelope is the diagonal.

The restriction on second-order budget deviations is substantive. If $h_n = b_n - n\epsilon = o(n)$ and $h_n/\sqrt{n} \to +\infty$, each threshold of null size bounded away from one has alternative power tending to zero. Mixture weights then imply the no-audit limit δ, attained by a coin. If $h_n/\sqrt{n} \to -\infty$, the threshold near $na - h_n/2$ has null size tending to zero and worst alternative power tending to one. Thus a common limiting missing fraction does not determine the zero-audit critical limit.

C PROOF OF THE FINITE ONE-QUERY OBSTRUCTION

Fix the parameters in Proposition 6. Average any admissible procedure over a fresh uniform permutation of identities, mapping its query back to the original population. In the iid subproblem, each permuted mask corresponds to a legal original mask. Averaging preserves iid-null validity and the all-alternative power floor, and cannot decrease worst-mask power. We do not claim it preserves validity over the full ordered conditional-null class; this is only a necessary-constraint upper-bound argument.

Pad a no-query branch by an unused query. Conditional on the permuted native configuration and the original procedure's seed, the unexposed part of the permutation places pending losses uniformly into the remaining slots. The selected slot has miss probability $(K-j)/b$, independently of that conditioned information. Averaging the two possible final decisions gives $f(j,0), f(j,1) \in [0,1]$, independent of p and K, with expected rejection

$$h_j(K) = \left(1 - \frac{K-j}{b}\right) f(j,0) + \frac{K-j}{b} f(j,1), \quad j \leq K \leq j + b. \tag{A6}$$

This also covers original policies with nonuniform identity-sensitive choices.

For every fixed legal mask, the iid rejection probability is a polynomial in p. Its power floor for all $p > a$ and size bound at a imply, by continuity, size exactly δ at a. Compare two count-choice kernels differing only at one total K. Since every $\mathrm{Bin}(n, a)$ mass is positive, $h_j(K)$ must agree across all feasible j. For fixed j, (A6) is affine in K on the integer interval $[j, j + b]$. Adjacent intervals overlap in at least two distinct counts because $b \geq 2$, so the affine functions are identical. They cover $0, \ldots, n$; hence $h_j(K) = A + BK/n$ throughout. When $b = n$ there is only one interval and the same conclusion is immediate.

Exact iid size gives $A + Ba = \delta$. The all-alternative power floor implies $B \geq 0$. Values at $K = 0, n$ lie in $[0, 1]$, so $B \leq \min(\delta/a, (1 - \delta)/(1 - a))$. The symmetrized procedure's power is therefore at most H, bounding the original procedure's worst-mask power as well. The HT construction in the main text gives equality and is valid for the conditional null since $\mathbb{E}K/n \leq a$. The argument fails at $b = 1$, where a query recovers all labels and adjacent intervals share only one count.

D Finite computation and its limitations

The implementation uses exact rational hypergeometric probabilities for small cases and an exact binomial count cap. For any specified response $f_j(z)$, let $w_K = \max_j \mathbb{E}_{\mathrm{HG}(b,K-j,k)} f_j(Z)$ over feasible j. Its support over the *conditional*, not just iid, null is computed backwards from $B(n, K) = w_K$ by

$$B(t, r) = \max\{B(t + 1, r), (1 - a)B(t + 1, r) + aB(t + 1, r + 1)\}. \tag{A5}$$

The conditional success probability at each history ranges over $[0, a]$, so linearity puts a maximizer at 0 or a. Backward induction shows that the count-state recursion is sufficient for this symmetric terminal payoff. An independent enumeration over full binary histories checks this claim on small cases; it is not an invalid symmetrization of the null class.

For $k = 1$, the worst-mask power of the *specified* finite rule can also be evaluated without a quadratic scan. With cap H and residual size s, the conditional response for total K and native count $j \leq H$ equals $s(K - j)/b$ if $j \leq H - b$. Otherwise it is

$$\frac{s(K - j)}{H - j} \quad \text{when } (H - j)/b > s, \qquad s + \frac{(1 - s)(K - H)}{b - H + j} \quad \text{when } (H - j)/b \leq s.$$

These functions are monotone on each interval between $H - b$, $H - bs$, and H; for $j > H$ rejection is certain. It is therefore enough to check feasible interval endpoints and adjacent integers to those breakpoints. Summing the minimum against $\mathrm{Bin}(n, p)$ yields the rule's exact mathematical worst-mask power. The large-n implementation evaluates this formula in floating point, not certified interval arithmetic. It is not an optimization over all finite-n tests.

Certifying a continuous nuisance minimum. For a cell $[l, r] \subset I_\beta$ of width at most d, calibrate a monotone binomial NP test at null parameter r. For every $u \in [l, r]$, it has size at most δ under parameter u, and its power at $u + d$ is at least its power at $l + d$. Consequently

$$R_k(r, l + d; \delta) \leq \inf_{u \in [l,r]} R_k(u, u + d; \delta).$$

Take the minimum of these lower bounds over a covering partition; the ROC at any feasible point is an upper bound on the continuous minimum. Our budget certificate uses exact rational probabilities on 1,000 equal cells of $[0, 1/5]$, with $d = 1/5, \delta = 1/20$. It certifies the lower bound above $19/20$ at $k = 56$, and supplies the exact $k = 55, u = 1/5$ witness below $19/20$. More observations cannot decrease optimal power, since a test can ignore them. This proves that 56 is the smallest fixed query budget for that limiting target, without assuming the worst composition is a grid point. The same certificate bounds the $k = 14$ frontier. These are limiting-budget statements, not uniform finite-cohort performance bounds.

E An exact finite query-budget threshold

The limiting frontier need not give a sufficient finite budget, and a conservative finite procedure need not identify the information limit. Here we separate these effects in the declared experiment

$n = 1000, b = 500, a = 1/10, p_\star = 1/5, \delta = 1/20$. The experiment is truthful, single-deadline, with an outcome-aware committed mask and private audit randomness. The following bounds apply both to the original ordered-null/iid alternative problem and to its $\mathcal{C}_0(a), \mathcal{C}_1(p_\star)$ extension. The exact budget threshold is a design-point result, not a general finite minimax formula.

E.1 A FINITE UPPER BOUND FOR EVERY ADAPTIVE POLICY

Proposition 2 (Legal two-world comparison)**.** *Fix n, b, k, a, p, δ in the truthful experiment. Choose a fixed integer j_0 and define*

$$j(K) = \min\{\max\{j_0, \max(0, K-b)\}, \min(K, n-b)\}.$$

For $\theta \in \{a, p\}$ let

$$q_\theta(j, z) = \sum_{K: j(K)=j} \mathbb{P}\{\mathrm{Bin}(n, \theta) = K\}\mathbb{P}\{\mathrm{HG}(b, K-j, k) = z\}.$$

Then $V_{n,k}(\delta)$ is at most the ordinary randomized NP power for the joint masses q_a versus q_p. The same upper bound holds for the count-tail-class problem.

Proof. Take iid Bernoulli latent vectors in both worlds. Given K, release exactly $j(K)$ errors and $n - b - j(K)$ correct labels, uniformly within type. This is a legal kernel for every K, without conditioning away rare populations. Given $K, J = j$, the native ledger is uniform on its $\binom{n}{b}\binom{n-b}{j}$ possible patterns; its law depends only on j. Given K and that ledger, the pending error subset is uniform of size $K - j$. After t distinct adaptive queries with z errors, every unqueried pending identity therefore has conditional error probability $(K - j - z)/(b - t)$.

Pad an early-stopped audit privately to k queries. Its total Z has the hypergeometric law above. Given Z, the ordered bits are uniformly arranged, independently of K and the hypothesis. For a fixed policy a common Markov kernel takes (J, Z) to the native pattern, independent seed, ordered bits, adaptive identities, and the original stopping transcript. Consequently no full transcript can improve NP power for this pair. A test valid for the whole null is valid at this iid null; its worst-alternative power is no larger than its power at this chosen iid alternative and mask. Supremizing over policies gives the bound. The iid worlds also belong to the respective count-tail classes. No permutation invariance of the ordered null class, or validity of the witness NP rule for other masks, is being asserted. □

We use $j_0 = 0$. When $K \le b$, $J = 0$; when $K > b$, $J = K - b$ and $Z = k$. Both branches are retained exactly. The latter branch's likelihood ratios exceed every $J = 0$ ratio: the latter are null-weighted averages of the increasing iid total-count likelihood ratio over $K \le b$. Its null mass is below δ at the certified design, so all its atoms are rejected by NP and its exact mass can be aggregated. This is not a tail approximation. An independent implementation instead constructs and sorts every joint atom without aggregation or monotone-likelihood-ratio assumptions.

E.2 EXACT NULL RECALIBRATION OF A FEASIBLE RESIDUAL RULE

Fix a cap H and common stage s. For $j \le H$, use the randomized upper-tail test $f_{j,s}(z)$ against $\mathrm{HG}(b, \min(b, H-j), k)$, rounding its boundary rejection probability downward to the grid $2^{-80}\mathbb{Z}$. For $j > H$ reject directly. Uniformly sample k distinct pending indices. The supplied integer table implements a boundary probability $g/2^{80}$ by comparing an independent uniform 80-bit word with g; it does not implement an uncertified floating approximation.

Write $F_s(j, K-j) = \mathbb{E}_{\mathrm{HG}(b,K-j,k)} f_{j,s}(Z)$, and let $A_K(s)$ and $B_K(s)$ be its minimum and maximum over every feasible $\max(0, K-b) \le j \le \min(K, n-b)$. Define

$$S(H, s, k) = \sum_{K=0}^{n} \mathbb{P}\{\mathrm{Bin}(n, a) = K\}B_K(s), \qquad L(H, s, k; p) = \sum_{K=0}^{n} \mathbb{P}\{\mathrm{Bin}(n, p) = K\}A_K(s). \tag{11}$$

These are complete-population sums. Their envelopes are computed over all feasible masks, not only endpoint compositions.

Validity beyond iid. The response is nondecreasing in z and j: decreasing the reference error count lowers its upper tails, and common downward rounding preserves the order. For a fixed mask and uniformly drawn subset, turning a latent correct label into an error cannot decrease rejection. Taking extrema over the same mask collection preserves this order. Symmetry makes both extrema functions only of K, so both A_K and B_K are nondecreasing. Stochastic count domination then bounds size under $\mathcal{C}_0(a)$ by S and power under $\mathcal{C}_1(p)$ below by L. The iid boundary laws and countwise extremizing masks attain these sums. Thus the size and power quantities in (11) are exact for this rule and these classes, not plug-in estimates of a conditional pending distribution.

Calibration and selection. For fixed H, k with $S(H, 0, k) \leq \delta$, select the largest stage $s \in [0, 1] \cap 2^{-80}\mathbb{Z}$ with $S(H, s, k) \leq \delta$; otherwise discard that cap. All reported caps pass this feasibility check. Increasing s increases every row, justifying integer bisection in s. The final stage and its next lattice point are checked by exact integer inequalities; floating proposals are not acceptance criteria. In particular, the stage need not itself be at most δ: it is the complete worst-mask expectation S that controls size. No monotonicity in k of this procedure is used.

The declared candidate profile was $H \in \{116, 120, 124, 128, 136, 144, 160\}$, with every integer budget from 56 to 100 screened, plus the illustrative budget 14. Selecting $H = 144$ used the planning alternative $p_\star = .2$, not observed target labels. For fixed H, k, the stage calculation uses null inputs only. This distinction precludes calling the selected design p-agnostic. With the fixed p-agnostic cap $H = 124$, null-only recalibration already certifies power .952138999000587 at 59 queries (58 gives .948671911040563). Neither cap selection nor the witness upper pair is claimed globally optimal.

E.3 Certified threshold, matched comparisons, and sensitivity

Corollary 5 (Minimum budget at the declared finite design)**.** *For $n = 1000, b = 500, a = .1, p_\star = .2, \delta = .05$, the minimum hard query budget permitting worst-case power at least .95 is exactly 57, under either the original null/iid alternative or the count-tail extension.*

Proof. Exact rational evaluation of Proposition 2 and (11) gives the outward bounds

$$.948587615767662 \leq V_{1000,56}(.05) \leq .949008177854816,$$
$$.952436876791621 \leq V_{1000,57}(.05) \leq .952722276007193.$$

The lower rules have exact size at most .05 over $\mathcal{C}_0(.1)$ and the displayed power over $\mathcal{C}_1(.2)$; the upper witnesses lie inside the narrower original classes. Hard-budget policy classes are nested because extra queries can be ignored. Hence the 56-query upper bound excludes every smaller budget, while the 57-query lower rule meets the target. The exact numerator/denominator witnesses and independent full-sum replay are provided in the numerical supplement. □

For the decisive $k = 57, H = 144$ rule, the stage is $573556229773914321842136/2^{80}$. Its exact size is below $1/20$, whereas the next stage integer has size above $1/20$. The table SHA-256 is `ef3addc7dc707de9342a72d5f6b346b69a175d83b160889d30c9799a7943bd77`. The unresolved power gaps are about .000421 at 56 and .000285 at 57; exact budget minimality does not identify either exact minimax power.

Table 4: Matched finite powers, rounded for display. Every method uses the same truthful, single-deadline count-tail experiment and size ceiling .05. The upper column bounds every adaptive policy, not just these procedures. All reported residual-rule powers have complete-sum integer certificates.

k	Cap-tail $H = 124$	Recalibrated $H = 124$	Recalibrated $H = 144$	Upper
14	.335562	.507937	.523647	.527928
56	.799016	.944738	.948588	.949009
57	.807873	.946614	.952437	.952723
100	.955061	.995456	.995925	.995926

The cap-tail calibrated stage is about .0442464, but its actual worst-mask sizes at budgets $14, 56, 57, 100$ are only .0218880, .0120388, .0118573, .00920764. Recalibration addresses this un-

used size, explaining most of the improvement; it does not acquire extra information. The 100-query sufficient certificate remains valid for that cap-tail calibrated rule (Appendix T).

Same-contract simple baselines. The exact passive ROC hull has power .05 at this design: all 500 nonconstant threshold points satisfy $\mathbb{P}\{\mathrm{Bin}(1000, .2) \geq r + 500\} \leq \mathbb{P}\{\mathrm{Bin}(1000, .1) \geq r\}$, checked by integer sums, so the hull lies on the diagonal after randomization. With any $k \geq 1$, uniform sampling makes $\widehat{r} = (J + bZ/k)/n$ unbiased for K/n. The HT test $\phi = \delta + \frac{1}{2}(\widehat{r} - a) = \widehat{r}/2$ is in $[0, 1]$, with worst size .05 and power .10 in the stated classes. An independent .05 coin gives power .05. These are contract-matched baselines, not claims that HT is the strongest established procedure: recalibrated residual sampling in Table 4 is itself a standard stronger comparator. Feature-assisted or sequential methods with different information/error guarantees are not forced into this table.

Table 5: Worst-mask power of the same frozen $k = 57, H = 144$ integer table. No stage or cap is retuned to actual p. The design alternative is .2.

p	.12	.15	.18	.19	.20	.21	.22	.25
Power	.011549	.477821	.857861	.915060	.952437	.975082	.987793	.999047

The severe .12 result is below both the .05 coin and HT's .06. The 95% claim is therefore not a guarantee against all $p > a$. Monotonicity of A_K extends it to $\mathcal{C}_1(.2)$, including iid $p \geq .2$; an average-risk inequality alone does not justify that class. The supplied implementation establishes arithmetic validity, not truthful adjudication, cohort recruitment, mask commitment, or repeated-audit control.

F Two-target design and complete-information limits

This section uses the same truthful count-tail experiment as Appendix E. It asks how a finite certificate can guide a choice when the actual alternative risk is unknown. Convex randomization, binomial NP testing, and the coarsening argument below are standard tools; the results here are certified design consequences, not new testing primitives.

F.1 One sample, a private choice of two rules

Let $f_{124}(j, z)$ and $f_{144}(j, z)$ be the frozen 80-bit response tables from exact null calibration with $n = 1000, b = 500, k = 57, a = .1, \delta = .05$. Both are nondecreasing in j, z and have size at most $1/20$ under $\mathcal{C}_0(.1)$. Define the mixed response

$$f_{\mathrm{mix}}(j, z) = \tfrac{5}{12} f_{124}(j, z) + \tfrac{7}{12} f_{144}(j, z).$$

The cap and mixture weight are fixed before target labels are inspected. After mask commitment, privately select a branch: generate a uniform four-bit word; discard and regenerate if it is at least 12; select $H = 124$ for values 0–4 and $H = 144$ for values 5–11. This samples exactly $5/12$ and $7/12$, not an 80-bit approximation. It terminates almost surely with $4/3$ expected words, not a fixed random-bit bound. Draw one uniform sample of 57 distinct pending positions, obtain their truthful labels, and use the selected table's independent uniform 80-bit boundary word. The branch, sample, and boundary randomness must be mutually independent and hidden until the mask is committed. There are at most 57 costly label requests; random-word rejection consumes no labels. The supplied interface freezes the ledger and checks inputs, but software cannot establish these independence or truthfulness assumptions.

For any common law and mask, rejection probability is the same weighted sum of component probabilities. Hence its infimum is at least the weighted component infima; its null supremum is at most the weighted component suprema. Using downward-rounded component certificates gives

$$\inf_{\mathcal{C}_1(.12),\,\kappa} \mathbb{E} f_{\mathrm{mix}} \geq \frac{37538707463231}{750000000000000} > .05,$$

$$\inf_{\mathcal{C}_1(.2),\,\kappa} \mathbb{E} f_{\mathrm{mix}} \geq \frac{2850031853990003}{3000000000000000} > .95.$$

Table 6: Two-target tradeoffs at $n = 1000, b = 500, a = .1, \delta = .05$. Certified worst-case lower powers, rounded down, under truthful $\mathcal{C}$ classes; every rule has size at most .05. Caps and mixing weight are fixed before target labels are seen.

Rule	Queries	$p = .12$	$p = .20$
Exact null, fixed cap $H = 124$	57	.10395	.94661
Exact null, planning cap $H = 144$	57	.01154	.95243
Private $5/12$–$7/12$ mixture	57	.08661	.95017
Exact null, fixed cap $H = 124$	59	.10598	.95213

These conservative bounds prove both targets without requiring coincident least-favourable masks. They do not equal the actual mixed-rule powers.

For an exact evaluation, let

$$F_h(j, K-j) = \mathbb{E}_{\mathrm{HG}(500,K-j,57)} f_h(j, Z), \qquad h \in \{124, 144\}.$$

Form $(5F_{124} + 7F_{144})/12$ *before* minimizing or maximizing over the common feasible j at each K. The response has denominator $12\,2^{80}\binom{500}{57}$. Enumerating all 251,001 feasible (K, j) pairs over all 1,001 totals, followed by the complete binomial sums, gives

$$\sup_{\mathcal{C}_0(.1),\,\kappa} \mathbb{E} f_{\mathrm{mix}} \in [.049195054457213747575129,\ .049195054457213747575130],$$

$$\inf_{\mathcal{C}_1(.12),\,\kappa} \mathbb{E} f_{\mathrm{mix}} \in [.086612158153241,\ .086612158153242],$$

$$\inf_{\mathcal{C}_1(.2),\,\kappa} \mathbb{E} f_{\mathrm{mix}} \in [.950175402109023,\ .950175402109024].$$

Monotonicity in j, z implies monotone population-total extrema, as in Appendix E; stochastic count domination proves the class transfer. Boundary iid laws with countwise extremizing masks attain these specified-rule sums. Thus mixing the per-mask responses, not their separately computed extrema, is essential for the exact result.

The all-policy 56-query upper bound is below .95 at $p = .2$. Consequently the minimum budget satisfying *both* a .05 floor over $\mathcal{C}_1(.12)$ and .95 power over $\mathcal{C}_1(.2)$ is exactly 57. This is budget minimality for the two targets, not optimality of the mixture weight or the power function. At .11 the same rule has power in $[.016951039730228, .016951039730229]$, so it has no universal .05 floor for all $p > .1$. Its .12 power also remains far from the full-information ceiling below. The separate $k = 59, H = 124$ comparison changes both the budget and cap relative to $k = 57, H = 144$; it does not isolate a two-label effect.

Independent checks. A separate arithmetic implementation reconstructs the frozen component tables. Both it and the standard-library producer enumerate every common mask count and compare all exact sensitivity fractions, terminal extrema, and extremizing mask witnesses. The independent calculation does not import the producer for this arithmetic. A separate operational check loads its query interface to test branch rejection sampling, boundary decisions, input failures, and the hard query budget. Both retain all binomial mass. The supplement distinguishes the exact results from rounded presentation and contains 13-risk sensitivity values, including adverse near-null cases.

