TESTING HYPOTHESES WITH KNOWLEDGE NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

When does a prerequisite network make a hypothesis cheaper to test? We study one conjunction of binary hypothesis under supplied prerequisite rules and fresh noisy queries, with validity required at every legal state. On a directed acyclic graph allowing at most r failed rules, the verification cost at a supplied legal ideal I is $\Theta_\eta(C_r(G[I], S)\log(1/\delta))$ at every confidence level $\delta \leq 1/4$, uniformly in the graph, where $C_r(G[I], S)$ is a fractional cover of the low cut missing sets inside I. When the mastered witness is unknown and is located through a diagnostic table E, the cost is $\Theta_\eta(\min\{k\log(1/\delta), \mathcal{H}_\eta(E) + \log(1/\delta)\})$, where $\mathcal{H}_\eta(E)$ is the cost of hitting a witness without recognizing it: discovery and confidence add. Over all tables, h diagnostic pairs for 2^h witnesses leave a discovery cost of order $h\log(h+1)$, while a constant factor surplus of pairs lowers it to order h, even against unbalanced tables and adaptive proposals. The laws survive two relaxations. With w redundant witnesses and at most $r < w$ failed rules, only the confidence term grows, by $w/(w-r)$. With error rates unknown but bounded by a known margin, the optimized resource law is unchanged; this covers sampled outputs graded by exact checkers under a pass rate gap. The converse also requires every grade pattern satisfying the witness inclusions and diagnostic pair incompatibility to be realizable. Pairwise implications of a union closed knowledge space preserve full mastery certificates, but not partial profile ones. Preregistered experiments reproduce the additive law and the diagnostic resource law in simulation and on sampled outputs of language models, and locate where they stop: under persistent errors, and under rules read off difficulty labels.

1 INTRODUCTION

Success on one advanced task can certify several simpler capabilities at once: if an advanced capability logically requires every target, observing it certifies the whole claim. An evaluator that uses such implications needs fewer queries, but only if it can find the evidence and trust the prerequisite rules behind it. We study both costs as the noisy query complexity of testing one conjunctive claim under supplied prerequisite rules. The instance to keep in mind is a system whose outputs are sampled and graded by exact checkers: a query draws one fresh output and returns one checker's grade of it. The statistical guarantee is uniform over every allowed state, even when cost is measured on a favorable subset.

Consider k basic targets and t advanced witnesses, each requiring all targets. At a profile mastering exactly one witness, an evaluator given its location needs order $a = \log(1/\delta)$ fresh noisy queries. Without the location, the optimal cost has order $\min\{ka, t + a\}$: direct target testing competes with witness search followed by confirmation. Supplying the favorable family does not supply the answer. The same evaluator must also be correct when no witness, or some target, is mastered.

The paper follows the resulting testing problem in four steps.

- **From structure to verification.** Missing sets with small outgoing edge cuts describe the null alternatives when some original rules may fail. Their fractional cover gives a valid sequential test and its matching cost at a supplied legal ideal (Theorem 3.1); on two layer networks the cover is an explicit program in which shared prerequisites and redundant rules determine the cost. The supplied graph is justified by a knowledge space result: the pairwise implications of a union closed family keep every full mastery certificate, though not partial profile ones (Appendix C).

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.

- **From verification to discovery.** For unknown favorable profiles on a general DAG, covers of informative and certifying departures give a discovery sandwich (Theorem 4.2). For a private witness diagnostic table E, discovery and confidence add exactly up to constants: $\min\{ka, \mathcal{H}_\eta(E)+a\}$ (Theorem 4.1). Here a search need only *hit* the actual witness; it need not know that it has done so.
- **Which diagnostics reduce discovery.** With $t = 2^h$ witnesses, we optimize over all tables of at most m diagnostic pairs (Theorem 5.1). The sharp budget law covers duplicate rows, unbalanced tables and adaptive proposals. It separates the ambiguity remaining below an h-bit address from the reliability gained by spending surplus pairs on a code.
- **Where the laws hold.** With redundant witnesses, failed rules change only the confidence term, by the factor $w/(w-r)$, under a global or a per group failure budget (Theorem 6.1). Unknown error rates within a known margin leave the resource law unchanged (Theorem 6.2); graded samples under a pass rate gap fall in this class, and incompatible checks produce the complementary failures the law uses.

Preregistered measurements reproduce the additive and resource laws in simulation and on language model outputs, and the need for the factor $w/(w-r)$ in simulation; Section 7 shows where their assumptions stop, and Appendix L gives every design.

The distinctive converse concerns unrecognized hitting in an arbitrary fixed noisy table, and its connection to testing that remains valid beyond the favorable profiles. Fractional certificates, sequential information allocation and error correcting codes are established ingredients (Gilmer et al., 2016; Garivier & Kaufmann, 2016; Naghshvar & Javidi, 2013); Appendix A discusses related work. Throughout, Θ_η hides positive constants depending only on a fixed noise level η.

2 Model

2.1 The supplied graph and the query model

Let V be a finite set of capabilities, and let a state $x \in \{0,1\}^V$ record which are mastered. Knowledge space theory describes feasible states by a family closed under unions (Doignon & Falmagne, 2015); for the conjunction question studied here, such a family and its pairwise implications have the same certificates at complete mastery, but not at partial profiles (Proposition C.1, Appendix C). We therefore take a finite simple DAG $G = (V, \mathcal{E})$ as the supplied model. An original edge $u \to v$ asserts $x_v \le x_u$; the exact legal states are the order ideals

$$\mathcal{K}_0(G) = \{x : x_v \le x_u \ ((u,v) \in \mathcal{E})\}. \tag{1}$$

To study trust in the rules, allow a nonnegative integer budget r of failed original edges at the tested state:

$$B_G(x) = \sum_{(u,v)\in\mathcal{E}} (1-x_u)x_v, \qquad \mathcal{X}_r(G) = \{x : B_G(x) \le r\}. \tag{2}$$

The budget counts original rules, with no parallel copies; adding transitive edges can change $\mathcal{X}_r$ for $r > 0$, although it preserves $\mathcal{K}_0$.

A node A_n, selected using the past, returns a fresh noisy bit:

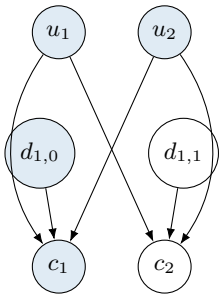
$$Y_n = x_{A_n} \oplus Z_n, \qquad Z_n \sim \mathrm{Bern}(\eta), \qquad \ell_\eta = \log\frac{1-\eta}{\eta}, \quad \kappa = (1-2\eta)\ell_\eta, \quad 0 < \eta < 1/2. \tag{3}$$

The level η is known; Section 6.2 replaces it by a known margin. Each flip is independent of all past flips, the algorithm's randomness and the selected node. Repeated queries cost one each and supply new noise. Put $p_1 = 1-\eta, p_0 = \eta$, so $\kappa = \mathrm{kl}(p_1, p_0)$; logarithms are natural unless marked otherwise.

For a nonempty target set S with $k = |S|$, the hypotheses are

$$H_0 = \{x \in \mathcal{X}_r : \exists s \in S, x_s = 0\}, \qquad H_1 = \{x \in \mathcal{X}_r : x_s = 1 \ \forall s \in S\}. \tag{4}$$

For $0 < \delta < 1/2$, a δ-correct procedure stops almost surely at every state in $\mathcal{X}_r$ and errs with probability at most δ at every state of $\mathcal{X}_r$. Write $a = \log(1/\delta)$. For a supplied exact legal ideal $I \supseteq S$,



Targets: the claim is $x_{u_1} = x_{u_2} = 1$.

Diagnostics: help choose a witness; they do not certify the targets.

Witnesses: either mastered leaf certifies both targets under exact rules.

Shading shows I_1; I_2 switches the diagnostic and witness.

Figure 1: The private architecture with $k = t = 2, m = 1$ and rows 0 and 1. Edges point from prerequisite to dependent. The favorable profiles guide the efficiency objective, while correctness covers every legal ideal.

and for a known finite nonempty family $\mathcal{F}$ of such ideals, define

$$\mathfrak{T}_r(G, S, \delta; I) = \inf_{\mathcal{A}\ \delta\text{-correct on } \mathcal{X}_r} \mathbb{E}_{\mathbf{1}_I} \tau_{\mathcal{A}}, \tag{5}$$

$$\mathfrak{D}_r(G, S; \mathcal{F}, \delta) = \inf_{\mathcal{A}\ \delta\text{-correct on } \mathcal{X}_r} \max_{I \in \mathcal{F}} \mathbb{E}_{\mathbf{1}_I} \tau_{\mathcal{A}}. \tag{6}$$

The first infimum may choose a different valid evaluator for each supplied I; the second chooses one evaluator without knowing the actual member of $\mathcal{F}$. Neither may assume that the tested state belongs to that favorable family when making its decision. Abbreviate $\mathfrak{T}_r(G, S, \delta) = \mathfrak{T}_r(G, S, \delta; V)$.

2.2 Private witnesses and diagnostic tables

The sharpest results concern one family of networks (Figure 1). A table $E : [t] \to \{0, 1\}^m$ specifies target roots u_i, diagnostic roots $d_{j,0}, d_{j,1}$, and witness leaves c_v, with edges and favorable profiles

$$u_i \to c_v\ (i \in [k], v \in [t]), \qquad d_{j,E(v)_j} \to c_v\ (j \in [m], v \in [t]), \tag{7}$$

$$I_v = S \cup \{d_{j,E(v)_j} : j \in [m]\} \cup \{c_v\}, \qquad v \in [t]. \tag{8}$$

Each witness implies every target, while the diagnostics tell the evaluator where to look. Denote the family cost (6) at $r = 0$ by $\mathfrak{D}_{E,k}(\delta)$. The graph's full legal class also permits both or neither members of a diagnostic pair and multiple witnesses; only the favorable profiles have the complementary failures in (8). Section 6.2 shows how contract design produces them.

Definition 2.1 (Unrecognized witness discovery)**.** Let $\mathcal{H}_\eta(E)$ be the smallest positive integer q for which a procedure that knows E and uses at most q diagnostic or witness queries, target queries being free, queries c_v with probability at least $3/4$ at every I_v. It may query adaptively and repeatedly and need not recognize a hit.

Target responses have the same law on all these profiles and can be simulated internally; $1 \le \mathcal{H}_\eta(E) \le t$, and any other fixed success level changes only constants, by independent repetition. The budget m counts available pairs and q counts observations: a coded diagnostic is a separately available node, not a parity of noisy answers.

3 Network structure determines verification

3.1 Null alternatives are low cut missing sets

For a missing set $D \subseteq V$, write $\partial^+_G(D) = \{(u, v) \in \mathcal{E} : u \in D, v \notin D\}$ and $x^D = \mathbf{1} - \mathbf{1}_D$. Then $B_G(x^D) = |\partial^+_G(D)|$, so the opposite answer differences at full mastery are

$$\mathcal{D}_r(G, S) = \{D : D \cap S \neq \varnothing, |\partial^+_G(D)| \le r\}.$$

A target can be absent while its descendants remain present only by spending the edge budget on the boundary of the missing set. A set of mastered nodes certifies the claim without noise exactly when it meets every such D, which by Menger's theorem means $r + 1$ edge disjoint directed paths from each

unqueried target into the set (Proposition E.1). The noisy counterpart allocates fractional evidence to those same sets:

$$C_r(G,S) = \min_{\lambda \geq 0} \left\{ \sum_{v \in V} \lambda_v : \lambda(D) \geq 1 \quad (D \in \mathcal{D}_r(G,S)) \right\}, \tag{9}$$

with $\lambda(D) = \sum_{v \in D} \lambda_v$. This is fractional certificate complexity for the promise function (4); its LP dual is fractional block sensitivity (Gilmer et al., 2016). An optimum exists in $[0,1]^V$, $1 \leq C_r \leq k$, C_r is nondecreasing in r, and it is at most the minimum noiseless certificate size. At $r = 0$, the constraints reduce to $\lambda(\mathrm{Desc}_G(s)) \geq 1$, $s \in S$, with each node among its own descendants. For positive r, separating its weighted covering inequalities is NP hard when r is part of the input, which does not settle the complexity of the number C_r itself; for each fixed r the sets reachable from a target after deleting at most r edges generate every constraint, an explicit polynomial size program (Appendix M).