F.2 A COHORT-SIZE OBSTRUCTION EVEN WITH COMPLETE LABELS

Proposition 3 (Complete-information benchmark)**.** *In the truthful experiment, allowing all pending labels to be recovered gives minimax power $R_n(a, p; \delta)$ under $\mathcal{C}_0(a), \mathcal{C}_1(p)$. The same value holds for the original ordered null and iid alternative.*

Proof. The randomized upper-tail binomial NP test of total K is nondecreasing. It has size δ for $\mathrm{Bin}(n, a)$ and power $R_n(a, p; \delta)$ for $\mathrm{Bin}(n, p)$, so stochastic order transfers both bounds to the count-tail classes. The ordered null is included in $\mathcal{C}_0(a)$. Conversely, iid Bernoulli vectors of rates a, p are legal, and their full-vector likelihood ratio is a nondecreasing function only of K. Their ordinary NP power bounds any valid procedure, even after all labels are observed. For this hard pair choose a

common fixed mask; its query record is generated from the latent vector and hypothesis-independent private randomness and cannot improve full-vector NP power. □

For $n = 1000, a = .1, p = .12, \delta = .05$, reject for $K > 116$ and reject at $K = 116$ with probability

$$\gamma = \frac{.05 - \mathbb{P}\{\mathrm{Bin}(1000, .1) > 116\}}{\mathbb{P}\{\mathrm{Bin}(1000, .1) = 116\}} \in [.658616929756346748, .658616929756346749].$$

Exact binomial arithmetic gives power in $[.653175248288369, .653175248288370]$. The equality covers truthful full information, not unknown label noise or mean-risk-only classes. Therefore querying all 500 pending labels cannot deliver 95% power at this gap and cohort size. A larger cohort or a different declared target is required; this obstruction does not establish an optimal partial-audit rule.

F.3 Exact inversion of the complete-information cohort requirement

At $a = 1/10, p = 3/25, \delta = 1/20$, the smallest complete-label cohort with power at least $19/20$ is $n = 2645$ under Proposition 3's population classes. For both $n = 2644$ and 2645, the exact NP rule rejects if $K > 290$ and randomizes at $K = 290$. With

$$q_n(j) = \binom{n}{j} 9^{n-j}/10^n, \qquad r_n(j) = \binom{n}{j} 3^j 22^{n-j}/25^n,$$

the randomization and power are

$$\gamma_n = \frac{1/20 - \sum_{j>290} q_n(j)}{q_n(290)}, \qquad R_n = \sum_{j>290} r_n(j) + \gamma_n r_n(290).$$

Exact integer sums establish $0 \le \gamma_n \le 1$ and the outward intervals

n	lower power endpoint	upper power endpoint
2644	.949974516611648439	.949974516611648440
2645	.950045945355148722	.950045945355148723

so $R_{2644} < 19/20 < R_{2645}$. No tail mass is discarded. Minimality follows because the iid NP experiment with $n + 1$ observations can ignore one observation, so its value is nondecreasing in n. Proposition 3 equates that value to the $\mathcal{C}$ minimax value at each n. This does not assume the $\mathcal{C}$ classes are closed under taking a subcohort. The threshold is a necessary cohort size for partial auditing, not a sufficient partial-label budget or a result for $\mathcal{P}$ or noisy labels. An independent standard-library producer uses integer recurrences; separate tests use direct combinatorial sums at both large sizes and enumerate all ordered binary observations for 24 small NP problems.

F.4 A supplied bound on the pending count

Proposition 4 (Padding a bounded pending set)**.** *Suppose the primary truthful experiment replaces exactly b pending items by an arbitrary outcome-aware number $B \le b$ on every path, with the same latent-law classes. A valid fixed-b, at-most-k procedure can be implemented using at most k new label requests and retaining its size and worst-case power guarantees. The bound b is supplied in advance.*

Proof. Before exposing any private audit randomness, enlarge the committed pending set to size b by temporarily hiding $b - B$ already known truthful labels. Choose the additional positions by a fixed rule (or independent randomness). The coarsened ledger is itself an admissible outcome-aware fixed-b kernel of the original latent vector; the vector's unconditional law has not changed. Run the fixed-b procedure on this ledger, deliberately withholding the extra stored values from its decisions, including their contribution to J. When it queries a padded position, return the stored truthful value; otherwise request the unknown true label. The transcript supplied to the fixed-b procedure has exactly the law of this legal fixed-b experiment, and only unknown positions incur new requests. No conditioning or cohort selection on B is used. □

This simple reduction gives an operational implementation of a known pathwise bound, not a way to certify the bound from one observed deadline. The supplied query interface accepts exactly $b = 500$ pending positions; padding is a proved preprocessing protocol, not a separately implemented random-B adapter. Since the class of kernels with $B \leq 500$ contains those with $B = 500$, the 57-query minimum remains exact for its minimax problem at the declared design. It is not a lower bound for every realized smaller pending set. If the guarantee $B \leq b$ is unavailable, conditioning on its occurrence can alter the population count law. The stated power promise then does not follow. No .95 power over an enlarged class with unbounded B is proved.

Operationally, recruit the cohort under a justified risk law; record the frozen predictor and deadline; commit the ledger and supplied missing-count bound; use private uniform query randomness; obtain quality-controlled truth; and apply the certified table once. Population assumptions, gold-label reliability, and the supplied bound are external inputs, not outputs of a successful arithmetic replay. No feature-assisted advantage, repeated-audit guarantee, or empirical deployment validation is asserted.

G MISSINGNESS AND UNCERTAIN LABEL RELIABILITY

Unknown label noise can destroy identification without further restrictions (Scott et al., 2013; Liu et al., 2023). Here we specialize that principle to the full selective-audit experiment. The result is an exact criterion for asymptotic power to equal the blind level, not the entire nonblind power frontier. No new general identifiability principle is claimed.

Proposition 5 (Asymmetric reliability criterion)**.** *Fix $0 < a < p < 1$, $0 < \delta < 1$, $0 \leq u < v \leq 1$, and an integer $k \geq 1$. Suppose $b_n/n \to \beta \in (p - a, 1)$, put $d = (p - a)/\beta$, and define*

$$G = \min_{x \in I_\beta} \{v(x + d) - u - (1 - u)x\},$$

where I_β is given in (3)*. Under the conditional envelope* (7)*, if $G \leq 0$ then $V_{n,k}^{\mathcal{C},[u,v]}(\delta) \to \delta$. If $G > 0$, a finitely valid procedure using one uniform pending query gives*

$$\liminf_n V_{n,k}^{\mathcal{C},[u,v]}(\delta) \geq \delta + \min(\delta, 1 - \delta)G > \delta.$$

Both assertions hold also with $\mathcal{P}$ in place of $\mathcal{C}$, and with the original conditional-null/iid-alternative population classes.

Proof. **Response intervals.** At pending error fraction t, a homogeneous admissible channel has false-positive probability $q_0 \in [0, u]$, sensitivity $q_1 \in [v, 1]$, and positive-response mean $(1 - t)q_0 + tq_1$. Its possible means fill

$$A(t) = [vt, u + (1 - u)t].$$

The lower and upper endpoint channels are $(0, v)$ and $(u, 1)$; interpolating them fills the interval. Both endpoints increase with t. Consequently the minimum distance between $A(x)$ and $A(x + d)$ is $[v(x + d) - u - (1 - u)x]_+$.

Admissible full-record hard pair. Fix x strictly inside I_β, and write $y = a - \beta x$. Then y lies strictly inside $[\max(0, p - \beta), \min(a, 1 - \beta)]$, an interval of positive length. Use iid Bernoulli(a) and Bernoulli(p) populations, which are legal under each pair of population classes in the statement. Put $j_n = \lfloor ny \rfloor$. On the branch

$$\max(0, K - b_n) \leq j_n \leq \min(K, n - b_n),$$

reveal j_n uniformly chosen errors and $n - b_n - j_n$ uniformly chosen correct labels; on its complement use any legal mask. The fallback probability vanishes in both worlds. Indeed, interiority gives linear slack, while K/n converges to the respective risk and $b_n/n \to \beta$.

On the feasible branch, the ternary observed record is uniform over all records with those counts, independently of K and the world. Conditional on K and the record, the pending error positions are uniform among all $(K - j_n)$-subsets. Choose a homogeneous memoryless channel separately in each world so that its limiting means $r_0 \in A(x)$ and $r_1 \in A(x + d)$ have minimum separation. They coincide when the intervals overlap. Both channels obey the conditional envelope given the full latent vector, history and queried identity; their response randomness is fresh. Different legal nuisance

channels across hypotheses are permitted by the robust experiment. Neither mask anticipates the private audit seed.

Adaptive transcripts. Generate a length-k without-replacement hidden-truth sequence from the uniform pending composition and pass it through the chosen channel. Given the common observed record and the independent audit seed, feed the next response to whatever distinct pending identity the policy selects. Unused responses are discarded if it stops. Exchangeability of unqueried identities, also after noisy previous responses, makes this sequential simulation correct. Thus the full transcript, including identities, their order and stopping, is a common postprocessing of the response sequence, record and seed. This does not condition on the completed adaptive identity path and then assert independent responses.

For $\theta \in \{a, p\}$ let $t_\theta = (\theta - y)/\beta$. The hidden sequence on the feasible branch is within $k(k-1)/(2b_n)$ in total variation of independent Bernoulli$((K - j_n)/b_n)$ draws. After the channel this error cannot increase. Coupling Bernoulli parameters adds at most

$$k\mathbb{E}\left[|(K - j_n)/b_n - t_\theta| \mid \text{feasible}\right] = o(1).$$

Adding the vanishing fallback mass, the entire transcript is uniformly over policies within $o(1)$ of common postprocessing of k independent Bernoulli(r_θ) responses and the common record/seed. The two response-sequence laws have total variation at most $k|r_1 - r_0|$. Hence every robust-valid procedure has power at most

$$\delta + k[v(x + d) - u - (1 - u)x]_+ + o(1).$$

If the minimizer of G is at an endpoint, approximate it by fixed interior x_m, take $\limsup_n$, and only then let $m \to \infty$. This avoids assuming a negligible fallback at a boundary composition. For $G \le 0$ the upper bound becomes δ, including equality. An independent level-δ rejection coin attains the matching lower bound.

Constructive converse. Put $t_n = n^{3/4}$, $H_n = \lfloor na + t_n \rfloor$, $L_n = \lceil np - t_n \rceil$, and $\eta_n = \exp(-2t_n^2/n)$. For sufficiently large n, $0 \le H_n < L_n \le n$, $b_n \ge 1$, and $\eta_n < \delta$; for all other n never reject. Binomial tail domination under $\mathcal{C}$, or predictable-risk Hoeffding concentration under $\mathcal{P}$, gives $\mathbb{P}_0(K > H_n) \le \eta_n$ and $\mathbb{P}_1(K < L_n) \le \eta_n$. The original classes obey these bounds too. Set $s_n = (\delta - \eta_n)/(1 - \eta_n)$ and $c_n = \min(s_n, 1 - s_n)$. Reject if $J > H_n$; otherwise sample one pending identity uniformly and, after its response W, reject with probability

$$f_n(J, W) = s_n + c_n[W - U_n(J)], \qquad U_n(j) = u + (1 - u)\frac{\min(b_n, H_n - j)}{b_n}.$$

For $j \le H_n$, $U_n(j) \in [0, 1]$; hence $0 \le f_n \le 1$. Conditional on every whole null population and mask with $K \le H_n$, $\mathbb{E}W \le U_n(j)$, even for identity-dependent admissible channels. Unconditional size is therefore at most $\eta_n + (1 - \eta_n)s_n = \delta$.

Under $K \ge L_n$, either $j > H_n$ and power is one, or $\max(0, L_n - b_n) \le j \le \min(H_n, n - b_n)$. In the latter case $H_n - j < b_n$ because $L_n > H_n$ and $L_n - j \le K - j \le b_n$, so the clipping branch is inactive. Since $\mathbb{E}W \ge v(K - j)/b_n \ge v(L_n - j)/b_n$, conditional power is at least

$$s_n + c_n\left[v\frac{L_n - j}{b_n} - u - (1 - u)\frac{H_n - j}{b_n}\right].$$

Every sequence of feasible j has a subsequence with $j/n \to y$ in $[\max(0, p - \beta), \min(a, 1 - \beta)]$; its bracket tends to $v(x + d) - u - (1 - u)x$ for $x = (a - y)/\beta \in I_\beta$. Compactness gives a uniform lower limit G. Multiplying by $\mathbb{P}_1(K \ge L_n) \ge 1 - \eta_n$ proves the claimed bound. Direct rejection cannot reduce it because $\delta + \min(\delta, 1 - \delta)G \le 1$. Extra queries are ignored. □

Symmetric envelope and endpoint calculation. Writing $I_\beta = [x_-, x_+]$, the affine minimum uses x_+ when $u + v \le 1$ and x_- when $u + v \ge 1$. At $(u, v) = (e, 1 - e)$ it is $G = (1 - e)(p - a)/\beta - e$, proving Corollary 4. If $(p - a)(1 - e)/e \ge 1$, the blind region is empty for $\beta < 1$. The result gives neither the full nonblind minimax frontier nor a high-power guarantee for a small budget.

Known-channel contrast at the same nominal reliability. If a common channel (q_0, q_1) is known, with $q_1 > q_0$, replace $U_n(j)$ above by $q_0 + (q_1 - q_0)\min(b_n, H_n - j)/b_n$. The identical finite-validity and concentration argument gives, under $\mathcal{C}$ or $\mathcal{P}$,

$$\liminf_n V_{n,k}^q(\delta) \ge \delta + \min(\delta, 1 - \delta)(q_1 - q_0)(p - a)/\beta > \delta.$$

This is a lower bound, not an optimal known-channel frontier. At $a = .1, p = .2, \beta = .5, e = .2, y = .05$, the hidden fractions are $.1, .3$. Channels $(q_0, q_1) = (1/5, 1)$ and $(2/35, 4/5)$ respectively yield response mean $7/25$ in both worlds. Both meet the $.2/.8$ conditional bounds. The known common symmetric channel $(1/5, 4/5)$ instead gives $.26, .38$. Knowing more about quality can therefore restore positive guaranteed information without improving the nominal truth-conditional accuracy floor.

What the result does not say. The limit fixes k, a, p, δ, u, v while n grows. Its coupling error contains terms depending on k, so it does not establish blindness for growing budgets, finite cohorts, or complete recovery. Repeated labels for the same item, concurrent calibration controls, new native releases and informative proxies can change the experiment. The envelope is conditional on the full latent vector, history and identity, not a marginal accuracy estimate. The positive-gain rule is finite-valid under $\mathcal{P}$, but its power statement is asymptotic; it does not transfer the finite 57-query certificate from $\mathcal{C}$ to $\mathcal{P}$.

Independent exact checks. The supplement checks interval endpoints and channel interpolation on 3,060 rational parameter tuples, including 140 endpoint-only overlap cases, plus 288,288 finite conditional inequalities for the constructive rule. It also checks the displayed common-mean example, the equality boundary $\beta = 2/5$, and a nonblind case $\beta = 3/10$ with $G = 1/15$. These deterministic checks are not a substitute for the asymptotic proof.

G.1 Additional lineage and operational motivation

Exact nuisance calibration is classical (Lehmann and Scheffé, 1950); bounded Bernoulli observations give Bernstein response polynomials (Goyal and Sigman, 2012). Our power-floor analysis specializes these ideas by reducing the growing observed cohort to a degree-k equalizer, not by introducing interpolation or baseline-coin relaxation. NIST Total Recall retrieves preassigned labels (NIST, 2016); content-safety measurement uses scores, weights and quality controls (Dobi et al., 2026). They motivate access to extra labels but do not validate this paper's risk laws, fixed deadline or reliability transport.

H Exact query-budget frontier with uncertain audit reliability

The response-interval criterion in Appendix G can be strengthened to the full above-critical, fixed-query power frontier. The result below turns the separation of query quantity from knowledge of label quality into a general asymptotic design calculation. It retains the conditional envelope, committed mask, and single-deadline experiment; ordinary binomial testing and residual sampling remain its ingredients.

For $0 \le r, s \le 1$, define

$$R_k^+(r, s; \delta) = \begin{cases} \delta, & s \le r, \\ R_k(r, s; \delta), & s > r, \end{cases}$$

where R_k is the randomized binomial NP power used in Theorem 1. The first branch represents overlapping nuisance response classes, not the power of testing two known reversed Bernoulli laws.

Theorem 5 (Robust above-critical frontier)**.** *Fix $0 < a < p < 1$, $0 < \delta < 1$, $0 \le u < v \le 1$, and an integer $k \ge 1$. Suppose $b_n/n \to \beta \in (p - a, 1)$, put $d = (p - a)/\beta$, and retain I_β from* (3)*. Set*

$$U(x) = u + (1 - u)x, \qquad L(x) = v(x + d).$$

Under the conditional response envelope (7)*,*

$$\lim_n V_{n,k}^{\mathcal{C},[u,v]}(\delta) = \min_{x \in I_\beta} R_k^+(U(x), L(x); \delta). \tag{12}$$

The same equality holds under $\mathcal{P}$, and under the original conditional-null/iid-alternative risk classes. The upper bound permits all adaptive distinct-query policies using the full native record. A finitely valid uniform-query test attains the limit when $G = \min_{x \in I_\beta}(L(x) - U(x)) > 0$; when $G \le 0$, a level-δ coin attains the limit δ.

A finite-valid rule and a finite lower certificate. For $b \geq k$ and a null cap H with $\mathbb{P}_0(K > H) \leq \eta_0 < \delta$, put $z = (\delta - \eta_0)/(1 - \eta_0)$. Reject when $J > H$. Otherwise, draw k distinct pending identities uniformly, and use the randomized increasing upper-tail test of their positive-response count at reference size z under $\mathrm{Bin}(k, U_{n,k}(J))$, where

$$U_{n,k}(j) = \min\left\{1, u + (1-u)\frac{\min(b, H-j)}{b-k+1}\right\}. \tag{13}$$

This reference includes a conservative finite-population correction. It is not a claim that responses actually are independent binomial draws.

If an alternative cap satisfies $\mathbb{P}_1(K < A) \leq \eta_1$ and $A > H$, define

$$L_{n,k}(j) = v \max\left\{0, \frac{A-j-k+1}{b-k+1}\right\}, \qquad \mathcal{J} = \{j \in \mathbb{Z} : \max(0, A-b) \leq j \leq \min(H, n-b)\}.$$

Let $B_k(r, s; z)$ be the power under $\mathrm{Bin}(k, s)$ of the increasing randomized upper-tail test calibrated to size z under $\mathrm{Bin}(k, r)$, including when $s < r$. The rule has size at most δ and worst-case power at least

$$(1-\eta_1) \min_{j \in \mathcal{J}} B_k(U_{n,k}(j), L_{n,k}(j); z). \tag{14}$$

The empty-set minimum is defined as one. This is a finite guarantee for a specified conservative rule, not the finite minimax value. Valid C or P concentration bounds may supply the caps; binomial-tail caps must not be transferred from C to P.

Proof. **Finite conditional stochastic bounds.** Condition on the entire latent vector and committed native record with $K \leq H$ and $J = j$. Before the tth identity is selected, at most $\min(b, H-j)$ errors remain among $b - t + 1$ pending items. Uniform sampling and the conditional response envelope give, conditionally on previous identities and responses,

$$\mathbb{P}(W_t = 1 \mid \text{past, latent vector, ledger}) \leq \min\left\{1, u + (1-u)\frac{\min(b, H-j)}{b-t+1}\right\} \leq U_{n,k}(j).$$

The fresh private sampling is not exposed to the response mechanism in advance. This calculation integrates the envelope over the next uniform identity; it does not condition on the complete final identity path. A sequential uniform-variable coupling therefore dominates the response sum by $\mathrm{Bin}(k, U_{n,k}(j))$, even when the responses are causal and dependent. The rejection rule is increasing in that sum, so its conditional size is at most z. The cap event gives total size at most $\eta_0 + (1-\eta_0)z = \delta$.

On $K \geq A$ and $j \leq H$, either $\mathcal{J}$ contains j or the record is infeasible. If the original pending set contains $h = K - j$ errors, after $t-1$ queries at least $h - t + 1$ remain. For $h \leq b$, the ratio $(h-t+1)/(b-t+1)$ is nonincreasing in t. Thus the conditional positive response probability at each $t \leq k$ is at least $L_{n,k}(j)$. Sequential coupling now gives stochastic domination in the reverse direction. Conditional rejection probability is at least the displayed B_k. Records with $j > H$ reject surely; multiplying by the probability of $K \geq A$ proves (14).

Asymptotic achievability outside the blind region. Take $t_n = n^{3/4}$, $H_n = \lfloor na + t_n \rfloor$, $A_n = \lceil np - t_n \rceil$, and $\eta_n = \exp(-2t_n^2/n)$. The C and P concentration arguments of Proposition 5 give both cap failures at most η_n; the original risk classes also obey these bounds. Use the rule whenever $b_n \geq k$, $H_n \leq n$, and $\eta_n < \delta$; otherwise never reject. These operational conditions use only null-side parameters. Eventually $H_n < A_n$ as well, which is needed only for the power analysis and not to decide which test to run.