3.2 A VALID SEQUENTIAL VERIFIER AND ITS COST

Theorem 3.1 (Verification at a supplied legal ideal)**.** *Let I be an exact legal ideal with $\varnothing \neq S \subseteq I$. For any integer $r \geq 0$ and $0 < \delta < 1/2$, put $C = C_r(G[I], S)$, $a_\delta = \log(2/\delta)$, and $M = \lceil 2Ca_\delta \rceil$. Then*

$$\frac{C}{\kappa}\,\mathrm{kl}(1-\delta,\delta) \leq \mathfrak{T}_r(G,S,\delta;I) \leq \frac{M(\log 2 + \ell_\eta)}{\kappa} + \frac{\delta(a_\delta + \ell_\eta)}{2\kappa}. \tag{10}$$

For fixed η, the cost is $\Theta_\eta(Ca)$, uniformly in the graph and $0 < \delta \leq 1/4$. For fixed G, S, I, r, η, its ratio to a tends to C/κ as $\delta \downarrow 0$. The finite upper bound queries only I and stops in finite expectation at every state in $\mathcal{X}_r(G)$.

The lower bound is a stopped change of measure: against each x^D, $\kappa \sum_{v \in D} \mathbb{E}_{\mathbf{1}} N_v(\tau) \geq \mathrm{kl}(1-\delta,\delta)$, so normalized expected query counts form a feasible fractional certificate. At a supplied ideal no edge enters I: each induced missing set extends by zeros outside I to a valid null, and restriction to I cannot increase violations, so $C_r(G[I],S)$ is the right quantity (Appendix E.4). The upper bound (Algorithm D.3) draws a node from λ/C in each of M blocks and runs a fresh local test with thresholds $\log 2$ and $-\log(2kM/\delta)$; a lower exit calls a fresh direct test of the targets, and M upper exits accept. On any null, each block samples a missing coordinate with probability at least $1/C$ and detects it with probability at least $1/2$, so missing every alarm has probability at most $(1 - 1/(2C))^M \leq \delta/2$. At the supplied reference each block has constant expected cost and the fallback runs with probability at most $\delta/(2k)$. The fallback is needed because an absent nontarget alone must not reject a true conjunction. Appendix D proves the finite constants, stopping integrability and the exact limit.

Which network features lower the coefficient. On a two layer graph with all edges from k basic targets to t advanced nodes, C_r is an explicit linear program of size $O(|V| + |\mathcal{E}|)$: a missing target forces all but at most r of its neighbors to be missing, so the binding constraint uses the target and its cheapest $(d_u - r)_+$ neighbors (Theorem E.2). When every target has degree q and every advanced node has indegree b, the coefficient is $\min\{k, t/(q-r)\}$ for $r < q$ and k otherwise (Corollary E.3): shared prerequisites and degree redundancy, not finer topology, determine it. Fractional and logical certificates differ: on $K_{k,t}$ with $k = 64$, $t = 32$, $r = 16$, the smallest noiseless certificate has 17 nodes while $C_r = 2$ (Proposition E.4). A single advanced witness gives $C_0 = 1$, but one failed rule raises its coefficient to k: structural savings require a specified degree of trust in the supplied implications, which Section 6.1 buys back with redundancy. Appendix E states and proves these results.

4 FINDING USEFUL EVIDENCE

The supplied ideal cost does not charge for learning which reference is useful. This section removes that information while retaining exact rules, $r = 0$: the graph and the favorable family are known, the actual state is not. Discovery gains remain conditional on that family, while validity still covers every legal state.

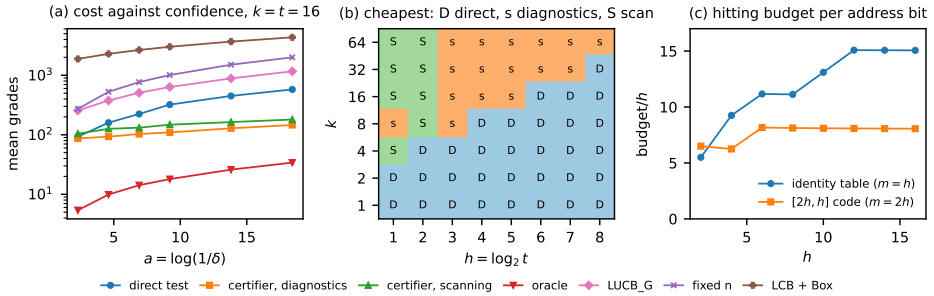


Figure 2: The laws on the symmetric channel with $\eta = 1/4$. (a) Mean grades against $a = \log(1/\delta)$ at $k = t = 16$, 1000 runs per point (200 for LUCB_G and LCB + Box). (b) The cheapest procedure at $\delta = 0.01$ over k and $h = \log_2 t$, 300 runs per cell. (c) Hitting budgets per address bit of majority reads of the identity table ($m = h$, exact) and of maximum likelihood decoding of a systematic $[2h, h]$ code ($m = 2h$, 4000 Monte Carlo trials).

4.1 Discovery adds to confidence

Theorem 4.1 (Discovery adds to confidence)**.** *For fixed $0 < \eta < 1/2$, uniformly over every table E, $k, t \geq 1$, $m \geq 0$ and $0 < \delta \leq 1/4$,*

$$\mathfrak{D}_{E,k}(\delta) = \Theta_\eta(\min\{ka, \mathcal{H}_\eta(E) + a\})\,. \tag{11}$$

The upper bound stops in finite expectation on the full legal class.

A search policy proposes witnesses; independent asymmetric checks turn its unrecognized hits into valid evidence. False witnesses have constant expected checking cost. On a favorable profile, deduplication leaves only one true witness to pay the high confidence threshold. Fresh constant success cycles, raced against direct testing, cost $O_\eta(\min\{ka, \mathcal{H}_\eta(E) + a\})$. For the converse, delete the private witness while keeping its diagnostic labels. This is a legal positive baseline, and deleting any target from it gives a null. A valid test must either acquire order ka target information or hit the private witness with constant probability. Truncation and restarts give an admissible bounded horizon search (Appendix G). Its finite policy LP is exponential in the horizon, so this is not an efficient algorithm for an arbitrary table.