For any sequence $j_n \in \mathcal{J}_n$, a subsequence satisfies $j_n/n \to y \in [\max(0, p-\beta), \min(a, 1-\beta)]$. For $x = (a-y)/\beta \in I_\beta$, the two references converge to $U(x)$ and $L(x)$; the reference size converges to δ. If $G > 0$, compactness makes their ordering uniformly strict eventually. The increasing binomial test is then NP. Continuity of the finite binomial ROC and compactness of I_β turn (14) into the matching lower limit in (12). This procedure on the nonblind parameter region does not use p or v; these enter its power analysis. The choice of a constant test instead in a blind planning configuration may use p; no universal alternative-agnostic robust attainer over both regions is asserted.

Full-transcript upper bound. Fix a strict-interior $x \in I_\beta$ and $y = a - \beta x$. Apply the common native-ledger construction in Proposition 5 to iid Bernoulli(a) and Bernoulli(p) populations. On the

feasible branch, reveal $j_n = \lfloor ny \rfloor$ uniformly selected errors and $n - b_n - j_n$ uniformly selected correct entries. The complete native record has the same law in both worlds, and the fallback probability vanishes. Conditional on that record and K, pending truths are exchangeable.

If $L(x) > U(x)$, use homogeneous memoryless channels $(u, 1)$ under the null and $(0, v)$ under the alternative. The existing fixed-k coupling simulates any adaptive policy's full transcript as common postprocessing of the record, private seed and k responses. Uniformly over policies, these response sequences are $o(1)$ in total variation from iid Bernoulli($U(x)$) and iid Bernoulli($L(x)$). NP and data processing bound limiting power by $R_k(U(x), L(x); \delta)$. An $o(1)$ change in size is harmless by continuity of the finite-experiment ROC at $0 < \delta < 1$.

If $L(x) \le U(x)$, the admissible response intervals overlap. Choose legal homogeneous channels with the same limiting response mean. The same full-transcript coupling gives limiting power at most δ. Such hypothesis-specific nuisance channels are permitted in the robust experiment. The iid hard populations belong to all three risk-class pairs.

Finally, minimize over interior x and approximate an endpoint minimizer by fixed interior points, taking the cohort limit before the endpoint limit. Continuity gives the closed-interval bound, including equality $G = 0$. If $G \le 0$, the independent size-δ coin supplies the matching lower bound. This completes the proof. □

Corollary 6 (Known-channel comparison)**.** *For a known common channel $0 \le q_0 < q_1 \le 1$, the corresponding above-critical value under the same three risk-class pairs is*

$$\lim_n V^q_{n,k}(\delta) = \min_{x \in I_\beta} R_k\big(q_0 + (q_1 - q_0)x, q_0 + (q_1 - q_0)(x + d); \delta\big).$$

Proof. Use the same argument with the affine known response mean $q_0 + (q_1 - q_0)t$ in both worlds. The finite upper and lower response references replace $u + (1 - u)t$ and vt by that mean, with the same without-replacement bounds. Their limiting separation is $(q_1 - q_0)d > 0$, so the nonblind argument applies throughout I_β. □

Inverting the frontier. For a target $\rho \in (\delta, 1)$ define

$$k_\star(\rho) = \min\left\{k \ge 1 : \min_{x \in I_\beta} R^+_k(U(x), L(x); \delta) \ge \rho\right\},$$

with value infinity if the set is empty. In the blind region no fixed budget attains ρ. If $G > 0$, the set is nonempty: uniform Bernoulli concentration for a midpoint threshold gives both errors vanishing as k grows, uniformly over x, because the separation is at least G. This inverts a family of fixed-k limits. It does not interchange n and k limits or establish a growing-budget theorem.

Numerical evaluation and its scope. For $a = .1, p = .2, \beta = .5, \delta = .05$ and $(u, v) = (.1, .9)$, $I_\beta = [0, .2]$ and $L(x) - U(x) = .08$. For one query, $U(x) \ge .1 > \delta$, so the exact worst power is $\delta(.36/.28) = 9/140$, compared with the earlier valid lower bound .054. The accompanying computation evaluates several reliability bounds and the known common channel with the same $(e, 1 - e)$ parameters. Its budget search and continuous nuisance minimization are numerical, not rational finite certificates. All displayed budgets concern limiting power. They include neither the cost of establishing reliability bounds nor evidence that those bounds transport to real pending cases.

A conservative finite design under P. A separate increasing Hoeffding-threshold rule makes the finite scope concrete. Take $n = 10^6$, $b = 500000$, $k = 1200$, $a = .1$, $p = .2$, $(u, v) = (.1, .9)$, $\delta = .05$, $H = 101858$, $A = 198142$, and $\eta_0 = \eta_1 = .001$. Both integer cap failures require a deviation of at least 1859, so each has probability at most $\exp(-2 \cdot 1859^2/10^6) < .001$ under P as well as C. For every feasible $0 \le j \le H$, the finite response bounds above are unclipped and their gap is the same rational number

$$g = \frac{9}{10}\frac{A - H - k + 1}{b - k + 1} - \frac{1}{10} = \frac{59494}{831335}.$$

Reject directly if $J > H$; otherwise reject when the fraction of positive audit responses exceeds $U_{n,k}(J) + c$, where $c = 8861/250000 = .035444$. Conditional Hoeffding bounds yield

$$\text{size} \le \eta_0 + (1 - \eta_0)e^{-2kc^2} < .05, \qquad \text{power} \ge (1 - \eta_1)\big(1 - e^{-2k(g-c)^2}\big) > .9553.$$

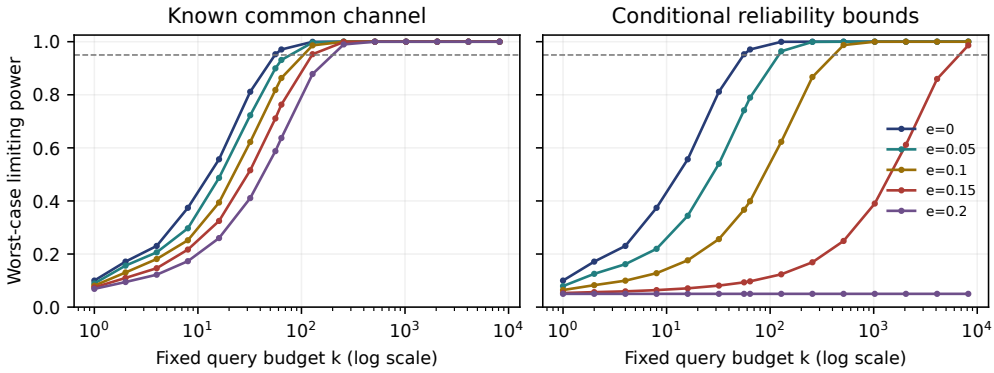


Figure 2: Numerical limiting power at $a = .1, p = .2, \beta = .5, \delta = .05$: known common channels versus conditional reliability bounds. The latter can require many more queries or permit no fixed-budget asymptotic gain. The finite-cohort P certificate is a separate supplied-rule guarantee.

Table 7: Approximate finite minimax values (absolute error below 10^{-6}), $a = .1, p = .2, \delta = .05$, $b = .1n, k = 1$. Bounded reliability uses conditional $u = .1, v = .9$, not average accuracy. Rules are designed at $p = .2$; none is uniformly optimal over p. Exact outward brackets are in Appendices N, P, and Q.

n	Truthful	Known symmetric noise	Bounded reliability
100	.381305	.312183	.299323
1,000	.660514	.428168	.419333

These inequalities allow causal label noise: the finite response bounds hold conditionally on history, so martingale Hoeffding applies. The supplied rule uses a conservative cohort of one million and is not finite minimax; in particular, 1200 is not claimed necessary. The computation certifies the cap, size and power comparisons using only rational arithmetic: $e^{-z} \le (\sum_{i=0}^{96} z^i/i!)^{-1}$ for $z \ge 0$. Its size upper bound is below .049993599 and power lower bound exceeds .9553801256. This finite P example is separate from every asymptotic budget in the numerical comparison and from the 57-query truthful C certificate.

I Additional finite-design evidence

All numerical evidence is deterministic, with rational certificates where stated, not deployment data. Specified-rule power differs from minimax power: the p-agnostic asymptotic attainer need not be efficient at finite n.

Is the weakness intrinsic? For $k = 1$, rational primal/dual witnesses bracket finite minimax power between an iid-null upper and a full-conditional-null lower bound (Table 7). Solver tolerances are not certificates. At $n = 100, b = 10$, rule (5) with the binomial-quantile cap schedule has power .02005, far below the approximately .38130 optimum. At $n = 1000, b = 100$, approximately .66051 is achievable, but even an arbitrary identity-sensitive one-query policy cannot guarantee .67 against all legal masks. Thus a substantial implementation loss and a genuine finite-cohort ceiling coexist. Appendix N gives the reduction and certificates.

The finite cost of imperfect labels. Table 7 separates truthful, known 10% symmetric-noise, and conditionally bounded audit responses. The two bounded-reliability bracket widths are below 1.3×10^{-8}. Thus robust and known channels have the same critical limit, but not the same finite performance. These are exact count-envelope lower and all-policy upper certificates, not finite certificates for $\mathcal{P}$ or $\mathcal{M}$, Monte Carlo estimates, or plug-in Gaussian limits. Appendices Q and P give certificates, additional channels, and exact replay checks.

Calibration can consume size and change feasibility (Appendix R).

A design-point guarantee is not uniform power. At $n = 1000, b = 100$, the one-query truthful rule designed at .2 has power only .00155 at .12, below HT's .06; its known-noise analogue gives .00705. With truthful queries and $b \geq 2$, requiring a floor δ for every $p > a$ and every legal mask forces a one-query rule to the HT envelope, but a vanishing relaxation restores fixed-k critical consistency without weakening size (Appendix J). Thus this is not a general impossibility of useful robust design.

The $b = 500, k = 57$ rule likewise has only .01155 worst-case power at .12. A standard private mixture raises this to .08661 while retaining .95017 at .2 (Appendix F). This remains low power, not a general repair: at .11 it falls to .01695. These negative results motivate checking off-design behavior rather than presenting one planned alternative as universal performance.

When more labels cannot meet the target. At $n = 1000, a = .1, p = .12, \delta = .05$, even observing all 1,000 true labels gives optimal power about .6531752 (Appendix F). This complete-information NP ceiling holds for the original experiment and the $\mathcal{C}$ extension. A 95% requirement at this weak change cannot be met merely by increasing the pending-label budget: the cohort or design target must change. The ceiling does not excuse a rule's much lower power; it separates that procedural loss from insufficient population information. Exact inversion gives a minimum complete-label cohort of **2,645**: at $n = 2644$ the optimal power is below .95, and at 2645 it exceeds .95 (Appendix F.3). This is a necessary cohort size for any partial audit, not a sufficient partial-label budget.

Using the benchmark to plan an audit. Before inspecting labels, fix a, δ, planning target $p_\star > a$, desired power, and the cohort/protocol. Justify both risk classes and label reliability independently; $p_\star$ is not known actual risk. Use certified finite power when available, not only the limiting frontier, and check off-design sensitivity. The 57-query result uses truthful $\mathcal{C}$ classes, not $\mathcal{P}$ or merely mean risk. The bounded-pending extension needs the supplied pathwise bound of Section 2, not conditioning on an observed random B. Nonrejection certifies neither acceptable risk nor the realized cohort's count.

J UNKNOWN-RISK DESIGN AND EXACT POWER FLOORS

Here "unbiased" means power at least the test level against *every* $p > a$ and every legal mask, not unbiased risk estimation.

Proposition 6 (Finite one-query rigidity)**.** *Fix $n \geq b \geq 2$, $a, \delta \in (0, 1)$. A p-agnostic procedure with at most one trusted query that has size at most δ for every mask under iid Bernoulli(a), and power at least δ for every mask and every iid Bernoulli(p), $p > a$, has worst-mask power at most $H(a, p, \delta)$ for every $p > a$. The HT rule attains this bound and is valid under* (1)*.*

The proof (Appendix C) forces exact size under every mask by continuity at a. The resulting one-query response must be globally affine, with slope bounded by (42). Adding a few queries does not remove the *exact* restriction, but allowing a vanishing deficit changes the answer.

Theorem 6 (Exact versus vanishing power floors)**.** *Fix $a, \delta \in (0, 1)$, $k \geq 1$, and $b_n/n \to \beta \in (0, 1)$. If each p-agnostic rule is iid-a valid for every mask and has power at least δ for every $p > a$ and every mask at every n, then for each fixed $p \in (a, 1)$ its worst-mask power has limit superior strictly below one. In contrast, with $b_n = \lfloor n(p - a) \rfloor$, there are p-agnostic rules valid under* (1) *whose worst-mask power at p tends to one and whose power is at least $\delta - \xi_n$ for* every *alternative and mask, where $\xi_n = O((\log n/n)^{k/2})$.*

Appendix K gives an explicit upper bound, the complete all-adaptive-policy reduction, and a finite-valid construction for the second claim. The latter mixes a baseline coin with a gated all-miss audit event. It retains the exact null guarantee; only the power floor is relaxed. It does not attain the full frontier at every missing fraction. The original attaining rule with the stated quantile calibration has a uniform asymptotic floor $\delta - O(\sqrt{\log n/n})$ at fixed k and positive β. Thus exact similarity is a brittle constraint, not a requirement to suffer a fixed positive power deficit. Neither result makes our finite calibration optimal or establishes the smallest necessary relaxation rate.

K EXACT POWER FLOORS, POLYNOMIAL EQUALIZERS, AND A VANISHING RELAXATION

This section proves Theorem 6, including its sensitivity to an exact power-floor requirement. Similarity and bounded-observation response polynomials have classical precedents (Lehmann and Scheffé, 1950; Goyal and Sigman, 2012). The reduction needing proof is why access to $n - b$ selectively revealed native labels still leaves a degree-k response under the additional universal floor. We do not assert a new generic interpolation or Bernoulli-factory principle.

K.1 A GLOBAL FINITE-POPULATION EQUALIZER

Fix $n \geq b \geq k + 1$ and one p-independent procedure with at most k queries. Assume iid-a size at most δ under every mask and power at least δ under every mask and every iid $p > a$. The full ordered null is stronger; only its necessary iid constraint is used for the upper bound.

Average the procedure over a fresh uniform permutation of identities, mapping its physical queries back to the original population. For each permutation, conjugating a legal mask gives a legal mask. In the iid experiment, this averaging preserves both requirements and cannot decrease worst-mask power. It is not asserted to preserve the larger ordered conditional-null class.

Conditional on total count K and native miss count j, the permuted native configuration is uniform with a law independent of K. Its remaining pending miss locations form a uniform $(K - j)$-subset. After t distinct queries and z observed misses, any selected unseen position has miss probability $(K - j - z)/(b - t)$. This is true for an identity-sensitive adaptive policy too. Pad a stopped policy with unused queries to length k. A simulator can generate a without-replacement binary sequence and feed successive bits to the positions requested by the policy. Native configuration, seed, identities, stopping and final decision are common postprocessing of that sequence. Conditional on its total, the sequence's order is uniform independently of K. Hence there are coefficients $f_j(z) \in [0, 1]$, independent of K, p, with conditional expected rejection

$$G_j(K) = \sum_{z=0}^{k} f_j(z) Q_{j,z}(K), \qquad Q_{j,z}(K) = \frac{\binom{K-j}{z}\binom{b-K+j}{k-z}}{\binom{b}{k}}, \quad j \leq K \leq j + b. \tag{15}$$

Here $0 \leq j \leq n - b$. For algebra outside the physical interval, use the polynomial convention $\binom{x}{r} = x(x-1)\cdots(x-r+1)/r!$, including negative or noninteger x; never truncate it to zero.

For any fixed mask kernel, the iid power of a fixed procedure is a polynomial in p, obtained by summing over the finitely many latent binary vectors. The floor as $p \downarrow a$ and the size bound force exact size δ. Compare two kernels choosing feasible $j(K)$ uniformly within label type and differing only at one K. Every $\mathrm{Bin}(n, a)$ mass is positive, so $G_j(K)$ agrees across every feasible j at that K. Each G_j in (15) has degree at most k. Adjacent intervals $[j, j + b]$ and $[j + 1, j + b + 1]$ share $b \geq k + 1$ integers, identifying their polynomials. They cover $0, \ldots, n$, so all G_j are one polynomial G_n of degree at most k, with $0 \leq G_n(K) \leq 1$ on these integer counts. If $b = n$ there is only one interval, and the conclusion is immediate.

The symmetrized procedure consequently has mask-independent power $P_n(p) = \mathbb{E}_{\mathrm{Bin}(n,p)} G_n(K)$, with $P_n(a) = \delta$. Its power dominates the original worst-mask power $m_n(p)$. This is the native-information collapse: the original data contain a growing native cohort, not merely k independent binary observations.

K.2 A FINITE UPPER BOUND AND ITS FIXED-k LIMIT

Fix any legal j and define signed weights

$$w_{j,z}(q) = \mathbb{E}_{\mathrm{Bin}(n,q)} Q_{j,z}(K). \tag{16}$$

They are polynomial expectations over all counts, not hypergeometric laws outside the physical interval. Generalized Vandermonde gives $\sum_z Q_{j,z}(K) = 1$, hence $\sum_z w_{j,z}(q) = 1$. The global identity above gives $P_n(q) = \sum_z f_j(z) w_{j,z}(q)$. Whenever every $w_{j,z}(p) > 0$, set

$$D_{n,j}(a, p) = \max_{0 \leq z \leq k} \frac{|w_{j,z}(a)|}{w_{j,z}(p)}. \tag{17}$$

The nonnegative deficits $1 - f_j(z)$ imply

$$1 - \delta = \sum_z (1 - f_j(z)) w_{j,z}(a)$$
$$\leq D_{n,j}(a,p) \sum_z (1 - f_j(z)) w_{j,z}(p) = D_{n,j}(a,p)(1 - P_n(p)).$$

Thus the explicit finite bound, uniform over all eligible policies, is

$$m_n(p) \leq 1 - \frac{1-\delta}{D_{n,j}(a,p)}. \tag{18}$$

Strict positivity of the alternative weights is a real condition. For $n = 10, b = 3, k = 2, j = 3, p = 1/2$ they are $(5/12, -1/6, 3/4)$, even though np lies inside $(j, j+b)$. In this case (18) cannot be used.

Now fix $0 < a < p < 1$, $0 < \beta < 1$, and a fixed k. Choose any y such that $0 \leq y \leq 1 - \beta$ and $0 < t_p = (p - y)/\beta < 1$, and let $j_n/n \to y$ be legal integer counts. Such a y always exists: project $p - \beta/2$ onto $[0, 1 - \beta]$. Set $t_a = (a - y)/\beta$ and

$$R = \max\left\{ \frac{|t_a|}{t_p}, \frac{|1 - t_a|}{1 - t_p} \right\} > 1. \tag{19}$$

For fixed k, the coefficients of $Q_{j_n,z}(nx)$ converge to those of $B_{k,z}((x - y)/\beta)$, where $B_{k,z}(t) = \binom{k}{z} t^z (1-t)^{k-z}$. Convergence of the finitely many normalized binomial moments then gives $w_{j_n,z}(q) \to B_{k,z}((q - y)/\beta)$ for $q = a, p$. All alternative limits are strictly positive; positivity in (17) therefore holds eventually. As there are only $k + 1$ terms,

$$D_{n,j_n}(a,p) \longrightarrow \max_z \frac{|B_{k,z}(t_a)|}{B_{k,z}(t_p)} = R^k.$$

The endpoint terms $z = 0, k$ give the last equality. Equation (18) therefore proves the first part of Theorem 6, more precisely

$$\limsup_{n \to \infty} m_n(p) \leq 1 - \frac{1-\delta}{R^k} < 1. \tag{20}$$

This argument controls every procedure at each eligible n before taking a limit; it does not exchange a policy supremum with pointwise convergence. The value t_a may be outside $[0, 1]$. Only polynomial extrapolation, with absolute values, is used there. At $a = .1, p = .2, \beta = .1, y = .15$, $R = 3$. The bound is then $1 - .95/3^k$ for $\delta = .05$, not a sharp practical budget bound; the one-query HT bound is stronger in that case.

Why the positivity and overlap restrictions matter. A bounded polynomial alone does not prove (20). The function $H(x) = 1 - (50/49)(x - 4/5)^2$ lies in $[0, 1]$ on $[0, 1]$, has $H(1/10) = 1/2$, stays at least $1/2$ for every $x > 1/10$, and reaches one at $4/5$. Its degree-two Bernstein coefficients on $[3/4, 17/20]$ are $(391/392, 393/392, 391/392)$. The middle coefficient is not a legal response probability. The inherited coefficient constraints, not just the degree, exclude this counterexample. Also, if $b = k < n$, querying every pending label recovers the entire cohort: with $n = k+1$, rejecting exactly at $K = n$ has size a^n and power $p^n > a^n$, with a full-count response of degree $n > k$. Thus the overlap condition cannot generally be weakened to $b = k$.

K.3 A VANISHING RELAXATION RESTORES CRITICAL CONSISTENCY

Fix integers $0 \leq L \leq H \leq n$ and $1 \leq k \leq b$. Choose an upper bound η on $\sup_{Q \in \mathcal{H}_0(a)} \mathbb{P}_Q(K > H)$. The exact binomial tail or a conditional Hoeffding bound can supply it. Let

$$\rho = \frac{\binom{\min(b, H-L)}{k}}{\binom{b}{k}}, \qquad \alpha = \eta + (1 - \eta)\rho. \tag{21}$$

Integer binomial coefficients in this subsection use the ordinary convention: $\binom{m}{k} = 0$ for $0 \leq m < k$. Query k uniformly selected distinct pending labels and let Z count misses. Use the signal

$$A = \{J > H\} \cup \{J \geq L,\ Z = k\}. \tag{22}$$

If $\alpha < \delta$, reject with conditional probability

$$\phi = q + (1-q)\mathbf{1}_A, \qquad q = \frac{\delta - \alpha}{1 - \alpha}. \tag{23}$$

If $b < k$ or $\alpha \geq \delta$, use a size-δ independent coin instead; no division by $1 - \alpha$ is used in that branch.

On a fixed whole population and committed mask with $K \leq H$, direct rejection is impossible. If $J \geq L$, the pending miss count is at most $H - L$, so the all-miss probability is at most ρ. Otherwise the signal is false. Hence $\mathbb{P}_Q(A) \leq \alpha$ uniformly over the conditional null and all masks, and $\mathbb{E}_Q\phi \leq q + (1-q)\alpha = \delta$. Pointwise $\phi \geq q$, so the power floor holds even beyond iid alternatives, with deficit

$$\xi = \frac{(1-\delta)\alpha}{1-\alpha} \quad (\alpha < \delta), \qquad \xi = 0 \quad \text{in the coin branch.} \tag{24}$$

This is exact null validity, not a relaxed type-I error guarantee.

For consistency fix $p > a$ and $b = \lfloor n(p-a) \rfloor$. Take $\sqrt{n} \ll t_n \ll n$, $H = \min(n, \lceil na + t_n \rceil)$ and $L = \max(0, \lfloor na - t_n \rfloor)$. With $\eta_n = \exp(-2t_n^2/n)$, the conditional upper tail is at most η_n (if $H = n$ it is zero). Under iid p, the event $K \geq np - t_n$ has probability at least $1 - \eta_n$. On this event every mask satisfies $J \geq K - b \geq na - t_n$ and therefore $J \geq L$. If $J \leq H$, the number of pending *hits* is

$$b - K + J \leq b - K + H \leq 2t_n + 1.$$

Put $d_n = \min(b, \lceil 2t_n + 1 \rceil)$ and $r_n = \binom{b-d_n}{k} / \binom{b}{k}$. Both direct and nondirect branches give the finite critical power lower bound, when $\alpha < \delta$,

$$\inf_{\kappa} \mathbb{E}_{p,\kappa}\phi \geq q + (1-q)(1-\eta_n)r_n. \tag{25}$$

For fixed k, $r_n \to 1$ and $\alpha_n \to 0$; the coin fallback is eventually unused. Choose $t_n = \sqrt{n(k+1)\log n/2}$ for $n \geq 2$. Then $\eta_n = n^{-(k+1)}$, $H - L \leq 2t_n + 2$, and $\rho_n \leq (\min(b, H-L)/b)^k$ imply $\xi_n = O((\log n/n)^{k/2})$. This proves the second part of the theorem. The rule uses n, b, a, δ, k, not p; only the power analysis invokes the critical relation. The faster floor-deficit rate for larger k is not a consistency rate: $1 - r_n \leq kd_n/b$ only gives the stated construction a conservative $O(\sqrt{\log n/n})$ power-deficit bound.

Why this does not contradict the exact equalizer. For a relaxed universal floor $\delta - \xi$, continuity forces only a size interval $[\delta - \xi, \delta]$ for every mask. Choosing extremal feasible counts separately at each total gives the necessary weighted inequality

$$\sum_{K=0}^{n} \mathbb{P}\{\mathrm{Bin}(n, a) = K\} \big[\max_j G_j(K) - \min_j G_j(K)\big] \leq \xi.$$

At $\xi = 0$ this gives exact equality at every count. With $\xi > 0$, counts typical under a separated alternative have exponentially small null mass; the inequality does not force uniform equality there. This explains the failure of a naive stability argument. It is not a sharp necessary slack rate. The relaxed construction need not attain (4) above the critical missing fraction, and neither construction is claimed finite-sample optimal.

K.4 The original attaining rule is uniformly asymptotically unbiased

The original rule (5), not just the gated construction, admits a useful weaker floor. Let $L \leq H$ be deterministic integer counts, let its reference size be s, and suppose $\mathbb{P}_{\mathrm{Bin}(n,a)}(K < L) \leq \eta_-$. For every iid $p \geq a$, the same lower-tail probability is at most η_- by binomial dominance. On $K \geq L$, a branch with $j > H$ rejects. In any other branch its reference pending miss count $M_0 = \min(b, H - j)$ and true count $M = K - j$ satisfy $(M_0 - M)_+ \leq H - L$.

Couple the two finite populations by changing at most $H - L$ pending labels, and query the same uniform k-subset. The probability of hitting any changed label is at most $k(H-L)/b$. If $M \geq M_0$, monotonicity gives rejection probability at least s directly. Otherwise the coupling gives at least $s - k(H-L)/b$. Uniformly in $p \geq a$ and the mask, the original rule therefore satisfies

$$\mathbb{E}_{p,\kappa}\phi_{\mathrm{original}} \geq (1 - \eta_-) \max\{0, s - k(H-L)/b\}. \tag{26}$$

With a Hoeffding cap, or the specified binomial quantile capped by the same Hoeffding bound, choose $H - na = O(\sqrt{n \log n})$ and $na - L = O(\sqrt{n \log n})$, with both tails $O(n^{-2})$. At fixed k and positive b/n, $s = \delta - O(n^{-2})$, so this floor is $\delta - O(\sqrt{\log n / n})$. This is a uniform floor over alternatives, not merely pointwise convergence of the frontier. It does not erase the finite underperformance in Table 8 or yield a sharp rate.

L A MISSING-LABEL BUDGET ALONE GIVES NO MINIMAX VALUE TO HIT STATUS

Power against a fixed alternative reveal law differs from power minimized across all legal reveal laws. We ask what can be guaranteed when only the number of unavailable labels is controlled, without a trusted query. The following exact cohort result applies to every cohort size, erasure budget, and test level, for a class of exchangeable alternatives. It provides the passive comparator used to characterize the trusted-query intervention.

L.1 EXPERIMENT AND QUANTIFIERS

Fix integers $n \geq 1$ and $0 \leq b \leq n$, a real number $0 < a < 1$, and a deterministic deadline after all n events have been committed. Write $m = n - b$ and $K = \sum_i V_i$. No label is revealed before this deadline. At the deadline a kernel can inspect the realized vector $V \in \{0,1\}^n$ and erase at most b entries. The remaining entries are truthful. All erased labels are revealed at the next tick. This is causal cohort-end release, with eventual full revelation; it is not a bound of one tick on delay from individual issuance.

Full status observes $S \in \{\perp, 0, 1\}^n$, including identities. Miss-only observation is $X_i = \mathbf{1}\{S_i = 1\}$ plus n, b and the deterministic ledger. It receives no hit notification or erasure indicator. Null laws obey $\mathbb{E}[V_i \mid V_{<i}] \leq a$ in event order. Null and alternative kernels range over the *same* erasure-budget class, but need not be the same kernel. All fitted objects are conditioned on and fixed.

The alternative latent law is exchangeable. Let $P_k = \Pr(K = k) > 0$ and $Q_k = \binom{n}{k} a^k (1-a)^{n-k}$, and assume

$$\frac{P_k}{Q_k} \text{ is nondecreasing in } k. \tag{27}$$

This includes iid Bernoulli(p), $a < p < 1$, and mixtures of iid Bernoulli laws with parameter in $[a, 1)$ having positive count probabilities. The latter need not have independent latent errors. The theorem does not cover arbitrary nonexchangeable alternatives.

For observation Z, define

$$\mathcal{R}_Z(\delta) = \sup_{\phi:\, \sup_{Q \in \mathcal{H}_0,\, \kappa \in \mathcal{K}_b} \mathbb{E}_{Q,\kappa}\phi(Z) \leq \delta} \inf_{\kappa \in \mathcal{K}_b} \mathbb{E}_{P,\kappa}\phi(Z), \qquad 0 \leq \phi \leq 1.$$

The alternative reveal law is minimized here. This quantifier is different from power against a fixed alternative channel. Write $\mathcal{R}_{\text{count}}$ for the same optimization restricted to nondecreasing functions of $\sum_i X_i$.

L.2 EXACT MAXIMIN ROC

Theorem 7 (Cohort budget boundary)**.** *Under the preceding experiment and* (27)*, for every* $0 \leq \delta \leq 1$*,*

$$\mathcal{R}_S(\delta) = \mathcal{R}_X(\delta) = \mathcal{R}_{\text{count}}(\delta).$$

The common value is the largest vertical coordinate at horizontal coordinate δ *in the convex hull of*

$$(0,0), \quad (1,1), \quad \left(\Pr_Q(K \geq k), \Pr_P(K \geq k + b) \right), \qquad 1 \leq k \leq m. \tag{28}$$

It is attained by a randomized nondecreasing function of the revealed miss count. For $b = n$ *the value is* δ*. One iid Bernoulli(*a*) null reveal kernel, independent of* δ*, and one alternative reveal kernel attain the ordinary Neyman–Pearson upper bound simultaneously for the entire ROC.*

The randomization in (28) may mix thresholds; it is not in general a single hard threshold with only one count randomized. The endpoint $(1, 1)$ is essential: the formal $k = 0$ expression with $k + b$ is not the always-reject test.

Proof. First suppose $m > 0$. Let $J_+ = (K_a \wedge m)$ and $J_- = (K_a - b)_+$ for $K_a \sim Q$, with pmfs q^+ and q^- on $0{:}m$. Let $J_P = (K_P - b)_+$, $K_P \sim P$, with pmf r. All entries of these three pmfs are positive. Pooling the first $b + 1$ adjacent bins of an ordered likelihood ratio preserves its order, so r_j/q_j^- is nondecreasing.

Apply the pool-adjacent-violators construction to the ratios r_j/q_j^+, merging adjacent blocks whose ratios decrease. For each final contiguous block B set

$$\theta_B = \frac{r(B)}{q^+(B)}, \qquad q_j^* = \frac{r_j}{\theta_B} \quad (j \in B).$$

Thus q^* is a pmf, and $r_j/q_j^* = \theta_B$ is nondecreasing. In a pooled block, every prefix has its original cumulative r mass at least θ_B times its cumulative q^+ mass. This follows inductively when two decreasing-ratio blocks merge, or from the lower convex minorant of the cumulative (q^+, r) diagram. Consequently

$$F_{q^+}(j) \le F_{q^*}(j). \tag{29}$$

At each block endpoint $F_{q^*} = F_{q^+} \le F_{q^-}$. Inside a block, $q_j^*/q_j^- = (r_j/q_j^-)/\theta_B$ is nondecreasing, so the increments $q_j^* - q_j^-$ change sign at most once, from negative to positive. The cumulative difference attains its maximum at a block endpoint. Hence

$$F_{q^+}(j) \le F_{q^*}(j) \le F_{q^-}(j) \quad \text{for every } j. \tag{30}$$

Use one uniform random variable to couple K_a and $J^* \sim q^*$ through their quantiles. The stochastic bounds in (30) imply

$$\max(0, K_a - b) \le J^* \le \min(K_a, m)$$

almost surely. Under iid null losses, conditional on $(K_a, J^*) = (k, j)$, uniformly choose j of the k misses and $m - j$ of the $n - k$ hits to reveal. Exactly b entries are erased. This defines a legal randomized kernel conditional on the realized K_a, not a change in the latent null law.

Under the exchangeable alternative, reveal $(K_P - b)_+$ misses and enough hits to reveal m labels in total, choosing uniformly within each label type. This erases as many misses as possible, then hits if necessary. In either construction, conditional on the revealed miss count j, the full status vector is uniform over its $\binom{n}{j}\binom{n-j}{m-j}$ configurations. Indeed for fixed (k, j) a particular such status has probability

$$\frac{\binom{b}{k-j}}{\binom{n}{k}\binom{k}{j}\binom{n-k}{m-j}} = \frac{1}{\binom{n}{j}\binom{n-j}{m-j}},$$

independent of k. Thus the full-status likelihood ratio for these two worlds is exactly $r_j/q_j^* = \theta_B$.

The Neyman–Pearson test at any size δ is nondecreasing in j. At a boundary ratio, randomize with the same probability on every tied block. Its test function f is constant on each pooled block. Since q^* and q^+ have equal block masses, its size under both is δ. Extend $f(j) = f(m)$ for $j > m$. Under any conditional-risk null, the full latent count is stochastically dominated by K_a; the revealed miss count is no larger than that latent count. Therefore

$$\mathbb{E}_{Q,\kappa} f(\textstyle\sum_i X_i) \le \mathbb{E}_Q f(K) \le \mathbb{E} f(K_a) = \mathbb{E}_{q^+} f = \delta.$$

No permutation symmetry of the conditional-risk null class is assumed. For every alternative mask with at most b erasures, the revealed count is at least $(K_P - b)_+$, so its power is at least $\mathbb{E}_r f$. The constructed two worlds give exactly that power and supply an upper bound for *every* full-status robust test. This proves all three equalities and gives a common least-favorable pair for every δ.

Finally, every nondecreasing f on $0{:}m$ is a mixture of the always-reject test, upper count thresholds, and the never-reject test. Their size/power pairs are precisely (28); the maximizing curve is its upper concave hull. If $m = 0$, erasing every label in both worlds leaves no information and an independent size-δ coin attains power δ. □

L.3 A SEPARATE SET-BASED PROOF OF THE UPPER BOUND

For an independent route, restrict the null to iid Bernoulli(a). Only this subproblem is permutation invariant. Symmetrizing a full-status test does not increase its worst null size or decrease its worst alternative power. It suffices to consider exactly b erasures: the adversary can use that budget, and any test can handle fewer erasures by randomly erasing extra labels first. The visible miss count j is then sufficient for the symmetric test. Conditional on latent count k, it ranges over $I_k = [\max(0, k-b), \min(k, m)] \cap \mathbb{Z}$.

For a level set $A \subseteq \{0, \dots, m\}$, put $D(A) = \{k : I_k \cap A \neq \varnothing\}$ and $E(A) = \{k : I_k \subseteq A\}$. For any nonnegative slope λ the support objective is $P(E(A)) - \lambda Q(D(A))$. A component $[l, u]$ of A has dilation $[l, u+b]$ and erosion $[l+b, u]$, with the erosion's left endpoint replaced by zero if $l = 0$, and its right endpoint by n if $u = m$. An interval with left endpoint above its right endpoint is empty.

Fill any gap of at most b missing ranks between components: dilation is unchanged and erosion can only grow. Remaining dilations are disjoint. Delete components with nonpositive support objective. If none remain, the empty set is at least as good. Otherwise the first remaining component $[l, u]$ has positive score. For $h = u + b$, monotone P_k/Q_k and $E([l, u]) \subseteq D([l, u]) \subseteq [0, h]$ imply $P_h/Q_h > \lambda$. Extend the remaining set to the upper ray $[l, m]$. Every newly added dilation index lies above h, and also belongs to the newly added erosion. Its contribution is at least $P_k - \lambda Q_k \geq 0$; any other added erosion mass is nonnegative. Thus an upper ray, the full set, or the empty set is optimal for every supporting slope.

For a randomized test, layer-cake identities express the maximum and minimum over I_k as integrals of the corresponding level-set indicators. The same support bound applies. It yields the hull (28). Count threshold mixtures attain that hull under the entire conditional-risk null, completing the independent upper-bound argument.

L.4 CONSEQUENCES, EVIDENCE, AND LIMITS

The result separates *possible usefulness* of hit status under a fixed reveal law from *guaranteed usefulness* under a budget-only service promise. A backlog cap alone cannot guarantee a full-status power gain in this cohort model. Such a gain requires additional structure that excludes the constructed least-favorable reveal kernels, a different alternative class, or a different observational experiment. It does not say that hits should be discarded in actual deployments.

This is exact fixed-deadline maximin testing, including randomized tests. A rule testing once at that pre-fixed deadline has a global ever-alarm budget δ. Repeating independently calibrated deadline tests does *not* inherit that budget. Neither proof yields a single online procedure optimal at every deadline, general interleaved-delay rates, unknown-onset uniform optimality, or a universal no-value theorem for hit information. The alternative count law is known to the oracle test.

Small floating-point diagnostics compare full ternary, miss-only, and count tests under the genuine conditional-risk null; they are not the proof. A separate rational-arithmetic checker verifies the constructive null, its marginal and bounded-displacement constraints, blockwise test size and power, and the set-based support argument on finite cases. Neither numerical agreement nor internal proof review establishes literature priority.

Sample-aware subtractive contamination with explicit missing positions is already studied by Vallinayagam et al. (2026), including Section 7. The indexed ledger and adversarial erasure model are therefore not new. The result here is the exact cohort ROC collapse and its extension to the event-order conditional-risk null, with a common attainable null kernel. The PAV, stochastic-ordering, and Neyman–Pearson tools are standard. This model-specific boundary delimits what passive label retention can guarantee before an independently randomized query changes the experiment.

M CAN UNUSED SIZE EXPLAIN THE FINITE-SAMPLE WEAKNESS?

Consider the specified one-query rule with deterministic cap H and residual size $s \in (0, 1)$. Write $g(K, j)$ for its rejection probability on a population with total K and feasible native miss count j,

and set

$$U(K) = \max_{\max(0,K-b)\le j\le \min(K,n-b)} g(K,j), \qquad S = \mathbb{E}_{\mathrm{Bin}(n,a)} U(K).$$

This S is its exact worst-case size over the full conditional null, not just an iid proxy. To see this, its rejection table is nondecreasing in the native count and the queried binary outcome. Increasing the native count lowers the hypergeometric reference miss fraction; its randomized upper-tail test at fixed s is then pointwise no less rejecting. For any fixed mask and queried index, flipping one latent zero to one either increases the native count or leaves it fixed and increases an audited outcome. Rejection cannot decrease. Taking the maximum over masks shows that $U(K)$ is nondecreasing in K.

Under the conditional null, K is stochastically dominated by $\mathrm{Bin}(n,a)$, giving size at most S. Conversely, iid Bernoulli(a) and a kernel choosing a maximizing feasible j at each total attain S. Thus, if $S \le \delta < 1$, the standard correction

$$q = \frac{\delta - S}{1 - S}, \qquad \phi^+ = q + (1-q)\phi$$

has size exactly δ and pointwise dominates the original rule. It uses no alternative parameter and no extra label queries. If the original worst-mask power is P, the corrected power is $q + (1-q)P$. Since that power is bounded below by the original attaining power and above by $V_{n,1}(\delta)$, it retains the same limiting frontier by squeezing. This is ordinary use of unused test size, not a new testing method.

We evaluated only the ten existing configurations in Table 8. At $n = 100$, the original worst sizes are about .04407053 for $b = 10$ and .02176029 for $b = 50$. Their corrected powers are .02612726 and .04611719. At $n = 1000$ the corrected powers are .11596536 and .08321562; at $n = 10^6$ they are .97988977 and .09891008. The critical-regime gain is small, and every displayed above-critical value remains below the HT value .10. These results rule out the claim that simply spending the remaining size solves the reported weakness; they do not rule out better calibration or a different finite-sample test.

For exact small cases, the checker verifies 672 envelope/monotonicity/size configurations and 54 corrected tests. The large-count calculation uses the same constant-size set of piecewise-monotone candidate native counts as the original power calculation, now taking a maximum. It is a chunked linear scan in K, evaluated in float64. This is not a certified numerical decision engine: exact size S proves validity algebraically, whereas an implementation needs exact arithmetic or a certified upper bound $c \in [S, \delta]$ and $q = (\delta - c)/(1 - c)$ for guaranteed size. No data simulation was used.

We do not instead claim that a largest feasible residual size always exists. For the literal upper-tail implementation, at reference miss fraction one, $s \uparrow 1$ still accepts a queried hit, but $s = 1$ rejects that zero-null-mass outcome too. The endpoint can jump. For $H = b = n$ and $a < \delta$, every $s < 1$ has size $sa < \delta$, whereas $s = 1$ has size one. The baseline correction avoids that endpoint problem. Neither procedure is asserted finite-minimax optimal or a cure for the paper's unresolved practical-performance limitation.