Measured. Costs count queries, called grades. On the symmetric channel with $\eta = 1/4$ at favorable profiles with $k = t = 16$, the cost of direct testing grows by 30.5 grades per unit of a and that of the certifier with diagnostics by 3.7 (Figure 2(a)). The certifier is cheaper than direct testing, scanning, the LCB + Box test of the lowest mean (Kaufmann et al., 2018) and LUCB_G witness search (Kano et al., 2019) at every δ from 10^{-1} to 10^{-8}. Its advantage over scanning comes from the diagnostics: in nine settings of a registered ablation, scanning's larger accepting boundary adds 3.2 to 10.9 grades, at most 21% of that advantage (Appendix L.1). The cheapest procedure switches where ka meets $\mathcal{H} + a$ (panel (b)), and every procedure with a guarantee errs in at most a fraction δ of runs for $\delta \in \{0.1, 0.05, 0.02, 0.01\}$ (Appendix L.1).

4.2 General networks: informative and certifying departures

Fix a legal baseline $B \supseteq S$, and a finite nonempty family $\mathcal{F}$ of legal ideals strictly containing B. Put $D_I = I \setminus B$, $Q = \bigcup_{I \in \mathcal{F}} D_I$, and $W = \bigcap_{s \in S} \mathrm{Desc}_G(s)$. A mastered node in W certifies the entire claim. Define

$$H_{\mathrm{diff}} = \min_{\lambda \geq 0} \left\{ \sum_{v \in Q} \lambda_v : \lambda(D_I) \geq 1(I \in \mathcal{F}) \right\}, \tag{12}$$

$$H_{\mathrm{cert}} = \min_{\lambda \geq 0} \left\{ \sum_{v \in Q \cap W} \lambda_v : \lambda(D_I \cap W) \geq 1(I \in \mathcal{F}) \right\}, \tag{13}$$

with infeasible value $+\infty$, and let $C_B = C_0(G[B], S)$. The first cover finds a difference from the baseline; the second finds a difference that suffices for acceptance. These LPs have polynomial size for an explicitly listed family, not necessarily for a succinct one.

Theorem 4.2 (General DAG discovery bounds)**.** *For fixed η, uniformly over these graphs and families and $0 < \delta \le 1/4$, the common evaluator cost $\mathfrak{D}$ satisfies*

$$c_\eta \min\{C_B a, H_{\mathrm{diff}} + a\} \le \mathfrak{D} \le C_\eta \min\{C_B a, H_{\mathrm{cert}} + a\}. \tag{14}$$

The upper bound stops in finite expectation at every legal state. The bounds match whenever $H_{\mathrm{diff}} = H_{\mathrm{cert}}$, in particular when $Q \subseteq W$.

The lower bound couples the unknown profile to B until the first departure; the dual of the difference cover makes any queried node a departure with probability at most $1/H_{\mathrm{diff}}$, and validity against nulls below B prevents accepting that positive baseline too early. For the upper bound, asymmetric checks sample from the certifying cover and race against verification of B (Appendix F). For $K_{k,t}$ and profiles consisting of the targets and one witness, $C_B = k$ and both covers equal t, giving $\Theta_\eta(\min\{ka, t + a\})$; for all profiles mastering exactly q of the t witnesses, the covers equal t/q, and this class realizes every finite hypergraph fractional transversal. Maximizing over *all* positive ideals instead forces direct testing of the maximal targets, so the favorable family is necessary to express a discovery gain (Proposition F.1).

4.3 Static structure does not determine discovery

The covers in (13) cannot by themselves resolve the diagnostic cost of Theorem 4.1. For $h \ge 2$, $t = k = 2^h$ and $m = 2h$, consider three balanced tables: one address bit repeated, the raw address written twice, and some systematic code. They have the same degrees, profile sizes and $C_B = t, C_0(G[I_v], S) = 1, H_{\mathrm{diff}} = 2, H_{\mathrm{cert}} = t$, but testing costs $\Theta_\eta(t + a)$, $\Theta_\eta(h \log(h + 1) + a)$ and $\Theta_\eta(h + a)$, respectively (Example F.2, which uses the budget law below).

The favorable profiles matter as well. Suppose every profile in the favorable family masters at least one witness, and the family includes all profiles that master every target, all $2m$ diagnostics and exactly one witness. Then, for every table and $0 < \delta \le 1/4$, the worst case cost is $\Theta_\eta(\min\{ka, t + a\})$ (Appendix F). Informative complementary failures are an efficiency assumption, not a consequence of the graph alone; Section 6.2 shows which measurement designs supply them.

5 Designing diagnostic capabilities

Fix $t = 2^h$ witnesses and allow at most m diagnostic pairs. Optimizing the table over this budget chooses both the available capabilities and their favorable profiles; it does not rewire a fixed population at unchanged states. The resulting optimum is the diagnostic resource law.

Theorem 5.1 (Optimal diagnostic resources)**.** *For $h \ge 1, m \ge 0$ integers, define*

$$\Psi(h, m) = \begin{cases} 2^{h-m} + m \log(m + 1), & m \le h, \\ h \log\left(2 + \dfrac{h}{m - h + 1}\right), & m > h. \end{cases}$$

For fixed $0 < \eta < 1/2$, uniformly over these budgets, $k \ge 1$ and $0 < \delta \le 1/4$, with the infima over tables with $t = 2^h$ rows and at most m diagnostic pairs,

$$\inf_E \mathcal{H}_\eta(E) = \Theta_\eta(\Psi(h, m)), \qquad \inf_E \mathfrak{D}_{E,k}(\delta) = \Theta_\eta(\min\{ka, \Psi(h, m) + a\}). \tag{15}$$

The infima permit unbalanced tables, duplicate rows and unused signatures. Balanced grouped signatures attain the first regime; balanced systematic codes attain the second and are constructible in expected polynomial time in the explicitly listed network size. If $k \ge t$, the testing rate simplifies to $\Theta_\eta(a + \Psi)$.