N CERTIFIED FINITE ONE-QUERY BRACKETS AND DESIGN-POINT SENSITIVITY

The asymptotic frontier does not identify the finite optimum. This section provides computable lower and upper bounds under exactly the same experiment, for $k = 1$. The optimization and weak-duality principle are standard tools; the scientific role is to distinguish procedural weakness from an unavoidable finite limitation, without presuming a general equality of two relaxations.

N.1 AN UPPER BOUND THAT PERMITS EVERY POLICY

Write $q_r(K) = \mathbb{P}\{\mathrm{Bin}(n,r) = K\}$ and $\mathcal{J}_K = \{\max(0, K-b), \dots, \min(K, n-b)\}$. For a response table $f_{jz} \in [0,1]$, define

$$g_{Kj} = (1 - (K-j)/b) f_{j0} + ((K-j)/b) f_{j1}, \qquad j \in \mathcal{J}_K.$$

The following finite linear program relaxes the full ordered conditional null to its iid-a member:

$$\begin{aligned} \text{maximize} \quad & \sum_{K=0}^{n} q_p(K) l_K \\ \text{subject to} \quad & g_{Kj} \le u_K, \quad l_K \le g_{Kj} \qquad (0 \le K \le n,\ j \in \mathcal{J}_K), \\ & \sum_{K=0}^{n} q_a(K) u_K \le \delta, \quad 0 \le f_{jz}, u_K, l_K \le 1. \end{aligned} \tag{31}$$

Its value upper-bounds the original finite minimax value. To see why this includes identity-sensitive and randomized query policies, symmetrize only in the iid subproblem. A fresh uniform permutation after mask commitment conjugates every legal mask to another legal mask and cannot lower worst-mask power. Conditional on the full native configuration, the unobserved permutation assigns the remaining misses uniformly among the b pending positions. Any chosen unseen identity therefore has miss probability $(K - j)/b$. A no-query branch can be padded with an ignored query. Averaging the final decision gives f_{jz} independent of K, just as in Appendix K. The full native configuration has a common conditional law given j. A mask can realize each feasible j separately at each K, so its null maximum and alternative minimum are respectively $\max_j g_{Kj}$ and $\min_j g_{Kj}$. Every full-null-valid policy consequently has worst-mask power no larger than (31). This argument never asserts permutation invariance of the ordered null.

N.2 A LOWER BOUND UNDER THE FULL CONDITIONAL NULL

Add the response inequalities

$$f_{j0} \le f_{j1}, \qquad f_{jz} \le f_{j+1,z} \quad (z = 0, 1).$$

For these monotone tables, $U(K) = \max_{j \in \mathcal{J}_K} g_{Kj}$ is nondecreasing. One direct coupling fixes a population, mask, and random queried identity and flips one latent zero to one. A native flip increases j; a pending flip either leaves the audit unchanged or increases z. The rejection probability cannot decrease. Relabeling a population of size $K + 1$ then transfers the maximizing size-K mask, proving the claim. Since conditional-risk nulls admit a coordinatewise coupling below iid-a Bernoullis, their count is stochastically bounded by $\mathrm{Bin}(n, a)$. Thus the *exact* worst-null size of this table is

$$S = \sum_{K} q_a(K) U(K).$$

The iid-a law and countwise maximizing masks attain it. The worst iid-p power is $P = \sum_K q_p(K) \min_{j \in \mathcal{J}_K} g_{Kj}$. This proves feasibility for the full null, not merely a numerical iid check.

For solver-proposed rows we first clip to $[0, 1]$, round to denominator 10^9, take cumulative maxima in each row, and raise f_{j1} to at least f_{j0} to enforce all displayed inequalities exactly. With their rational S, P, if $S > \delta$ multiply the table by δ/S. If $S \le \delta$, replace it by $q + (1 - q)f$, $q = (\delta - S)/(1 - S)$. Both transformations yield exact full-null size δ and a directly computable rational lower bound. They require no unverified floating-point size estimate. They are ordinary calibration, not new tests. Optimizing a table at p does not assert optimality elsewhere.

N.3 A RATIONAL UPPER CERTIFICATE FROM ARBITRARY DUAL PROPOSALS

Write (31) as maximizing $c^T x$ over $Ax \le d$, $0 \le x \le 1$. Every nonnegative rational vector y, regardless of whether a numerical optimizer calls it dual feasible, satisfies

$$c^T x \le y^T d + \sum_i \max\{0, c_i - (A^T y)_i\}. \tag{32}$$

Indeed $c^T x = y^T Ax + (c - A^T y)^T x$, and each box-constrained coordinate is bounded by the positive part of its coefficient. Quantize nonnegative dual proposals to denominator 10^{12} and evaluate every coefficient exactly. For rational a, p, all binomial masses and all coefficients of g_{Kj} are rational; integer arithmetic gives a valid upper bound. Bounds checking precedes integer conversion and accumulation in the implementation. Round the final lower downward and upper upward, never to

nearest. Saved response/dual proposal arrays and their full feasible-count geometry allow certificate replay without calling an optimizer. The replay repeats exact quantization and integer evaluation; solver tolerance plays no role.

We do not prove that monotone tables always attain the unrestricted program, or that iid-a is least favourable for arbitrary tables. Neither statement is needed: an exactly feasible lower witness and an exact upper certificate bracket the desired value. At the four reported design points they almost meet. The narrower exact critical intervals are

$$\begin{aligned} n = 100, b = 10: &\quad [.381304714391,\ .381304716993], \\ n = 1000, b = 100: &\quad [.660513866770,\ .660513886863]. \end{aligned}$$

At $b = .5n$, the exact HT lower bound is .1; the certified upper bounds are .100000000123 and .100000000352 at $n = 100, 1000$, respectively. The small positive gaps are certificate gaps, not proof that the value is strictly larger than .1.

N.4 Why the improved rule is not an unknown-risk solution

The critical designs use $p_\star = .2 = a + b/n$. That choice could be specified from the null and missing budget, but it cannot be confused with a guarantee at every actual $p > a$. Fixing the released rows and calibration, exact integer calculations give the following worst-mask powers:

n	$p = .105$	$p = .12$	$p = .2$	$p = .4$
100	.004755	.014283	.381305	.998964
1,000	.000299	.001548	.660514	.999999970

These powers are rounded summaries of rational evaluations, not new Monte Carlo trials or reoptimized alternatives. In particular, at $p = .12$ the larger cohort's design-point rule is worse than both HT (.06) and the independent coin (.05). It lacks the exact universal power floor of Theorem 6. This is a substantive limitation rather than a contradiction. The pointwise near-optimal table is evidence about what is possible at a specified alternative, not a uniformly preferable replacement for the p-agnostic asymptotic rule.

O Additional verification and adverse finite-rule results

We verify the mathematics with exact finite calculations, rather than substituting simulation volume for the general upper bound. The rational checker covers 168 population/audit/level configurations, 671 good-population calibration checks, and 20 enumerations of full-history conditional-null vertices. An independent check of the proof ingredients covers 12,285 exact upper-tail-versus-NP comparisons and 231 uniform full-configuration cases (6,664 paths). These are internal computational checks, not external peer review, empirical sample sizes, or proof-assistant certification. Additional exact checks cover 64 polynomial-overlap cases, 18 full-recovery exceptions, 280 signed-moment identities, and 249 relaxed-rule calibrations.

The finite implementation uses the exact binomial upper quantile with $\eta_n = \min(\delta/2, (n+1)^{-2})$, and its actual upper-tail mass in (5). Table 8 reports the worst-channel power of this specified one-query rule, computed by integrating its exact piecewise population formula in floating point. The result is *not* the optimum over all finite-n procedures. A separate conservative concentration-envelope lower bound is also retained in the reproducibility record, including zeros.

This adverse finite-sample behavior prevents interpreting the theorem as an off-the-shelf small-budget performance guarantee. The confidence band is an existence-proof device with a conservative finite implementation; it was not tuned to p. We tested whether unused type-I error budget alone explains the weakness. A standard p-agnostic baseline-coin correction spends the exact remaining size and improves every rule in the table. But at $n = 100$ the two powers become only .02613 and .04612; both remain below the .05 coin, and every displayed above-critical case remains below HT. Appendix M proves the correction and reports its scope. Thus spending unused size alone is insufficient.

Table 8: Deterministic power calculations, $a = .1, p = .2, \delta = .05, k = 1$. The implementation is conservative and can be weaker than a size-δ coin at small n. The last row is a proved minimax limit, not finite data.

Total cohort n	Critical $b = .1n$	Above critical $b = .5n$
100	.02005	.01776
1,000	.11421	.06757
10,000	.30979	.08532
100,000	.81573	.09426
1,000,000	.97989	.09793
HT rule, every n	.10000	.10000
Minimax limit	1	.10000

P THE JOINT CRITICAL EXPERIMENT WITH IMPERFECT AUDIT LABELS

We prove Theorem 2. Fix interior a, p, δ, a fixed hard budget $k \geq 1$, $b_n = n(p-a) + s\sqrt{n} + o(\sqrt{n})$, and a known channel $0 \leq q_0 \leq q_1 \leq 1$. There are no features, outcome proxies, repeated queries of the same item, or additional native releases. Query noise is generated freshly given the past, latent vector, and current identity. Write $Q_i = \mathrm{Bern}(q_i)^{\otimes k}$ and $R = R_k^q$. The count is sufficient for testing Q_0 against Q_1, including their singular endpoints. This appendix uses the ordinary randomized NP convention $R(x) = \sup_{0 \leq f \leq 1,\, Q_0 f \leq x} Q_1 f$.

P.1 ALL-RECORD UPPER BOUND

Fix any procedure valid under the full conditional null. Pad early stopping with ignored distinct queries up to k. For a necessary upper bound it suffices to use iid-a and iid-p latent laws. Conditional on total K and a chosen feasible native error count j, let a kernel choose the native configuration uniformly within the required types. Given j, the complete native ledger has a law independent of K; pending true labels are uniformly located subject to their total $K - j$.

Adaptively choosing identities cannot break exchangeability of the remaining pending labels. Generate an ordered sample without replacement, pass each label through fresh channel randomness, and reconstruct the chosen identities from the ledger, response sequence, and policy seed. These reconstructions are common postprocessing. Averaging over the common native ledger and seed gives a response $f_n(j, w_1, \ldots, w_k) \in [0, 1]$ independent of K. This reduction is only in the iid necessary-null subproblem; the ordered conditional-null class is not assumed to be permutation invariant.

Define $h_n(j) = Q_0 f_n(j, \cdot)$. The all-true-error audit law satisfies

$$Q_1 f_n(j, \cdot) \leq R(h_n(j)). \tag{33}$$

Choose $M_n \to \infty$, $M_n = o(\sqrt{n})$, for example $n^{1/8}$, and call j central when $|j - na| \leq M_n\sqrt{n}$. Define

$$H_n(x) = \max\{h_n(j) : j \text{ central},\ j \leq \lfloor na + x\sqrt{n} \rfloor\},$$

using zero for an empty set. Extend it constantly above the central window. It is nondecreasing and bounded by one.

For central null totals K, every central $j \leq K$ is feasible for large n: $a > 0$, $1 - p > 0$, and $b_n \sim n(p-a)$ give linear margins. There are at most $2M_n\sqrt{n} + O(1)$ hidden errors. The probability that any of k distinct queries encounters such an error is at most $k(K-j)/b_n$, uniformly over adaptive policies. Coupling on its complement bounds the total-variation distance from Q_0 by $O(kM_n/\sqrt{n}) = o(1)$. The null kernel chooses a maximizing j; a legal fallback handles noncentral totals. Since their probability vanishes,

$$\mathbb{E}H_n\big((\mathrm{Bin}(n, a) - na)/\sqrt{n}\big) \leq \delta + o(1). \tag{34}$$

Under the alternative, hide b_n true errors whenever possible and use a legal fallback otherwise. The exceptional probability vanishes since $p - (p - a) = a > 0$. On the feasible component, $j = K - b_n$

and all queried true labels are one, so their responses have law Q_1. On central j, (33) gives the bound $R(H_n((j-na)/\sqrt{n}))$. The worst-channel power is therefore at most

$$\mathbb{E}R\big(H_n\big((\mathrm{Bin}(n,p)-b_n-na)/\sqrt{n}\big)\big)+o(1). \tag{35}$$

The normalized binomial CDFs converge uniformly to those of X_0 and X_1. Integration against a bounded nondecreasing function is controlled by the uniform CDF error times its total variation. This applies uniformly to the changing functions H_n and $R\circ H_n$, both with variation at most one; weak convergence against arbitrary changing tests is not being assumed. Writing the right side of (6) as $F(\delta)$, (34)–(35) imply an upper bound $F(\delta+o(1))+o(1)$. Concavity of R and mixing feasible h show that F is bounded, nondecreasing, and concave. It is continuous at interior δ, which proves the limsup assertion.

P.2 An exactly valid finite construction

Fix any nondecreasing h feasible for $F(\delta)$. Select nested randomized upper-tail NP tests $\psi_x(z)$ with exact null size x and power $R(x)$. They can be chosen nondecreasing in both x and the noisy response count z. For $q_0=q_1$ the constant response x suffices. For $q_0=0<q_1$, use x at zero and one at positive counts. For $q_0<q_1=1$, use the upper-tail family with randomization at its threshold. Thus degenerate laws do not require an unspoken continuity assumption at zero size.

Sample k distinct pending identities uniformly with fresh randomness. At native count j and noisy count z, set

$$\phi_n(j,z)=\begin{cases}0, & j<na-M_n\sqrt{n},\\ \psi_{h((j-na)/\sqrt{n})}(z), & |j-na|\le M_n\sqrt{n},\\ 1, & j>na+M_n\sqrt{n}.\end{cases}$$

The lower gate is literal zero, not a size-zero NP test: the latter may have positive power on null-zero support. The table is nondecreasing in j and z. For feasible j at total K, let $g_n(K,j)$ be its exact expected rejection probability under uniform sampling and the channel, and set

$$U_n(K)=\max_j g_n(K,j),\qquad L_n(K)=\min_j g_n(K,j).$$

The maxima and minima range over $\max(0,K-b_n)\le j\le\min(K,n-b_n)$.

U_n is nondecreasing in K. Fix a maximizing population and mask at K, retain the same mask and sampled identities, and flip one true zero to one. A native flip increases j. A queried pending flip changes its response from $\mathrm{Bern}(q_0)$ to $\mathrm{Bern}(q_1)$, which increases under a shared uniform coupling. An unqueried flip leaves the response unchanged. Maximizing at $K+1$ cannot decrease this value.

Sequential uniform coupling of the conditional Bernoulli risks gives $K\preceq\mathrm{Bin}(n,a)$ under every allowed null. Thus the *exact integrated* worst-null size and worst iid-p power are

$$S_n=\mathbb{E}U_n(\mathrm{Bin}(n,a)),\qquad P_n=\mathbb{E}L_n(\mathrm{Bin}(n,p)).$$

The null equality follows because iid-a and countwise maximizing masks are allowed. This does not require size δ separately at each K. Multiplying the response table by $c_n=\min(1,\delta/S_n)$, with $c_n=1$ if $S_n=0$, gives exact finite validity under the full conditional null. These are mathematical expectations, not unverified floating-point calibrations. The construction is an existence result; no efficient exact representation of an arbitrary optimal h is claimed.

P.3 Lower convergence and consistency consequences

For central null K, every nonzero feasible response has central $j\le K$. Its query law is uniformly $o(1)$ from Q_0 by the preceding coupling. Monotonicity of h and feasibility of $j=K$, which hides only true zeros, therefore give

$$U_n(K)=h((K-na)/\sqrt{n})+o(1)$$

uniformly on central totals. For the alternative, when the minimum feasible native count $j_{\min}=K-b_n$ is central, every feasible central j is at least $j_{\min}$. The pending pool has at most $2M_n\sqrt{n}+O(1)$

true zeros; its noisy law is uniformly $o(1)$ from Q_1. The endpoint $j_{\min}$ has exact law Q_1, while the upper gate returns one. Hence

$$L_n(K) = R(h((K - b_n - na)/\sqrt{n})) + o(1).$$

Noncentral probabilities vanish. Uniform CDF convergence for the monotone functions $h, R \circ h$ gives $S_n \to \mathbb{E}h(X_0) \le \delta$ and $P_n \to \mathbb{E}R(h(X_1))$. Consequently $c_n \to 1$. Taking fixed, arbitrarily close-to-optimal h proves the liminf assertion without assuming that the supremum is attained. Such h may depend on p, s and the channel.

For $q_1 = 1$, $q_0 > 0$, write $r = q_0^k$. Since $R(x) = \min(x/r, 1)$, clipping h at r loses no power and lowers size. Setting $g = h/r$ shows that $F(\delta)$ is the passive monotone Gaussian ROC at $\min(\delta/r, 1)$. The two Gaussian laws are equivalent, and their passive ROC is strictly below one at every level below one. Thus consistency holds iff $r \le \delta$. Perfect queries have $r = 0$ and are handled directly by the gated construction.

If $q_0 < q_1 < 1$, Q_0 is absolutely continuous with respect to Q_1, so $R(x) < 1$ for every $x < 1$. If a sequence of feasible h had $\mathbb{E}R(h(X_1)) \to 1$, then $h(X_1) \to 1$ in probability. Equivalence of the two fixed Gaussian laws transfers convergence in probability to X_0; boundedness gives $\mathbb{E}h(X_0) \to 1$, contradicting $\delta < 1$. This also covers $q_0 = 0 < q_1 < 1$. An uninformative channel reduces directly to the passive ROC. These statements concern fixed k and known channel parameters, not a universal finite-data impossibility for noisy audits.

P.4 Reduction to classical product testing

The variational expression in Theorem 2 is not a new composition primitive. We give its exact reduction, including singular side-channel support. Let P_0, P_1 have positive continuous densities and continuous strictly increasing CDFs F_0, F_1 on $\mathbb{R}$. Let B_0, B_1 be any two laws on a finite set, with randomized NP power R. Set $A(u) = F_0(F_1^{-1}(u))$, with endpoints zero and one; let C be its least concave majorant on $[0, 1]$; and define

$$F_0^*(y) = C(F_1(y)).$$

Then, for every interior level δ,

$$\sup_{\substack{h \text{ nondecreasing},\ 0 \le h \le 1 \\ P_0 h \le \delta}} P_1 R(h) = \mathrm{ROC}_{P_0^* \otimes B_0,\, P_1 \otimes B_1}(\delta). \tag{36}$$

The repaired null depends only on the native laws, not on the level or side experiment. The right side is the standard independent product-testing operation; in type-II-error trade-off notation it is the tensor operation of Dong et al. (2019), with power complemented appropriately.

Here are the envelope facts and proof. The continuous increasing function A has a continuous, increasing, absolutely continuous concave majorant C with the same endpoints. It is affine on every open component of $\{C > A\}$, with contact at the component endpoints. It is strictly below one before its right endpoint: in the convex-hull representation of the envelope, a height-one convex combination can use only the graph point $(1, 1)$. Concavity then implies $C' > 0$ almost everywhere in $(0, 1)$. Consequently F_0^* is a CDF with density $f_0^*(y) = C'(F_1(y))f_1(y)$, and its native likelihood ratio $\ell(y) = 1/C'(F_1(y))$ has a nondecreasing version. Also $F_0^* \ge F_0$, so $P_0^* \preceq P_0$ in stochastic order.

For any feasible monotone h, choose at each y a side NP test of exact B_0-size $h(y)$ and power $R(h(y))$. Stochastic order gives $P_0^* h \le P_0 h \le \delta$, establishing the left-to-right inequality. Conversely, take a product NP test with one common boundary randomization. On shared side support it thresholds $\ell(y)B_1(z)/B_0(z)$. Reject B_1-only states freely; retain zero-ratio states when needed to fill size. Its conditional B_0-size $h(y)$ is nondecreasing and is constant on each open native interval corresponding to $C > A$, since ℓ is constant there. It is a side NP test at each y. A right-continuous version of h satisfies

$$P_0 h - P_0^* h = \int_{\mathbb{R}} (F_0^*(y) - F_0(y))\, dh(y) = 0.$$

Indeed, the integrand vanishes on the contact set, dh vanishes inside noncontact intervals, and boundary terms vanish at infinity. Continuous native laws remove any atom convention; side atoms and free power $R(0)$ are retained. Thus the product test is feasible in the original functional and proves the reverse inequality. This native-law repair is a limiting experiment identity, not a substitute for the finite-mask construction and exact conditional-null calibration above.

P.5 FINITE NOISY ONE-QUERY CERTIFICATES

For $k = 1$ the finite programs of Appendix N extend by replacing their payoff with

$$g_{Kj} = (1 - v_{Kj})f_{j0} + v_{Kj}f_{j1}, \qquad v_{Kj} = q_0 + (q_1 - q_0)(K - j)/b.$$

The iid-null relaxation remains an all-policy upper bound. Monotone rows give full-conditional-null lower bounds by the same shared-channel coupling used above. Rational q_0, q_1 keep every coefficient rational, so the primal calibration and nonnegative-dual box-residual certificate apply unchanged as mathematical principles. This is an extension of the certificate, not a claim that the monotone and unrestricted programs always coincide.

Table 9: Exact outward-rounded noisy one-query brackets at $a = .1, p = .2, \delta = .05, b = .1n$. The channel is known; rows are optimized at $p = .2$.

n	q_0	q_1	Certified lower	Certified upper
100	.1	.9	.312182786279	.312182788569
1,000	.1	.9	.428168138109	.428168151213
100	.1	1	.326060620748	.326060623168
100	.4	.4	.101047865086	.101047865755

The final witness replay checks all four integer proposals with the final source, independently of the optimizer. The uninformative-channel bracket contains the separately computed exact passive upper hull. At $q_0 = 0, q_1 = 1$, the same certificate implementation reproduces the earlier truthful $n = 100, 1000$ witnesses as identical rational numbers. Fixed-width integer overflow is checked before payoff arithmetic; larger coefficients use arbitrary-precision integer arithmetic. The supplement records both original proposal provenance and the final source-bound replay, rather than silently changing an earlier receipt. The symmetric-noise rule at $n = 1000$, without reoptimization or recalibration, has worst-mask power .00705072 at $p = .12$ and .99999997 at $p = .4$. These deterministic evaluations expose design sensitivity; they do not establish uniform superiority.

Q PROOFS FOR ALTERNATIVE RISK AND AUDIT-RELIABILITY CONTRACTS

All results retain the frozen cohort, truthful native ledger, outcome-aware exact-b mask committed before private audit randomness, and a hard budget of distinct pending queries. The classes in the main text change the latent risk or response-reliability contract, not this observation protocol. The arguments below use ordinary stochastic order, conditional Hoeffding bounds, and randomized NP testing.

Q.1 EXACT TRANSFER THROUGH TOTAL-COUNT ENVELOPES

We prove Corollary 2. Write

$$\mathcal{C}_0(a) = \{Q : K \preceq \mathrm{Bin}(n, a)\}, \qquad \mathcal{C}_1(p) = \{Q : K \succeq \mathrm{Bin}(n, p)\}.$$

The order compares every increasing function of the total, not just its expectation. Conditional risks at most a imply the first condition, and conditional risks at least p imply the second: generate V_i from its history-dependent conditional probability using independent uniforms, and use the same uniforms for iid Bernoulli variables with parameter a or p. This gives the corresponding coordinatewise coupling. No independence or exchangeability is required within $\mathcal{C}_0, \mathcal{C}_1$. These containments are not equalities. Sorting two iid Bernoulli(p) values in decreasing order preserves their binomial total, but the second conditional risk is zero after a first zero. The count-lower-envelope condition therefore need not imply a per-step lower bound.

Monotone mask-envelope lemma. Consider uniform sampling of k distinct pending indices and a response table $f(j, z) \in [0, 1]$ nondecreasing in native miss count j and audit positive count z. Responses may be truthful or use a fixed homogeneous binary channel with $q_0 \le q_1$. On a fixed population and mask the expected response depends only on K, j; denote it by $g(K, j)$. For

$$\mathcal{J}_K = \{\max(0, K - b), \ldots, \min(K, n - b)\}, \quad U(K) = \max_{j \in \mathcal{J}_K} g(K, j), \quad L(K) = \min_{j \in \mathcal{J}_K} g(K, j),$$

both U and L are nondecreasing.

For U, fix a maximizing population and mask at $K < n$, retain the same mask and uniform queried indices, and flip one fixed true zero to one. A native flip increases j. A queried pending flip increases its response under a shared uniform channel coupling; an unqueried flip changes nothing. This constructs a legal $K + 1$ configuration of no smaller response. For L, start instead from a *minimizing* configuration at $K + 1$ and flip a true one to zero. Its response cannot increase, so

$$L(K) \leq \text{response of the resulting configuration} \leq L(K + 1).$$

This reverse construction does not assume that optimal masks are nested.

It follows that the exact integrated size and power of this table are

$$\begin{aligned} \sup_{Q \in \mathcal{C}_0(a),\kappa} \mathbb{E}_{Q,\kappa} f &= \mathbb{E}U(\mathrm{Bin}(n, a)), \\ \inf_{Q \in \mathcal{C}_1(p),\kappa} \mathbb{E}_{Q,\kappa} f &= \mathbb{E}L(\mathrm{Bin}(n, p)). \end{aligned} \tag{37}$$

Stochastic order proves the inequalities. Equality holds because the iid laws are included and a legal outcome-aware mask can choose an extremizing native count separately at each K. The same equalities hold with the pointwise conditional-risk upper and lower classes. They are not a least-favourability claim for arbitrary nonmonotone procedures.

The truthful residual-cap rule of Appendix A has the required two-coordinate monotonicity. Increasing j lowers its reference hidden total, so a calibrated upper-tail test rejects no less at each audit count. To check boundary randomization, at a fixed count put $A = \mathbb{P}(Z > z)$ and $B = \mathbb{P}(Z \geq z)$: on a randomized boundary the response is $(s - A)/(B - A)$, nonincreasing in either tail probability. Both tails increase with the reference hidden total. The stated degenerate-law conventions and direct rejection above the cap preserve the order. The gated noisy tables of Appendix P are monotone as well. Equation (37) therefore preserves their finite worst-case values and all their derived limits. No CLT for arbitrary dependent laws is used: Gaussian limits are taken only for the exact binomial least-favourable values. The all-policy upper bounds remain valid because their iid submodels are still allowed.

The passive monotone native-count tests are the $k = 0$ case, so their exact hull also transfers. Every released monotone finite lower witness keeps its size and power, including frozen-table off-design evaluations; the iid upper relaxations still bound the enlarged problem. These facts preserve the reported brackets, not an unproved equality of arbitrary finite minimax values or of unrestricted and monotone linear programs.

Q.2 Conditionally bounded, heterogeneous audit responses

We prove Corollary 3. The bounds in (7) hold conditional on the entire latent vector, committed ledger, exposed past and current queried identity. The causal response kernel cannot anticipate future private audit randomness and may differ under null and alternative. These are conditional reliability bounds, not marginal annotator accuracy estimates.

Two different finite extremal channels. For a monotone table $f(j, z)$, sample uniform distinct indices independently of the response process. Generate responses sequentially with fresh uniforms. For a true zero, the actual response probability lies between 0 and u; for a true one, between v and 1. Use the same uniform for the actual history-dependent probability and both fixed endpoints. On the common sampled truth sequence this gives

$$W_t^{(0,v)} \leq W_t^{\text{actual}} \leq W_t^{(u,1)}, \qquad \mathbb{E}_{(0,v)} f \leq \mathbb{E}_{\text{actual}} f \leq \mathbb{E}_{(u,1)} f.$$

The actual probabilities may depend on earlier actual responses; the reference probabilities depend only on the current true label, so this dependence does not defeat the coupling. A tuple drawn in advance stays private until its entries are queried; equivalently each next unseen index may be drawn with fresh private randomness. No independence conditional on the complete adaptive identity path is asserted.

Let $g^+(K, j)$ use channel $(u, 1)$ and $g^-(K, j)$ use $(0, v)$, and put $U^+(K) = \max_j g^+(K, j)$ and $L^-(K) = \min_j g^-(K, j)$ over $\mathcal{J}_K$. Both fixed comparison channels are monotone, so the preceding

lemma gives increasing U^+, L^-. The exact extrema under the count-envelope classes are

$$S^+ = \mathbb{E}U^+(\mathrm{Bin}(n,a)), \qquad P^- = \mathbb{E}L^-(\mathrm{Bin}(n,p)).$$

The endpoint channels themselves are admissible, and iid laws and countwise extremizing masks attain these bounds. Multiplying f by $\min(1, \delta/S^+)$, with value one when $S^+ = 0$, gives exact finite validity. This calibration is also valid under the original conditional null. It cannot be replaced by known-(u,v) calibration: when $n = b = k = 1$, $a = u = .1$, $v = .9$, and $\delta = .05$, the rule $5W/18$ has size $1/20$ for the known channel, but size $19/360 > 1/20$ for the legal channel $(u, 1)$.

Critical limit. Suppose $u \le v$ and $(b_n - n(p-a))/\sqrt{n} \to s \in \mathbb{R}$. The homogeneous channel (u,v) is admissible. Restricting to it and the iid latent laws gives the full-policy upper bound of Theorem 2; arbitrary heterogeneous kernels need not be exchangeable. For the lower bound take a feasible nondecreasing h in (6), with R now the ROC of $\mathrm{Bin}(k,u)$ against $\mathrm{Bin}(k,v)$. Use the same gated table as Appendix P, with $M_n \to \infty$, $M_n = o(\sqrt{n})$, and nested NP tests of exact allocated size $h((j-na)/\sqrt{n})$ inside the window. The lower gate is literal zero, which matters when a size-zero NP test has free power.

For central null K, any nonzero feasible count j leaves at most $2M_n\sqrt{n} + O(1)$ true errors pending. Under $(u,1)$ its query law is within $O(kM_n/\sqrt{n}) = o(1)$ in TV of $\mathrm{Bern}(u)^{\otimes k}$; the feasible endpoint $j = K$ has that law exactly. For central alternative minimum count $K - b_n$, every central larger j leaves only that many true zeros pending. Under $(0,v)$ the query law is uniformly $o(1)$ from $\mathrm{Bern}(v)^{\otimes k}$, exactly so at $j = K - b_n$. The other gates cannot improve the adversary's extrema. Consequently

$$U^+(K) = h((K-na)/\sqrt{n}) + o(1), \qquad L^-(K) = R(h((K-b_n-na)/\sqrt{n})) + o(1)$$

uniformly on the respective central sets. Binomial CDF convergence against bounded monotone functions and negligible tails imply $S^+ \to \mathbb{E}h(X_0) \le \delta$ and $P^- \to \mathbb{E}R(h(X_1))$. The exact finite calibration tends to one. Fixed arbitrarily near-optimal h prove the matching lower limit without requiring an attained supremum. These designs may know p, s, u, v; p-agnostic noisy attainment is not implied.

If $v \le u$, choose a constant channel rate $c \in [v,u]$, independent of the truth. Any active transcript is then a common randomized postprocessing of the native ledger. Averaging over this channel turns every robust-valid active procedure into a valid passive test. Conversely, a passive test ignores responses. Thus, for every finite n, b, k,

$$V_{n,k}^{C,[u,v]}(\delta) = V_{n,0}^{C}(\delta) \quad (v \le u).$$

The same argument works for any other latent classes with this protocol. Combined with Appendix P's singular-channel analysis, the critical fixed-k consistency criterion is exactly $v = 1$ and $u^k \le \delta$. For $u > v$ one must not use a reversed-endpoint ROC: an uninformative channel is permitted.

Marginal error rates are insufficient. If all responses on true zeros equal one common Bernoulli(u) variable, each marginal false-positive rate is u, but k positives have probability u, not u^k. After a positive response the next conditional rate is one, violating (7). Estimating valid conditional envelopes from average-accuracy validation is outside these results.

Q.3 Predictable cumulative-risk budgets

We prove Theorem 3. Write $\mu_i = \mathbb{E}[V_i \mid V_1, \dots, V_{i-1}]$ and $A_n = \sum_i \mu_i$. The classes $\mathcal{P}_0(a), \mathcal{P}_1(p)$ impose the almost-sure compensator bounds in (8). Conditional Hoeffding's lemma gives

$$\mathbb{E}\exp\{\lambda(K - A_n)\} \le \exp(n\lambda^2/8), \qquad \lambda \in \mathbb{R}.$$

Since $A_n \le na$ under the null and $A_n \ge np$ under the alternative,

$$\sup_{\mathcal{P}_0(a)} \mathbb{P}(K > na + t) \le e^{-2t^2/n}, \qquad \sup_{\mathcal{P}_1(p)} \mathbb{P}(K < np - t) \le e^{-2t^2/n}. \tag{38}$$

For $t_n = n^{3/4}$ the truthful Hoeffding-cap procedure of Appendix A therefore has the same finite validity. On $K \ge \lceil np - t_n \rceil$ its lower proof is pointwise in the population: the feasible native interval, residual hypergeometric ROC and subsequence argument do not require iid outcomes. Thus all

truthful first-order fixed-k limits transfer. Iid-a and iid-p are permitted and give the same all-policy upper bounds.

These classes do not imply binomial stochastic envelopes. For $n = 2, a = .1$, independent risks $0, .2$ have $A_2 = .2$ but $\mathbb{P}(K \geq 1) = .2 > .19 = \mathbb{P}\{\mathrm{Bin}(2, .1) \geq 1\}$. Conversely, sorting two iid Bernoulli(a) outcomes in increasing order preserves the binomial total but has compensator $a^2 + 1 > 2a$ on the positive-probability branch where the first outcome is one. Hence $\mathcal{C}$ and $\mathcal{P}$ are not nested; both contain the original pointwise conditional-risk classes. Binomial-quantile calibration cannot be copied to $\mathcal{P}$.

Query-only critical upper bound. Put $D_n = n(p-a) - b_n = o(\sqrt{n})$ and choose

$$m_n = \left\lceil \sqrt{n(1 + D_n^2)} \right\rceil, \quad c_n = \lfloor na - m_n/2 \rfloor, \quad r_{0n} = \frac{na - c_n}{m_n}, \quad r_{1n} = r_{0n} + \frac{D_n}{m_n}.$$

Then $m_n \to \infty$, $m_n = o(n)$, $D_n^2/m_n \to 0$, and both $r_{0n}, r_{1n} \to 1/2$. Eventually $c_n \geq 0$ and $c_n + m_n \leq n - b_n$, by $a > 0$ and $p < 1$. Use a fixed pending set of size b_n, all zeros under the null and all ones under the alternative. Among native positions place c_n deterministic ones, m_n independent Bernoulli variables of parameter r_{0n} or r_{1n}, and deterministic zeros elsewhere. Their predictable sums are exactly

$$c_n + m_n r_{0n} = na, \qquad c_n + m_n r_{1n} + b_n = np.$$

Both laws are legal, independently heterogeneous populations. Their entire native-ledger laws differ in TV by $e_n \to 0$: the Bernoulli KL bound and Pinsker give

$$e_n^2 \leq \frac{m_n (r_{0n} - r_{1n})^2}{2 r_{1n}(1 - r_{1n})} = o(1).$$

This compares ordered vectors, not only counts, and handles noninteger na, np without conditioning on their realized totals.

Choose the legal homogeneous channel (u, v), $u \leq v$. Every pending truth is zero or one respectively. A full adaptive transcript is a common postprocessing of the native ledger, private seed and k iid responses, with ignored padding after early stopping. Replacing the alternative native law by the null native law costs at most e_n uniformly in the policy. That common native observation is ancillary, leaving exactly the query experiment. Every valid procedure consequently has worst power at most $R_k(u, v; \delta) + e_n$, proving the upper bound in (9). Identity-sensitive policies are included.

Uniform finite-valid lower rule. Use $t_n = n^{3/4}$, $\eta_n = e^{-2t_n^2/n}$ and, except for finitely many cases with $b_n < k$ or $\eta_n \geq \delta$, define

$$s_n = \frac{\delta - \eta_n}{1 - \eta_n}, \qquad q_{0n} = \min\left(1, u + (1-u)\frac{3t_n}{b_n - k + 1}\right).$$

Reject with probability zero if $j < na - 2t_n$, and one if $j > na + t_n$. Between the gates, draw k uniform distinct pending indices and use the canonical randomized upper-tail test of exact $\mathrm{Bin}(k, q_{0n})$ size s_n. Use zero rejection in the finitely many exceptional small-n cases. This rule requires neither p nor v.

On a null population with $K \leq na + t_n$, a central j leaves at most $3t_n$ true ones pending. Conditional on this entire population, mask and past transcript, a fresh uniform unseen query hits a true one with probability at most $3t_n/(b_n - k + 1)$. The conditional response-one probability is at most q_{0n}. A sequential uniform coupling bounds the response count by $\mathrm{Bin}(k, q_{0n})$, even with heterogeneous, history-dependent response probabilities. The direct upper gate is unavailable, and the literal lower gate cannot increase size. Equation (38) gives total size at most $\eta_n + (1 - \eta_n)s_n = \delta$.

On an alternative population with $K \geq np - t_n$, every feasible j satisfies $j \geq na + D_n - t_n \geq na - 2t_n$ eventually. The lower gate is therefore unavailable. In the central region, at most $2t_n - D_n \leq 2t_n + |D_n|$ pending labels are true zeros. Each response-one probability is at least

$$q_{1n} = v \max\left(0, 1 - \frac{2t_n + |D_n|}{b_n - k + 1}\right).$$

This bound is used only in analysis. The response count dominates $\mathrm{Bin}(k, q_{1n})$. Since $q_{0n} \to u$, $q_{1n} \to v$ and $s_n \to \delta$, for $u < v$ the test is eventually NP-optimal against this larger reference parameter. Its power is at least $(1-\eta_n)R_k(q_{0n}, q_{1n}; s_n) \to R_k(u, v; \delta)$. Positive-level finite-alphabet ROC continuity includes $u = 0$ and $v = 1$. For $u = v$, the same rule works because the two reference binomial laws approach each other in TV, so its power is at least $s_n - o(1)$ on the good event. No knowledge of v or wrong-direction NP optimality is needed. This completes the matching proof. Overlapping envelopes $v \le u$ instead allow an uninformative channel, giving limit δ at this scale. No result for a nonzero limiting root-n offset is asserted here.

For $a = .1, p = .2, n = 10m, b = m$, the upper construction can even use common deterministic native labels and pending zeros versus ones. The passive predictable-budget value is exactly δ, unlike the different count-envelope passive limit near $.1086685$ at $\delta = .05$. Thus the change of critical noisy experiment is real, not merely an unavailable proof under dependence.

A strict separation with informative noisy labels. The difference is not limited to an uninformative channel. Fix $a = .1, p = .2, \delta = .05, u = .1, v = .9, k = 1, s = 0$. Let $A = \overline{\Phi}(1)$, $B = \overline{\Phi}(3/4)$ and $r = (.05 - .1A)/(1 - A)$. Since $0 < A < 1/2$, $0 < r < .05$. For the count-envelope Gaussian experiment take

$$h(x) = r + (.1 - r)\mathbf{1}\{x > .3\}.$$

The null standard deviation is .3, hence $\mathbb{E}h(X_0) = .05$. Its values do not exceed .1, where the one-query ROC is $R_1(.1, .9; x) = 9x$. The alternative standard deviation is .4, giving feasible power

$$9\{r + (.1 - r)B\} = .45 + 9(.1 - r)(B - A) > .45.$$

The inequality is exact because $3/4 < 1$. The approximate value $.4863554232$ is only a normal-probability evaluation of this explicit lower bound, not an optimized or rationally certified frontier value. The predictable-budget optimum is exactly .45 by (9). With $k = 0$, the matching-native construction and the baseline coin give the asserted passive limit δ for every $b_n = n(p - a) + o(\sqrt{n})$, not only the integral subsequence above.

Q.4 What mean or marginal guarantees cannot provide

We prove Proposition 1 for truthful queries, $1 \le k \le b \le n$ and $0 < \delta \le a < p < 1$. Let Q_p be iid Bernoulli(p). The mean-null law

$$Q_0 = \frac{a}{p} Q_p + \left(1 - \frac{a}{p}\right) \delta_{0^n}$$

has every marginal risk a. Apply any fixed legal mask and any policy on both mixture components. Nonnegativity and null size at most δ imply $(a/p)\mathbb{E}_{Q_p}\phi \le \delta$. This upper bound applies even if all true labels are provided. Conversely, one uniform pending true label Z gives $T = (J + bZ)/n \in [0, 1]$ with $\mathbb{E}[T \mid V, S] = K/n$. The test $\phi = (\delta/a)T$ is valid for every $\mathcal{M}_0(a)$ law and has power at least $\delta p/a$ under every $\mathcal{M}_1(p)$ law. Iid-p attains equality, so $V^M_{n,k}(\delta) = \delta p/a$ exactly. The same argument works with coordinatewise marginal upper/lower bounds. No noisy-channel HT attainment or formula at $\delta > a$ is asserted.

Even retaining the original strong null but weakening only the alternative to marginal risks p prevents consistency. With $w = (p - a)/(1 - a)$, the exchangeable law

$$Q = (1 - w)\,\mathrm{Bern}(a)^{\otimes n} + w\delta_{1^n}$$

has each marginal p. The same legal mask on both components gives $\mathbb{E}_Q\phi \le (1-w)\delta + w < 1$, even with complete observation. At $a = .1, p = .2, \delta = .05$ this upper bound is $7/45$. These mixture/HT calculations are standard, but show why mean risk cannot be substituted for either conditional risk or count-tail control.