The all table converse. Fix a horizon q. The converse requires no balance or independence assumption on the table's rows. For every signature $z \in \{0, 1\}^m$, including unused ones, form a legal baseline that masters the targets and the diagnostics z but no witness. The baseline with signature

$E(v)$ and I_v differ only at c_v, so they give the same probability of first querying c_v by q. Diagnostic queries alone cannot distinguish profiles sharing a signature. Fix the policy's seed and a query index $n \le q$, and let $P_z(y)$ be the baseline law of the prefix before query n, padding stopped histories with a terminal symbol. If $a_v(y)$ indicates that the next query proposes witness v, then $\sum_v a_v(y) \le 1$, with all indicators zero after stopping. The average baseline probability of proposing the corresponding witness is

$$\frac{1}{t}\sum_{y,v} a_v(y) P_{E(v)}(y) \le \frac{1}{t}\sum_{y} \max_z P_z(y) = \frac{2^m}{t}\mathbb{E}_{\bar{P}} \max_z \mathbb{P}(Z = z \mid y), \tag{16}$$

where Z has an auxiliary uniform prior on the full cube and $\bar{P} = 2^{-m}\sum_z P_z$. For this fixed seed and prefix stage, $\log\bigl(\sum_y \max_z P_z(y)\bigr)$ is maximal leakage (Issa et al., 2019); the factor $1/t$ in (16) normalizes the witness average. Adaptive choices introduce no additional state dependent likelihood factor once the history is fixed, so the auxiliary posterior factors by coordinate. With n_j observations and posterior log odds L_j at bit j, $|L_j| \le \ell_\eta n_j$ and $\sum_j n_j \le q$, so

$$\max_z \mathbb{P}(Z = z \mid y) = \prod_j (1 + e^{-|L_j|})^{-1} \le \exp\left(-\tfrac{1}{2}\sum_j e^{-\ell_\eta n_j}\right) \le \exp\Bigl(-\tfrac{m}{2} e^{-\ell_\eta q/m}\Bigr).$$

The inequalities use $\log(1+x) \ge x/2$ on $[0,1]$ and Jensen. A union bound over the q baseline query steps, the equality of hit probabilities above, and integration over the seed give, for $m \ge 1$,

$$\frac{1}{t}\sum_v \mathbb{P}_{\mathbf{1}_{I_v}}\{\text{hit } c_v \text{ by } q\} \le \frac{q2^m}{t}\exp\Bigl(-\frac{m}{2}e^{-\ell_\eta q/m}\Bigr). \tag{17}$$

The product prior is auxiliary; the actual rows of the table need not be independent or uniformly distributed over signatures. The envelope has two readings. For $m \le h$, the prefactor $q2^m/t$ forces the ambiguity cost 2^{h-m} of the witnesses sharing a signature, and the exponential factor forces the reliability cost $m\log(m+1)$, up to constants depending on η, of learning every bit well enough to propose the right cell. For $m = h + b$, combining the envelope with the binary query tree bound yields $h\log(2 + h/(b+1))$. Appendix H gives uniform constants and handles padding a table up to a budget of at most m pairs.

Matching designs and their cost. The designs that attain Ψ mirror the two readings. For $m \le h$, estimate a grouped signature by repeated majority queries and then query its entire cell of 2^{h-m} witnesses. For $1 \le b = m - h \le h$, a random systematic code of length $h + b$ has distance $\Omega(\max\{1, b/\log(2 + (h+b)/b)\})$ with probability at least $1/2$, a Gilbert-type count (Gilbert, 1952); repeating each coordinate until its hard decision error is small relative to the code distance gives a constant success decoded witness proposal. For $m > 2h$, surplus pairs are left unused. The bounds hold for fixed noise, with constants depending only on η; they are not a noise uniform approximation ratio as $\eta \uparrow 1/2$.

At $m = h$ the order is $h\log(h+1)$; a linear surplus $b = \Theta(h)$ reduces it to h, while $b = o(h)$ leaves $\Psi/h \to \infty$, so linear query cost requires a constant factor surplus of pairs. The displayed Ψ is a rate, not a monotone finite cost curve. Table 1 in Appendix A places the converse next to the results it is closest to.

Measured. Figure 2(c) shows hitting budgets, the grades with which a policy hits the mastered witness with probability $3/4$ at every favorable profile of the symmetric channel. Majority reads of the identity table need 5 reads per address bit at $h = 2$ and 15 from $h = 12$ on; decoding a $[2h, h]$ code keeps about 8 from $h = 6$ to $h = 16$, hitting at $h = 16$ with 129 grades against 241. These budgets bound $\mathcal{H}$ from above; up to $h = 6$ scanning $3t/4$ witnesses has a smaller one (Table 5), and panel (b) compares complete procedures, which also pay to check each proposal. The surplus of pairs also pays on a language model. We made diagnostic reads of gpt-4o-mini mostly fresh by random selection clauses, stored 24 reads of every flag of 16 systems per h, and estimate success by resampling them, with an invalid answer (6.6% of reads) read as a fair coin fixed in advance. A random address stored as a systematic $[2h, h]$ code reaches success $3/4$ at every one of 16 addresses with 8 to 12 reads per address bit for $h \in \{4, 8, 12, 16\}$, while repeating the raw bits leaves 3 to 7 addresses below $3/4$ within 11 reads per bit, because rereading a persistently misread bit does not

repair it. At $h = 16$ and 96 reads the code succeeds with probability 0.821 and repetition with 0.414 (difference 0.41, 95% interval 0.32 to 0.50 over resamples of the systems), and coding still wins when invalid answers count as errors (Appendix L.6).

6 Failed rules and unknown error rates

The discovery laws above assume exact rules and a known symmetric channel. This section relaxes each assumption in turn: failed rules change only the confidence term, and a known margin leaves the optimized law unchanged.

6.1 Redundant witnesses make discovery robust to failed rules

A single witness gives $C_0 = 1$, but one failed rule raises its coefficient to k (Section 3.2). Redundancy restores the saving. Replace each witness c_v by w distinct leaves $C_v = \{c_{v,1}, \dots, c_{v,w}\}$, each keeping all its target and diagnostic edges, and call the graph $G_{E,w}$. The favorable profiles are $F_v = S \cup C_v \cup \{d_{j,E(v)_j} : j \in [m]\}$. The global class $\mathcal{X}_r(G_{E,w})$ allows at most r failed original edges in total; the per group class $\mathcal{Y}_r$ allows at most r failed edges into each group, so $\mathcal{X}_r \subseteq \mathcal{Y}_r$. Write $\mathfrak{D}_{\mathcal{X}}$ and $\mathfrak{D}_{\mathcal{Y}}$ for the common evaluator costs on these profiles with validity on the respective class.