Q.5 Finite certificates under reliability envelopes

For $k = 1$, a table f_{j0}, f_{j1} has two different endpoint payoffs:

$$g^+_{Kj} = (1 - r^+_{Kj})f_{j0} + r^+_{Kj} f_{j1}, \qquad r^+_{Kj} = u + (1 - u)(K - j)/b,$$
$$g^-_{Kj} = (1 - r^-_{Kj})f_{j0} + r^-_{Kj} f_{j1}, \qquad r^-_{Kj} = v(K - j)/b.$$

Writing $q_\theta(K) = \mathbb{P}\{\mathrm{Bin}(n,\theta) = K\}$, an iid necessary-subproblem upper program maximizes $\sum_K q_p(K)\ell_K$ subject to

$$g^+_{Kj} \le u_K, \quad \ell_K \le g^-_{Kj} \quad (j \in \mathcal{J}_K), \qquad \sum_K q_a(K)u_K \le \delta, \qquad 0 \le f_{jz}, u_K, \ell_K \le 1.$$

The fixed channels $(u, 1)$ under the null and $(0, v)$ under the alternative are admissible even for a nonmonotone table; they need not be its globally worst channels. Uniform within-type masks give the same all-policy one-query reduction as Appendix N. No symmetry of arbitrary count-envelope laws is assumed. Adding $f_{j0} \le f_{j1}$ and $f_{jz} \le f_{j+1,z}$ gives lower witnesses whose exact robust size and power are S^+, P^- above. Rational scaling or the baseline-coin correction preserves monotonicity and restores size δ. The nonnegative-multiplier box-residual certificate of (32) applies to the unrestricted upper program. Rational coefficients and exact integer arithmetic, not solver tolerances, certify Table 10.

Table 10: Outward-rounded robust one-query brackets under the conditional reliability envelopes, $a = .1, p = .2, \delta = .05, u = .1, v = .9$. The lower rules are design-point optimized. Both bounds apply to the count-envelope experiment, not the predictable-budget or mean-only classes.

n	b	Certified lower	Certified upper
100	10	.299323210630	.299323212692
1,000	100	.419333491786	.419333504619

Saved witnesses can be replayed without optimization; independent exact reconstruction checks the primal calibration and every dual residual. Numerical proximity at these settings is not a general equality theorem for monotone and unrestricted programs. The certificates are for $k = 1$ and specified p, u, v, not unknown-risk optimality or learned reliability. Their two endpoint channels differ from the known-channel certificates in Appendix P. None is a finite certificate for $\mathcal{P}$ or $\mathcal{M}$ without a different calibration.

Finally, these composite minimax results do not assert that native counts or miss-only observations are sufficient against every specified dependent alternative. A particular dependent law can make revealed-hit patterns informative, while the enlarged alternative class still contains an iid or heterogeneous least-favourable law. That distinction preserves the earlier dependence counterexamples rather than overturning them.

R Independent calibration: what it supplies and what it cannot establish

The reliability envelope in (7) is conditional on the selected item and the full history, not an average accuracy. This appendix gives one sufficient way to supply it, using standard confidence-set calibration. The new information frontiers do not depend on inventing a calibration method. Independent confidence validation is already used by Kato and Nakagawa (2026, Appendix B.5), and label-error validation with uncertainty propagation by Dobi et al. (2026, Section 6.2). Here the two finite endpoint channels determine which confidence failure affects false alarms and which affects a reported power bound.

R.1 A transport model, not a conclusion of validation

Freeze the predictor, the gold labeling standard, and a finite stratum rule $g \in \{1, \dots, G\}$. Gold controls identify whether a frozen prediction is correct ($V = 0$) or incorrect ($V = 1$). Every possible target query belongs to one of these strata. For each g, z fix an unknown number $q_{g,z} \in [0, 1]$. Assume each control response has positive probability $q_{g,z}$ conditional on its stratum, truth, the full truth arrays, current identity, and information exposed before that response. For target responses the same condition also includes the already committed ledger. Controls are not conditioned on a later ledger that may encode their responses. Responses use fresh private random numbers independent of the target truths and audit sampling. The same parameters apply to controls and deployment; they are fixed before calibration. The one-stratum case is a homogeneous channel.

This assumption excludes hidden within-stratum difficulty, drift, recognizing and treating controls differently, and shared latent service failures. Gold truths are assumed correct. Neither the model nor gold correctness is proved by observing a high validation accuracy. Take fixed counts $m_{g,0}, m_{g,1}$ of controls, independently of the target latent population. The pilot D comprises false-positive counts $X_g \sim \mathrm{Bin}(m_{g,0}, q_{g,0})$ and positive-response counts $Y_g \sim \mathrm{Bin}(m_{g,1}, q_{g,1})$. Conditional on any pilot, the target population remains in its stated $\mathcal{C}_0$ or $\mathcal{C}_1$ class. A mask may use both D and the entire target truth vector, provided it commits before private target-query randomness. Thus the argument allows pilot-dependent outcome-aware masks; it does not condition an unconditional latent-risk guarantee on arbitrary dependent validation data.

R.2 Confidence bounds and directional accounting

For $0 < \gamma < 1$, let $U_\gamma(x; m)$ be the one-sided binomial upper confidence bound: for $x < m$ it solves

$$\mathbb{P}_{U_\gamma}\{\mathrm{Bin}(m, U_\gamma) \le x\} = \gamma, \qquad U_\gamma(m; m) = 1.$$

Let $L_\gamma(y; m) = 1 - U_\gamma(m - y; m)$. Binomial-tail inversion gives

$$\mathbb{P}_q\{q > U_\gamma(X; m)\} \le \gamma, \qquad \mathbb{P}_q\{q < L_\gamma(Y; m)\} \le \gamma.$$

Indeed, $q > U_\gamma(x; m)$ implies $F_q(x) < \gamma$, where F_q is the binomial CDF; the event $F_q(X) < \gamma$ is an integer lower tail of probability at most γ. Complementing successes gives the second bound. These are ordinary exact binomial limits, not posterior probabilities at a specified observed count.

Choose confidence errors $\gamma_{g,0}, \gamma_{g,1}$ in advance with sums η_0, η_1. Define $U_g = U_{\gamma_{g,0}}(X_g; m_{g,0})$, $L_g = L_{\gamma_{g,1}}(Y_g; m_{g,1})$, and

$$U(D) = \max_g U_g, \quad L(D) = \min_g L_g, \quad E_0 = \{q_{g,0} \le U_g \ \forall g\}, \quad E_1 = \{q_{g,1} \ge L_g \ \forall g\}.$$

Union bounds give $\mathbb{P}(E_0^c) \le \eta_0$ and $\mathbb{P}(E_1^c) \le \eta_1$. Use vacuous bounds 1 and 0 for an unvalidated cell. Upper bounds may be rounded upward and lower bounds downward to a rational grid. Under the transport model, E_0, E_1 supply the corresponding conditional target envelopes simultaneously across items and histories. The model supplies this implication; the pilot only estimates its parameters.

Proposition 7 (Independent calibration of a monotone audit)**.** *Fix $0 < \eta_0 < \delta < 1$ and set $s = (\delta - \eta_0)/(1 - \eta_0)$. For every pilot d, choose a table $f_d(j, z) \in [0, 1]$, nondecreasing in both counts, and query k distinct pending items uniformly. Let $g_d^+(K, j)$ and $g_d^-(K, j)$ be its expected payoffs on that fixed population/mask under channels $(U(d), 1)$ and $(0, L(d))$. Put*

$$S_d^+ = \mathbb{E} \max_{j \in \mathcal{J}_K} g_d^+(K, j), \qquad K \sim \mathrm{Bin}(n, a),$$

$$B_d = \mathbb{E} \min_{j \in \mathcal{J}_K} g_d^-(K, j), \qquad K \sim \mathrm{Bin}(n, p).$$

If $S_d^+ \le s$ for every d, then under the preceding model

$$\sup_{Q \in \mathcal{C}_0(a), \kappa} \mathbb{P}\{\text{reject}\} \le \delta, \qquad \inf_{Q \in \mathcal{C}_1(p), \kappa} \mathbb{P}\{\text{reject}\} \ge \max(0, \mathbb{E}_D B_D - \eta_1).$$

The mask κ may depend on the pilot. With pilot probability at least $1 - \eta_0 - \eta_1$, the realized table has conditional null size at most s and alternative worst-mask power at least B_D. These are confidence statements over pilots, not guarantees conditional on every observed pilot.

Proof. Fix the physical channel and a pilot $d \in E_0$. Conditional on $D = d, V, S$, sequential uniform coupling bounds the actual responses above by channel $(U(d), 1)$; no sensitivity lower bound is needed. Monotonicity gives $\mathbb{E}[f_d \mid d, V, S] \le g_d^+(K, J)$. The mask-envelope lemma makes $\max_j g_d^+(K, j)$ increasing in K. Since pilot independence preserves the count-envelope null conditional on d, integration over the population bounds conditional size by $S_d^+ \le s$, uniformly over pilot-dependent masks. Writing $e_0 = \mathbb{P}(E_0^c) \le \eta_0$, the unconditional size is at most $s(1 - e_0) + e_0 \le s + (1 - s)\eta_0 = \delta$.

For $d \in E_1$, coupling with $(0, L(d))$ instead gives a lower payoff, and the increasing minimum over masks transfers through $\mathcal{C}_1$. Conditional power is at least B_d, even if $d \notin E_0$. Thus power is at least $\mathbb{E}[B_D \mathbf{1}_{E_1}] \ge \mathbb{E}B_D - \eta_1$, using $0 \le B_D \le 1$. The confidence statement follows by a union bound. □

Scaling an arbitrary nonnegative monotone table by $\min(1, s/S_d^+)$ enforces the premise; use multiplier one if $S_d^+ = 0$. The design can use both parts of the pilot, including an estimated sensitivity, because every chosen table is separately null-calibrated. The reported power threshold p remains fixed: a table optimized at a pilot-selected p_d must still have its certificate B_d evaluated at the reported p, not silently substitute p_d in a uniform power claim. It need not pass a gate $U < L$. A zero-rejection fallback is valid but costs power: deploying only on a pilot event A with $B_D \geq \tau$ yields at least $\tau \max(0, \mathbb{P}_D(A) - \eta_1)$, not unconditional power τ. For predictable-budget classes the confidence composition still works if a genuinely valid conditional stage bound is supplied; the binomial formulas above are not valid for that different latent contract.

R.3 Control costs and a finite decision changed by calibration

For one stratum, $\eta_0 = \eta_1 = .005$ and $\delta = .05$ give $s = 9/199$. With zero mistakes in m controls of each prediction-error truth type, $U = 1 - .005^{1/m}$ and $L = .005^{1/m}$. The smallest m giving $U \leq e, L \geq 1 - e$ is determined exactly by $(1-e)^m \leq .005 < (1-e)^{m-1}$:

Error tolerance e	.20	.10	.05
Controls per truth type	24	51	104
Total gold controls	48	102	208

These are sufficient sizes on the all-clean outcome, not a bound on the number needed to pass with high probability, nor a minimax sample-complexity claim. At a service with rates $(.1, .9)$ the all-clean 102-control outcome has probability $.9^{102}$, about 2.15×10^{-5}.

For the rounded envelope $(U, L) = (.1, .9)$, $n = 1000, b = 100, k = 1$ and $a = .1, p = .2$, the previously certified robust stage optimum at size .05 lies in $[.419333491786, .419333504619]$. After the above confidence charge, reoptimization at stage size $9/199$ gives the exact rational bracket

$$[.390842750727,\ .390842764014].$$

Thus a .40 *stage-power* target passes the supplied-envelope design but fails this conservative calibrated-envelope design, even after its stage table is reoptimized. The interval concerns the original robust response class at a smaller size, not a sharp joint pilot-and-audit optimum. The pilot contains more information than these rounded envelopes; exploiting its full law or tighter endpoints may improve on this modular reduction. The 102 controls are separate from the one target query, and the conditional stage comparison does not imply unconditional power .39084 over pilots.

No finite all-clean binomial pilot gives $L = 1$. It therefore cannot certify the exact perfect-sensitivity premise of fixed-k consistency by this route. If equally costly trusted adjudication is available for pending items, Theorem 1's one-query critical consistency is the pertinent alternative. A benefit from reusable external controls requires that distinct resource/access assumption, not uncounted gold labels.

R.4 Outcome-label validation is not error-response validation

For binary outcome Y, prediction h, and reported outcome $\widetilde{Y}$, our response is $W = \mathbf{1}\{\widetilde{Y} \neq h\}$, not $\widetilde{Y}$. A class-conditional outcome channel with false positives f and sensitivity r induces error-response rates

$$(q_0, q_1) = (f, r) \quad (h = 0), \qquad (q_0, q_1) = (1 - r, 1 - f) \quad (h = 1).$$

This requires the same conditional transport assumption within selected prediction groups. Bounds $f \leq U_f, r \geq L_r$ give $U = \max(U_f, 1 - L_r)$ and $L = \min(L_r, 1 - U_f) = 1 - U$. Both outcome-level confidence errors now contribute to the error-response false-positive failure budget η_0. Charging only the outcome-FPR error would misuse Proposition 7.

R.5 Failure examples and verification scope

Selection can defeat even infinite average-accuracy validation. A time-invariant service is correct on type A and always flips the error indicator on type B. If type B occupies fraction ρ of a representative control population, unlimited controls can identify average false positives ρ and

sensitivity $1-\rho$. A committed pending pool consisting of type B instead has rates $(1,0)$. Prespecified strata with worst-stratum bounds would expose this failure; pooled accuracy hides it. More strongly, controls drawn only from A are perfect with probability one, while a target all-correct B population always receives positive responses. For $n = b = k = 1, a = .1, \delta = .05$, an apparent $U \le .1$ gives the size-s rule $f(0,1) = s/[a + (1-a)U], f(0,0) = 0$. With $s = 9/199$ its actual rejection probability in the bad target population is at least $900/3781 > .05$. Independent controls without transport can lose validity, not just power.

History and target independence matter separately. If all true-zero responses equal one shared Bernoulli(u) variable, their marginal false-positive rate is u, but k positives occur with probability u, not u^k. Separately, let $n = 1, V \sim \mathrm{Bern}(a)$ and reuse $D = V$ as a pilot. Each fixed rule $\phi_0(V) = \delta$ and $\phi_1(V) = (\delta/a)V$ is valid for $\delta \le a$; selecting ϕ_D has size $\delta(1-a) + \delta > \delta$. Fixed-rule validity does not justify conditioning on a pilot that exposes the target truth.

The supplement uses rational-grid binomial inversion and exact arithmetic to check 984 coverage-grid cases, 45 pilot-adaptive small-cohort cases with every count-feasible mask, and 242 outcome-to-error mappings, as well as the integer control thresholds and counterexamples. The selected audit tables actually vary with pilot outcomes; no numerical optimality claim is made for this small family. The new finite stage bracket has saved primal/dual witnesses using the unchanged robust certificate implementation. These finite checks support the proofs, not substitute for their quantifiers.

S AUDIT-DESIGN CALCULATIONS AND RETAINED BOUNDARY EXAMPLES

S.1 MATCHED CALIBRATION CHOICES

Figure 3 compares three specified procedures, not three different information models. *Cap-tail + stage* spends a separate binomial tail allowance; *Exact null, fixed cap* keeps $H = 124$ and recalibrates the complete worst-mask expectation; *Exact null, planning cap* also chooses $H = 144$ for the declared alternative. The dominant gain comes from recovering unused size, not a new sampling method.

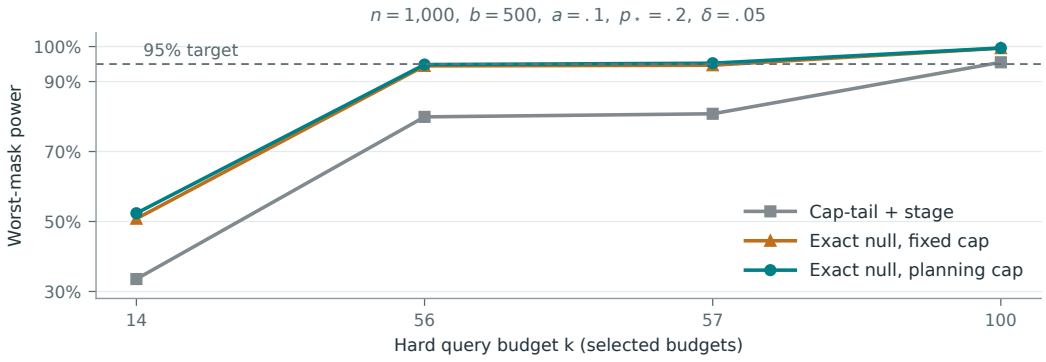


Figure 3: Matched null calibration at $n = 1000, b = 500, a = .1, p_\star = .2, \delta = .05$. All three rules are residual hypergeometric tests under the same truthful count-tail experiment. Exact integer certificates cover every displayed endpoint; lines only connect those budgets. Cap selection for the planning alternative is distinct from null-only recalibration of a fixed cap.

S.2 LIMITING BUDGET CALCULATION AND AN OPTIMISTIC SHORTCUT

For $a = .1, p = .2, \beta = .5, \delta = .05$, assuming all hidden null labels are correct suggests 14 queries, since $1 - .95(.8)^{14} > .95$. Under adversarial natural composition the exact minimum fixed budget for .95 limiting power is instead 56. At 55, the feasible composition $u = .2$ gives power below .94860. At 56, an exact rational interval certificate bounds the entire continuous minimum between .9523690 and .95266494. Ignoring extra queries proves monotonicity in k, excluding every smaller

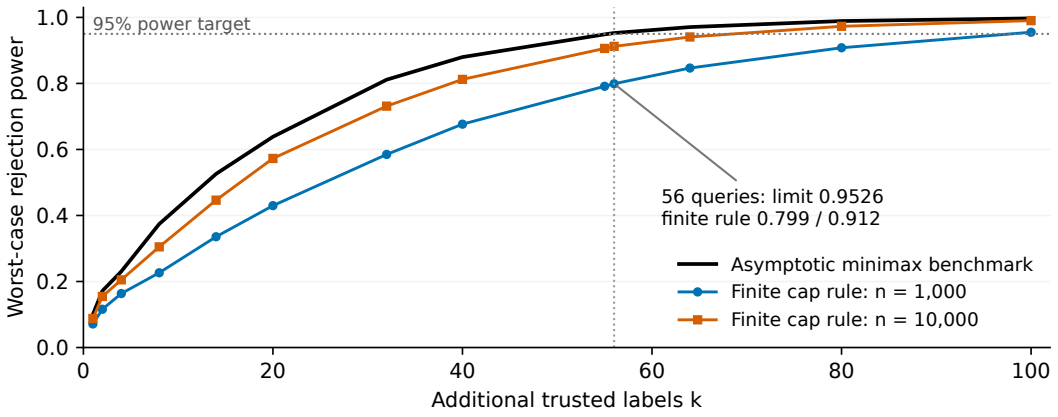


Figure 4: Retained conservative-rule diagnostic, superseded as a sufficient budget by the recalibration in Figure 3. Parameters are $a = .1, p = .2, b = .5n, \delta = .05$, truthful queries. Black: fixed-k asymptotic minimax benchmark. Colored curves: specified cap-tail calibrated worst-mask power, not finite optima. At $k = 56$ these are .7990 ($n = 1000$) and .9122 ($n = 10000$). The 56/100-query anchors have independent integer certificates (Appendix T); other points are floating diagnostics. Connecting lines do not assert growing-k limits.

budget. Fourteen queries guarantee only about .5253–.52594 in this experiment. The shortcut is an illustrative optimistic calculation, not a published baseline; the limiting budget is not a finite-n guarantee. Appendix D supplies the interval certificate.

S.3 Cap-tail rule curves versus limiting information

This subsection retains the earlier conservative-rule diagnostics in Figure 4, not the recalibrated, exactly certified comparisons in Figure 3 and Appendix E. The cap-tail calibrated finite curves use $a = .1, p = .2, b = n/2, \delta = .05$, $n \in \{1000, 10000\}$ and $k \in \{1, 2, 4, 8, 14, 20, 32, 40, 55, 56, 64, 80, 100\}$. Before evaluating power we fix $\eta_n^{\text{target}} = \delta/[2\sqrt{\max(1, n/100)}]$. Let H_n be the smallest integer whose binomial upper tail does not exceed this target; the actual tail, not the target, sets the stage level. This schedule uses no alternative risk or target-label data, vanishes with n, and has $H_n/n \to a$. It differs from the deliberately conservative schedule used in the earlier one-query checks; those adverse results remain in Appendix O.

The curves evaluate every feasible mask count, using hypergeometric probability tables. We truncate only the outer binomial sum, recording omitted probability below 10^{-13} per row. This controls truncation, not floating-point rounding; the displayed curves are numerical diagnostics, not certified decision software. An independent rational implementation checks 330 small cases, 70 direct labeled-population enumerations, and four larger rational power calculations. Interior-count and query-budget counterexamples below prohibit two tempting computational shortcuts.