Theorem 6.1 (Failed rules change only the confidence term)**.** *For fixed $0 < \eta < 1/2$, uniformly over tables E, integers $k, t, w \geq 1$, $m \geq 0$, $0 \leq r < w$ and $0 < \delta \leq 1/4$, with $R = w/(w-r)$,*

$$c_\eta \min\{ka, \mathcal{H}_\eta(E) + Ra\} \leq \mathfrak{D}_{\mathcal{X}} \leq \mathfrak{D}_{\mathcal{Y}} \leq C_\eta \min\{ka, \mathcal{H}_\eta(E) + Ra\}. \tag{18}$$

For $r \geq w$, both costs are $\Theta_\eta(ka)$. The upper bound stops in finite expectation at every state of $\mathcal{Y}_r$. Optimizing the table as in Theorem 5.1 replaces $\mathcal{H}_\eta(E)$ by $\Psi(h, m)$.

Deleting one target together with $w - r$ leaves of the active group gives a null with exactly r failed rules; averaging these information constraints forces either direct target evidence or group evidence with coefficient R. Deleting the whole active group gives a positive baseline, and the reduction of Theorem 4.1 applies unchanged, because on the favorable family a fresh query to any leaf simulates the original witness. For the upper bound, a hitting policy selects one group by an asymmetric screen, and only that group is validated, by $\lceil 2R \log(1/\epsilon) \rceil$ fresh blocks at uniformly drawn leaves. Under any null in $\mathcal{Y}_r$ the selected group has at least $w - r$ zero leaves, so validation passes with probability at most ϵ conditionally on the whole selection history. With $\mathcal{X}_r \subseteq \mathcal{Y}_r$, both laws follow (Appendix I). Efficiency is measured at intact active groups; validity holds everywhere. In simulation, validating a group without the factor accepts false claims in 627 of 1000 runs at $R = 4$, and with it in at most 18 (Appendix L.2).

6.2 Unknown error rates within a known margin

Fix $0 < \bar{\eta} < 1/2$. A response law belongs to $\mathcal{G}(\bar{\eta})$ if query n returns $Y_n \sim \mathrm{Bern}(p_n)$ given the past, with $p_n \geq 1 - \bar{\eta}$ when $x_{A_n} = 1$ and $p_n \leq \bar{\eta}$ when $x_{A_n} = 0$; the rate p_n is unknown and may depend on the node, the state and the past. The symmetric channel with parameter $\bar{\eta}$ is a member. Let $\mathfrak{D}^{\mathrm{gap}}_{E,k}$ require validity under every member and take the worst case cost over members, and let $\mathcal{H}^{\mathrm{gap}}_{\bar{\eta}}(E)$ be the hitting cost of a search using only diagnostics and witnesses that must succeed under every member, so $\mathcal{H}_{\bar{\eta}}(E) \leq \mathcal{H}^{\mathrm{gap}}_{\bar{\eta}}(E) \leq t$.

Theorem 6.2 (Channels with a known margin)**.** *For fixed $\bar{\eta}$, uniformly over E, $k, t \geq 1$, $m \geq 0$ and $0 < \delta \leq 1/4$,*

$$c_{\bar{\eta}} \min\{ka, \mathcal{H}_{\bar{\eta}}(E) + a\} \leq \mathfrak{D}^{\mathrm{gap}}_{E,k}(\delta) \leq C_{\bar{\eta}} \min\{ka, \mathcal{H}^{\mathrm{gap}}_{\bar{\eta}}(E) + a\}. \tag{19}$$

Over tables with $t = 2^h$ witnesses and at most m pairs, $\inf_E \mathcal{H}^{\mathrm{gap}}_{\bar{\eta}}(E) = \Theta_{\bar{\eta}}(\Psi(h, m))$ and $\inf_E \mathfrak{D}^{\mathrm{gap}}_{E,k}(\delta) = \Theta_{\bar{\eta}}(\min\{ka, \Psi(h, m) + a\})$.

The lower bound holds because the symmetric channel is a member. For the upper bound, use the reference BSC log score $L_n = \ell_{\bar{\eta}} \sum_{i=1}^{n} (2Y_i - 1)$ for a queried node. The processes e^{L_n} at label zero and e^{-L_n} at label one are nonnegative supermartingales. Ville's inequality keeps every check valid, and the majority votes of the designs obey Hoeffding's inequality in martingale form. For a fixed table the two sides differ only in the hitting term, by a factor at most logarithmic in $\mathcal{H}_{\bar{\eta}}(E)$ (Proposition J.2 in Appendix J); whether the logarithm can be removed is open.

Sampled outputs graded by exact checkers. Fix a probe and one deterministic contract per node: a target checks one component of the output, the two members of a diagnostic pair check incompatible properties of it, and a witness passes only if the targets, the diagnostics its row names and its own check all pass on the same output. A query draws a fresh output and returns one grade, so node v passes with a fixed rate p_v. Under a pass rate gap, $p_v \notin (\bar{\eta}, 1-\bar{\eta})$ at every node, the state $x_v = \mathbf{1}\{p_v \geq 1-\bar{\eta}\}$ obeys every rule and masters at most one member of each pair. A state with a mastered witness therefore masters exactly the diagnostics of its row: the complementary failures in (8) follow from the contract design. The responses form a member of $\mathcal{G}(\bar{\eta})$, so the procedures of Theorem 6.2 apply; Appendix J.2 writes out one such contract set for a program synthesis probe and shows where the gap and richness conditions enter. The symmetric channel is generally not such a sampled system; if every grade pattern satisfying the witness inclusions and diagnostic pair incompatibility is produced by some output, explicit sampled systems carry the converse, and both bounds hold when validity is required only for sampled systems (Proposition J.4). Mastery is then a pass rate, not a latent skill, and the rules come from declared contract inclusions rather than from correlations in generated answers.

Language models as sampled systems. Differences in knowledge realize this design: each of 16 systems is gpt-4o-mini with a knowledge base for one office: a random project codename is its witness, coded facts for four binary attributes are its diagnostic pairs, and one probe of 40 items is graded by exact contracts whose inclusions hold on every reply. At $\bar{\eta} = 1/4$, 14 of the 16 systems realize their favorable profile at temperature 1.0 and all 16 at 0.8. A replay answers each query with a uniformly drawn stored grade of its node. Averaged over the 14 favorable systems at $t = 16$, 1000 replays each, it certifies $k = 16$ targets at $\delta = 0.05$ with 32.2 grades when the four diagnostics are read first, against 67.8 for scanning and 66.4 for direct testing, with at most 5 errors in 14000 runs (Appendix L.4). The diagnostics are read without error here, so this population supports the additive law and the address reading, not the logarithmic factor of Theorem 5.1.

7 Where the assumptions stop

The favorable population is a property of the model: with gpt-4.1-mini or gpt-4.1-nano the design of Section 6.2 yields no favorable profile, since each misanswers one or two basic items. Fresh noise is an assumption, and computed answers violate it. When a diagnostic must be computed from the knowledge base, here divisibility of a four digit code by 3, gpt-4o-mini gives nearly the same answer on every sample and misreads 31 of 64 diagnostics. A system with a misread bit masters its witness but not a diagnostic of that witness's row, a state the rules forbid, so 14 of the 16 systems lie outside the class the laws cover; on them the certifier stays valid but costs 129.8 grades against 65.4 for direct testing, because only the raced direct test ends the run. Fresh noise is different: flipping each diagnostic grade of the knowledge configured systems with probability 0.2 raises the certifier's cost from 32.0 to 56.7 grades, still below scanning at 67.6 (Appendix L.5).

Rules must come from contract design, not from difficulty labels. On public grades of 13 open models on 128 MATH problems (Hendrycks et al., 2021; Brown et al., 2024), the rule that every Level 4 or 5 problem requires every Level 1 problem fails wherever it matters: every model that masters a hard problem fails some Level 1 problem, and a certifier that trusts the rule accepts the strongest model's false claim in 1.4% of runs at $\delta = 0.05$, below δ but not zero (Appendix L.7). Grading one sample against every target saves samples, not grades: at $k = 16$ on the knowledge configured systems, direct testing uses 6.4 samples but 66.4 grades.

Scope. The graph is supplied logical information, not learned from correlations. Optimizing diagnostics supplies new capabilities, whose construction cost is outside the query metric. The private witness law does not characterize all overlapping certificates, and its policy representation is exponential. Fresh observations are substantive: with one stored corrupted label at a single node, rereads leave minimax error η. An estimated margin adds the probability that it is violated (Corollary J.3); no calibration method is proposed. The model neither estimates benchmark average performance, as active testing does (Kossen et al., 2021), nor covers learned rules, batch grading of one sample, or general partial knowledge spaces.

AI USE STATEMENT

Large language models (LLMs) were used extensively to develop the mathematical model, claims, proof drafts and manuscript from the authors' research idea, and to reconstruct and critique proofs, check citations, implement finite calculations, simulate reviews and revise the exposition. At the authors' request, an LLM also designed, preregistered, implemented, ran and analyzed the experiments and drafted their text. LLMs were also the measured systems: gpt-4o-mini answered HumanEval tasks and synthetic knowledge probes, gpt-4.1-mini and gpt-4.1-nano answered the same knowledge probes, all graded by exact checkers, and public grades of 13 open models on MATH problems were reanalyzed.

Finite checks cover only their stated domains and are not independent human verification. The research idea is the authors', and the authors take full responsibility for the manuscript and its claims.

REFERENCES

Michael Ben-Or and Avinatan Hassidim. The bayesian learner is optimal for noisy binary search (and pretty good for quantum as well). In *2008 49th Annual IEEE Symposium on Foundations of Computer Science*, pp. 221–230. IEEE, 2008.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Ying Cheng. When cognitive diagnosis meets computerized adaptive testing: Cd-cat. *Psychometrika*, 74(4):619–632, 2009.

Herman Chernoff. Sequential design of experiments. In *Breakthroughs in Statistics: Methodology and Distribution*, pp. 345–360. Springer, 1992.

Hye Won Chung, Brian M Sadler, Lizhong Zheng, and Alfred O Hero. Unequal error protection querying policies for the noisy 20 questions problem. *IEEE Transactions on Information Theory*, 64(2):1105–1131, 2017.

Yuval Dagan, Yuval Filmus, Ariel Gabizon, and Shay Moran. Twenty (simple) questions. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 9–21, 2017.

Peter Damaschke. Deterministic versus randomized adaptive test cover. *Theoretical Computer Science*, 653:42–52, 2016.

Rémy Degenne and Wouter M Koolen. Pure exploration with multiple correct answers. *Advances in Neural Information Processing Systems*, 32, 2019.

Rémy Degenne, Wouter M Koolen, and Pierre Ménard. Non-asymptotic pure exploration by solving games. *Advances in Neural Information Processing Systems*, 32, 2019.

Jean-Paul Doignon and Jean-Claude Falmagne. Knowledge spaces and learning spaces. *arXiv preprint arXiv:1511.06757*, 54, 2015.

Kefan Dong and Tengyu Ma. Asymptotic instance-optimal algorithms for interactive decision making. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=oGVu9spZaJJ.

William Evans and Nicholas Pippenger. Average-case lower bounds for noisy boolean decision trees. *SIAM Journal on Computing*, 28(2):433–446, 1998.

Uriel Feige, Prabhakar Raghavan, David Peleg, and Eli Upfal. Computing with noisy information. *SIAM Journal on Computing*, 23(5):1001–1018, 1994.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pp. 998–1027. PMLR, 2016.

Edgar N Gilbert. A comparison of signalling alphabets. *The Bell system technical journal*, 31(3): 504–522, 1952.

Justin Gilmer, Michael Saks, and Srikanth Srinivasan. Composition limits and separating examples for some boolean function complexity measures. *Combinatorica*, 36(3):265–311, 2016.

Tomer Grossman, Ilan Komargodski, and Moni Naor. Instance complexity and unlabeled certificates in the decision tree model. In *11th Innovations in Theoretical Computer Science Conference (ITCS 2020)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2020.

Martin Grötschel, László Lovász, and Alexander Schrijver. The ellipsoid method and its consequences in combinatorial optimization. *Combinatorica*, 1(2):169–197, 1981.

Yuzhou Gu, Xin Li, and Yinzhan Xu. Tight bounds for noisy computation of high-influence functions, connectivity, and threshold. *arXiv preprint arXiv:2502.04632*, 2025.

Yuzhou Gu, Xin Li, and Yinzhan Xu. A unified lower bound on the noisy query complexity of boolean functions. In Steve Hanneke and Tor Lattimore (eds.), *Proceedings of Thirty Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 2940–2962. PMLR, 29 Jun–03 Jul 2026. URL https://proceedings.mlr.press/v336/gu26a.html.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL https://openreview.net/forum?id=7Bywt2mQsCe.