For the limiting curve, partition I_β into 4096 cells $[l, r]$. Stochastic ordering bounds the cell minimum below by $R_k(r, l + d; \delta)$, while evaluated diagonal points give upper bounds. These evaluations are floating point; the separate rational 1000-cell certificate for the 55/56 budget conclusion is unchanged.

S.4 Finite-cohort power of the specified audit rule

This calculation evaluates the standard residual-cap rule, not the optimum over all finite-cohort procedures. Let

$$\bar{\eta}_n = \frac{\delta}{2\sqrt{\max(1, n/100)}},$$
$$H = \min\{h \in \{0, \dots, n\} : \mathbb{P}(\mathrm{Bin}(n, a) > h) \le \bar{\eta}_n\},$$
$$\eta = \mathbb{P}(\mathrm{Bin}(n, a) > H), \qquad s = \frac{\delta - \eta}{1 - \eta}.$$

The distinction between the requested bound $\bar{\eta}_n$ and its actual binomial tail η avoids discarding available size. This schedule depends only on n, a, δ, not the true alternative risk or observed labels; $\eta \to 0$ and $H/n \to a$. It is one prespecified choice, not an optimized cap schedule. For $j \le H$, let f_j be the randomized upper-tail size-s test under $\mathrm{HG}(b, \min(b, H-j), k)$, including zero-null-mass upper support points; for $j > H$, set $f_j \equiv 1$.

Define $\mathcal{J}_K = \{\max(0, K-b), \ldots, \min(K, n-b)\}$ and

$$g(K) = \min_{j \in \mathcal{J}_K} \mathbb{E}_{Z \sim \mathrm{HG}(b, K-j, k)} f_j(Z), \tag{39}$$

$$P^{\mathrm{cap}}_{n,b,k}(p) = \sum_{K=0}^{n} \binom{n}{K} p^K (1-p)^{n-K} g(K). \tag{40}$$

Conditional on the entire latent vector, every feasible j can be realized by a committed mask. A minimizing j can therefore be selected separately for each K, before the independent audit seed. This proves that (40) is the exact mathematical worst-mask power of this specified rule under the iid alternative. The analogous maximum $w(K)$ gives its worst-mask iid-null size by averaging at a.

Why the count-envelope transfer applies. The rule is nondecreasing in both j and z. Monotonicity in z is part of its definition. To see monotonicity in j, increasing j decreases the reference hidden error count. For fixed z, both reference tail probabilities $\mathbb{P}(Z > z)$ and $\mathbb{P}(Z \ge z)$ then decrease. The randomized upper-tail p-value is uniform between these endpoints. Decreasing both endpoints cannot reduce its probability of being at most s, proving that $f_j(z)$ increases. Degenerate reference laws and the direct-reject branch obey the same ordering.

For a fixed mask and a fixed subset of queried positions, changing a latent zero to one can only increase the native or queried error count, or neither. It cannot decrease rejection probability. Averaging over query subsets and taking either a minimum or maximum over the same set of masks preserves this coordinatewise ordering. Symmetry makes those extrema depend only on K; thus both $g(K)$ and $w(K)$ are nondecreasing. Count-tail ordering consequently transfers the size bound and (40)'s power guarantee to $\mathcal{C}_0(a), \mathcal{C}_1(p)$. The iid laws and attaining masks also belong to these classes, so the displayed extrema are sharp for this rule within these classes. This is not a finite-power transfer to $\mathcal{P}$ or $\mathcal{M}$.

In particular, the specified rule's power is nondecreasing in the iid risk p. One may plan at a prespecified minimum meaningful violation $p_\star > a$: the same p-agnostic rule has at least this power at any iid $p \ge p_\star$, or under $\mathcal{C}_1(p_\star)$. This does not extend a certificate to smaller violations or imply uniformly optimal finite power.

Two shortcuts that fail. The minimum in (39) must not be restricted to endpoints. For $n = 7, b = 4, k = 3, a = 1/3, \delta = 1/5$ and requested cap tail $\delta/4$, one obtains $H = 4, s = 47/290$. At $K = 5$, the three feasible native counts $j = 1, 2, 3$ give respectively $94/145, 143/290, 529/870$. The strict minimum is at the interior count $j = 2$. Moreover, unlike optimal hard-budget power, this particular finite rule's power need not increase in k. For $n = 6, b = 5, a = .1, p = .2, \delta = .05$ and requested tail $\delta/4$, its powers at $k = 1, 2$ are respectively $92587/8322750$ and $360602/62420625$, approximately $.01112$ and $.00578$. When $K \le H$, the true pending error count can be below the reference, making a more informative upper-tail test reject less often. Neither shortcut is used in our curves.

Numerical status and verification. All feasible native counts are scanned. Alternative binomial-tail truncation records its omitted mass, below 10^{-13} in the reported calculations; the missing contribution lies between zero and that mass. Floating-point arithmetic adds an uncertified rounding error, so these are deterministic numerical evaluations, not outward-rounded certificates. An independent rational implementation checks 330 small cases, 70 direct enumerations of labeled populations and all masks/query subsets, and four larger exact-power cases up to $(n, b, k) = (200, 100, 56)$. It also verifies the two shortcut counterexamples and the prespecified cap schedule at four cohort sizes. The code separates these checks from the exact rational limiting-budget certificate for 56 queries. A first crossing of a target by a finite curve describes this rule at the displayed parameters, not the minimum budget possible for any valid method.

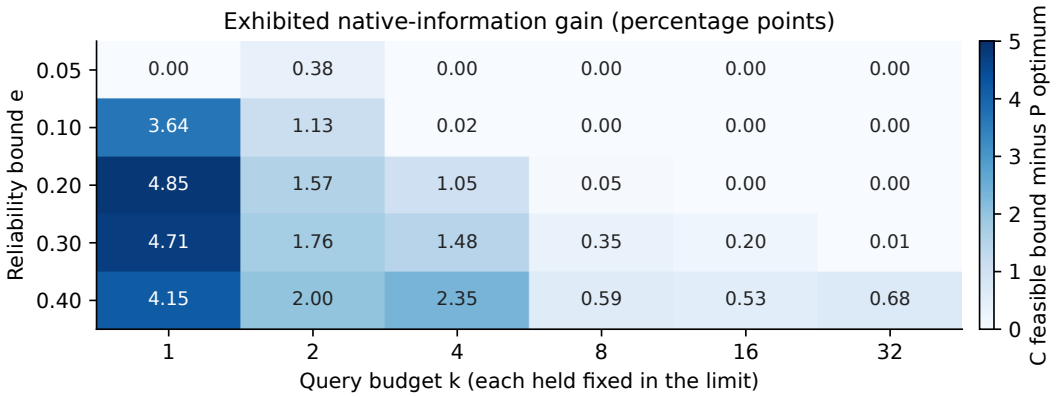


Figure 5: Native information can help beyond one query, but the exhibited gains are often small. Unlike the finite-budget example, this is the noisy critical regime $b/n \to p - a = .1$. At $a = .1, p = .2, \delta = .05$ and conditional $(u, v) = (e, 1 - e)$, cells subtract the exact predictable-risk $\mathcal{P}$ limit from an explicit count-tail $\mathcal{C}$ feasible lower bound, in percentage points. The latter is not the $\mathcal{C}$ optimum; zeros after rounding do not prove equality. Deterministic floating-point evaluation of two-level allocation witnesses; each k is fixed as n grows. Appendix S gives the construction.

S.5 Native-information sensitivity: feasible witnesses, not optima

For $a = .1, p = .2, \delta = .05$ and zero root-n offset, set $A = \mathbb{P}(X_0 > x)$, $B = \mathbb{P}(X_1 > x)$, where $X_0 \sim N(0, a(1-a))$ and $X_1 \sim N(0, p(1-p))$. A two-level nondecreasing allocation takes value r below x and t above it, with $r = (\delta - tA)/(1 - A)$ and $\delta \le t \le \min(1, \delta/A)$. Its feasible $\mathcal{C}$ limiting power is $(1 - B)R_k(u, v; r) + BR_k(u, v; t)$, whereas the exact $\mathcal{P}$ limit is $R_k(u, v; \delta)$. We retain the constant allocation and the original $x = \sqrt{a(1-a)}, t = .1$ witness, which gives .4863554232 for $k = 1, e = .1$. Thus this study extends, rather than replaces, the old example.

We evaluate $k = 1, 2, 4, 8, 16, 32$ and $(u, v) = (e, 1 - e)$ with $e = .05, .1, .2, .3, .4$. Thresholds in null-standard-deviation units range from -3 to 6 in steps of .05; feasible allocations are checked on 201 evenly spaced fractions of their interval, at $t = .1$ when feasible, and at ROC breakpoints for either t or induced r. Every retained witness meets its null budget in the numerical checks. Independent exact rational NP calculations verify the binomial ROC components for all 30 cells; a normal-series check independently reproduces the old witness. Gaussian probabilities and searched lower bounds remain floating-point evaluations, not rational certificates.

The exhibited gains persist at larger budgets, e.g. $e = .4, k = 32$ gives .747568 against .740778. They can be tiny as query-only power saturates, and a small searched gain does not upper-bound all possible native gains. The search knows the design alternative and reliability envelope; it supplies no finite-n power guarantee, nonzero-offset $\mathcal{P}$ theorem, growing-k limit, or empirical validation of the risk/reliability assumptions.

S.6 One-query formula and familiar baselines

For $k = 1$ and $\beta > \epsilon$,

$$R_1(u, u + d; \delta) = \begin{cases} \delta + (1 - \delta)d/(1 - u), & u \le \delta, \\ \delta(u + d)/u, & u \ge \delta. \end{cases} \tag{41}$$

This increases on the first branch and decreases on the second; thus the minimum over I_β is attained at an endpoint. The convention at $u = 0$ uses the first branch. For $a = .1, p = .2, \beta = .5, \delta = .05$, $I_\beta = [0, .2]$. Assuming $u = 0$ predicts power .24. The true worst composition $u = .2$ gives only .10. Thus a favorable all-correct hidden null can overstate the power of even a trusted audit. It is not sufficient to count the probability of encountering one of the excess hidden errors.

Let Z be the queried item's true error indicator. The standard expansion estimate $\widehat{r}_{\mathrm{HT}} = (J + bZ)/n \in [0, 1]$ satisfies $\mathbb{E}[\widehat{r}_{\mathrm{HT}} \mid V, S] = K/n$ (Horvitz and Thompson, 1952). The valid test

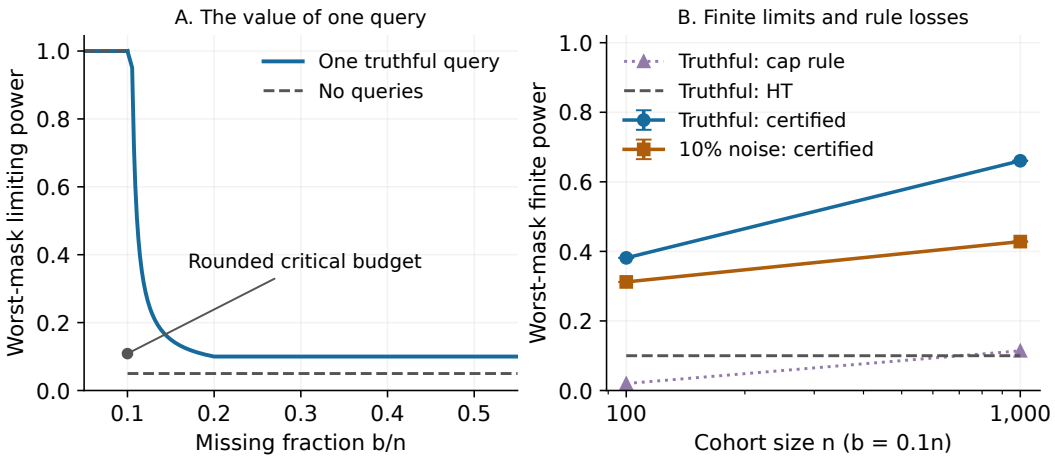


Figure 6: A one-query boundary case of the general fixed-budget theory. Fixed $a = .1, p = .2, \delta = .05$. Left: exact one-query limits versus the passive boundary; the isolated passive point uses $b_n = \lfloor .1n \rfloor$. Right: at $b = .1n$, certified finite minimax brackets for truthful and known 10% symmetric-noise queries, versus the truthful conservative rule and HT. Bracket widths are smaller than the markers; connecting segments guide the eye. The design-point lower rules need not perform well at other p. No Monte Carlo data are used.

$\phi_{\mathrm{HT}} = \delta + c_0(\widehat{r}_{\mathrm{HT}} - a)$, where $c_0 = \min(\delta/a, (1-\delta)/(1-a))$, has mask-independent power

$$H(a, p, \delta) = \delta + c_0(p - a). \tag{42}$$

This already equals the one-query frontier whenever $\beta \geq p, \delta \leq a$, or $\beta \geq 1 - a, \delta \geq a$. These are sufficient, not exhaustive, equality regions. In particular the .10 example is explained by a familiar linear test; its value is not evidence for a new audit method. At the critical budget, however, $H < 1$ while the optimal limiting power is one.

S.7 Retained passive boundary and one-query anchors

Proposition 8 (Passive cohort boundary)**.** *With no queries, iid Bernoulli(p) alternative and the conditional null* (1)*, the full-status maximin ROC is the upper hull of*

$$(0,0),\ (1,1),\quad \big(\mathbb{P}\{\mathrm{Bin}(n,a) \geq r\}, \mathbb{P}\{\mathrm{Bin}(n,p) \geq r + b\}\big),\quad 1 \leq r \leq n - b. \tag{43}$$

A randomized nondecreasing native miss-count test attains it. The same value holds with at most b rather than exactly b erasures.

Appendix L proves a broader count-MLR version and gives an attainable null construction. In the iid exact-budget subproblem, reducing full status to an unrestricted count test follows directly from permutation symmetry. The explicit monotone upper hull and its attainable masks yield the passive benchmark for the intervention.

Calibration can change a concrete decision. At $n = 1000, b = 100, k = 1$ with rounded $(U, L) = (.1, .9)$, spending $\eta_0 = .005$ changes the stage level from .05 to $9/199$. Its reoptimized robust value lies in $[.390842750727, .390842764014]$, below a .40 stage-power target. This is a cost of the stated modular envelope design, not an impossibility for joint pilot-and-audit methods or an unconditional power over pilots (Appendix R).

T CERTIFIED FINITE BUDGETS FOR AN INTEGER DECISION RULE

This appendix certifies a specified implementation of the existing residual-cap rule. It does not optimize finite minimax power, introduce a new sampling principle, or validate a deployed observation protocol. All four cases use $a = 1/10$, planning alternative $p_\star = 1/5$, $\delta = 1/20$, truthful responses, and $b = n/2$.

T.1 THE IMPLEMENTED RULE AND ITS EXACT GUARANTEE

Let $f_j(z)$ be the ideal cap rule of Appendix S.4 and set $L = 2^{80}$, $f_j^-(z) = \lfloor L f_j(z) \rfloor / L$. This pointwise downward rounding preserves monotonicity in both native and queried error counts, and can only lower null rejection probability. Hence size remains at most δ under $\mathcal{C}_0(a)$, for every legal committed mask. The same count-monotonicity argument transfers worst-mask power to $\mathcal{C}_1(p_\star)$. The per-transcript loss is smaller than $1/L$; minimization over masks and averaging over populations preserve this bound between the ideal and implemented powers.

To certify the cap without an approximate quantile, compute integer binomial masses and their denominator D_0. If T_H is the integer upper-tail numerator at cap H, its prescribed schedule condition is exactly

$$4 \max(100, n)\, \delta_{\mathrm{den}}^2 T_H^2 \;\le\; 100\, \delta_{\mathrm{num}}^2 D_0^2.$$

The selected H passes and its predecessor fails. This gives $H = 124$ for $n = 1000$ and $H = 1085$ for $n = 10000$. The stage size $s = (\delta - \eta)/(1 - \eta)$ is an exact rational, with $\eta = T_H / D_0$; no square-root approximation enters calibration.

Write $D = \binom{b}{k}$, $w_{m,z} = \binom{m}{z}\binom{b-m}{k-z}$ and $A_{m,c} = \sum_{z>c} w_{m,z}$. For $j \le H$, the reference hidden count is $m_j = \min(b, H - j)$. If $s = S/T$ and c_j is its exact upper-tail randomized cutoff, then

$$\gamma_j = \frac{SD - TA_{m_j,c_j}}{T w_{m_j,c_j}}, \qquad G_j = \lfloor L\gamma_j \rfloor.$$

The rejection table is zero below c_j, G_j/L at c_j, and one above it; $j > H$ rejects directly. Zero-null-mass upper outcomes are included. An independent private uniform 80-bit word U implements the boundary decision by $U < G_j$, avoiding a floating-point coin. This decoder assumes its certified table and honest uniformly sampled distinct pending indices. It does not establish label truthfulness, mask commitment, or unbiased sampling from an untrusted ledger.

T.2 EXACT POWER ARITHMETIC AND INDEPENDENT VERIFICATION

For a fixed total K and feasible $j \le H$, the implemented payoff is

$$\frac{LA_{K-j,c_j} + G_j w_{K-j,c_j}}{LD}.$$

All mask payoffs share the denominator LD, so their minimum is an exact integer comparison over *every* feasible j, not an endpoint shortcut. The outer $\mathrm{Bin}(n, p_\star)$ masses also share an integer denominator. The producer includes the contiguous range of totals whose mass is at least 2^{-80} and records the exact omitted mass; its contribution is between zero and that mass. If no such total exists at a coarser precision, the implementation sums every total. The omitted mass is less than $(n+1)2^{-80}$. Only the final display is rounded outward to decimals.

A separate verifier does not import the producer: it uses direct binomial coefficients for the hypergeometric weights, a backward binomial recurrence, and *all* $K = 0, \ldots, n$, without tail truncation. It checks the cap, every decision row, and both minimum/maximum mask envelopes. Thus it verifies power and the actual worst-case $\mathcal{C}_0$ size of the *same* implemented decision tables. Saved intervals below contain its exact full sums; the original ideal rule also lies in these displayed intervals.

Independent small-case checks compare against direct Fraction arithmetic, including coarse two-bit rounding: 720 power intervals, 5220 exact minimum/maximum terminal values, 20348 decision outputs and 720 ordered-null size calculations. The production-scale complete-sum replay evaluates 10987422 nondirect mask payoffs across the four cases. These checks are internal computational verification, not external human peer review or a machine-checked proof.

Table 11: Outward rational power certificates for the supplied 80-bit rule. All intervals have width 10^{-12}. These are specified-rule values, not finite minimax intervals. Parameters are those stated above.

n	b	k	Power lower	Power upper
1000	500	56	.799016366407	.799016366408
1000	500	100	.955061390857	.955061390858
10000	5000	56	.912227061827	.912227061828
10000	5000	100	.990263071780	.990263071781

Does unused false-alarm allowance explain the shortfall? The independent replay also computes the exact worst-case null size S. For $k = 56$ it is approximately .01204 at $n = 1000$ and .02579 at $n = 10000$, below .05. The standard independent-coin correction from Appendix M, $q = (\delta - S)/(1 - S)$, changes power P to $q + (1 - q)P$ and uses size exactly δ. Its exact outward power intervals here are $[.806738923610183, .806738923610184]$ and $[.914408614055009, .914408614055010]$, respectively, still below .95. Thus simply spending the unused size does not close this 56-query shortfall. This does not exclude a different cap, better recalibration, or a better finite procedure. The correction is established bookkeeping, not a new method, and the primary certificates above concern the uncorrected supplied tables.

T.3 Which finite-design conclusions follow?

The lower endpoint for $n = 1000, b = 500, k = 100$ exceeds .95. Thus 100 truthful queries suffice for this implemented rule at level .05 under $\mathcal{C}_0(.1)$ versus $\mathcal{C}_1(.2)$, in the frozen, single-deadline protocol. The rule itself does not use $p_\star$. Monotonicity in K extends the guarantee to all iid $p \geq .2$ and the corresponding count-tail alternatives, not to every population with average risk at least .2.

For example, independent Bernoulli risks $(.1, .3)$ have average .2 and belong to $\mathcal{P}_1(.2)$, but $\mathbb{P}(K \geq 2) = .03 < .04$ excludes them from $\mathcal{C}_1(.2)$. The assumption is about the risk-generating law, not a fact learned from the target cohort. Even complete labels need not allow rejection of every realized cohort with $K/n > a$, because such cohorts occur under acceptable-risk laws. Power is not the posterior probability that a particular alarm is correct, and nonrejection does not certify the realized error count.

The 100-query design is sufficient, not necessary or optimal. No interpolation, binary search, or monotonicity of this rule in k is assumed. The finite noisy budgets are not certified by these truthful tables: Figure 5 instead concerns noisy labels at $b/n \to p - a = .1$, not this $\beta = .5$ example. The supplied decoder is a reproducible statistical component, not a complete deployment system. Repeated deadlines require separate global error control.