Ibrahim Issa, Aaron B Wagner, and Sudeep Kamath. An operational approach to information leakage. *IEEE Transactions on Information Theory*, 66(3):1625–1657, 2019.

Brian W Junker and Klaas Sijtsma. Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25(3): 258–272, 2001.

Hideaki Kano, Junya Honda, Kentaro Sakamaki, Kentaro Matsuura, Atsuyoshi Nakamura, and Masashi Sugiyama. Good arm identification via bandit feedback. *Machine Learning*, 108(5): 721–745, 2019.

Richard M Karp and Robert Kleinberg. Noisy binary search and its applications. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pp. 881–890, 2007.

Emilie Kaufmann, Wouter Koolen, and Aurélien Garivier. Sequential test for the lowest mean: From thompson to murphy sampling. *Advances in Neural Information Processing Systems*, 31, 2018.

Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Tom Rainforth. Active testing: Sample-efficient model evaluation. In *International Conference on Machine Learning*, pp. 5753–5763. PMLR, 2021.

Jacqueline P Leighton, Mark J Gierl, and Stephen M Hunka. The attribute hierarchy method for cognitive assessment: A variation on tatsuoka's rule-space approach. *Journal of educational measurement*, 41(3):205–237, 2004.

ZITIAN LI and Wang Chi Cheung. Near optimal non-asymptotic sample complexity of 1-identification. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=nymm7XlXvZ.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation. *Advances in neural information processing systems*, 36:21558–21572, 2023.

Matthew L Malloy, Gongguo Tang, and Robert D Nowak. The sample complexity of search over multiple populations. *IEEE transactions on information theory*, 59(8):5039–5050, 2013.

Mohammad Naghshvar and Tara Javidi. Active sequential hypothesis testing. *The Annals of Statistics*, 41(6), December 2013. ISSN 0090-5364. doi: 10.1214/13-aos1144. URL http://dx.doi.org/10.1214/13-AOS1144.

Hirofumi Ota, Naoto Iwase, Yuki Ichihara, Junpei Komiyama, and Masaaki Imaizumi. Cite: Anytime-valid statistical inference in llm self-consistency. *arXiv preprint arXiv:2605.05873*, 2026.

Shuyin Ouyang, Jie M Zhang, Mark Harman, and Meng Wang. An empirical study of the non-determinism of chatgpt in code generation. *ACM Transactions on Software Engineering and Methodology*, 34(2):1–28, 2025.

Rüdiger Reischuk and Bernd Schmeltz. Reliable computation with noisy circuits and decision trees-a general n log n lower bound. In *[1991] Proceedings 32nd Annual Symposium of Foundations of Computer Science*, pp. 602–611. IEEE, 1991.

Martin Riddell, Ansong Ni, and Arman Cohan. Quantifying contamination in evaluating code generation capabilities of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14116–14137, 2024.

Zhijian Zhou, Zesheng Ye, Zhaorun Chen, Bo Li, and Feng Liu. Celeus: Certifiable and efficient llm evaluation via e-processes. *arXiv preprint arXiv:2606.20820*, 2026.

A Related work

Knowledge structure and certificates. Knowledge-space theory supplies the state family and surmise relation framework (Doignon & Falmagne, 2015); attribute hierarchies likewise model mastery through prerequisite ideals (Leighton et al., 2004). Cognitive diagnosis and its adaptive tests estimate whole attribute patterns (Junker & Sijtsma, 2001; Cheng, 2009); our question concerns one fixed conjunction. Proposition C.1 identifies what the classical pairwise envelope retains for this question at full mastery, with an explicit failure at partial profiles. Fractional certificates and their randomized verification interpretation are established (Gilmer et al., 2016); the contribution here is the low cut alternative family, its structural evaluations, and the uniform noisy verifier under the graph promise.

Sequential testing and search. The lower bound specializes the standard KL allocation principle (Chernoff, 1992; Garivier & Kaufmann, 2016): the analogous max–min information rate for our conjunction answer map is κ/C_r. Garivier–Kaufmann's theorem itself concerns a changed best arm, not this hypothesis. Without a graph, the question is whether every noisy target is above threshold, a pure exploration problem with multiple correct answers (Degenne & Koolen, 2019; LI & Cheung, 2025); the graph and the favorable profiles are what make discovery possible here. Its lowest mean form has sequential tests with matching lower bounds (Kaufmann et al., 2018), and good arm search returns one arm above a threshold (Kano et al., 2019); Section 4.1 uses both as baselines. Anytime valid evaluation of language models certifies an aggregate benchmark score (Zhou et al., 2026) or the modal answer to one prompt (Ota et al., 2026); the claims here are conjunctions of node states under supplied rules. Low-confidence estimation followed by optimized validation also appears in regret minimization (Dong & Ma, 2023). Active sequential hypothesis testing and search over populations separate acquisition from reliability costs (Naghshvar & Javidi, 2013; Malloy et al., 2013), while instance complexity already distinguishes cheap supplied evidence from expensive search (Grossman et al., 2020). Our private witness reduction adds full class validity and the direct testing alternative to an unrecognized hitting objective.

Fixed diagnostics and coding. Restricted question dictionaries have a noiseless identification theory (Dagan et al., 2017); noisy search and computation give classical amplification bounds (Feige et al., 1994; Gu et al., 2025). Reading n noisy bits, or computing parity, costs order $n \log n$ (Feige et al., 1994); the general form, order $I \log I$ for every function of total influence I (Gu et al., 2026), was preceded by bounds for nonadaptive algorithms and for random functions (Reischuk & Schmeltz, 1991; Evans & Pippenger, 1998). Coding also appears in noisy twenty questions (Chung et al., 2017). Unlike adaptively chosen bipartitions (Karp & Kleinberg, 2007; Ben-Or & Hassidim, 2008), even the best fixed table of h diagnostic pairs needs order $h \log(h+1)$ queries even to support unrecognized witness hits. Lower bounds for computing a function or recovering one signature do not give this: the likelihood envelope bounds multiple adaptive proposals for every table. The all table converse and its testing consequence are the increment.

Table 1 places the resource law next to its closest known results.

B Notation

Throughout, Θ_η, O_η, Ω_η hide positive constants depending only on η; $\mathbf{1}$ is the all one state and e_v the unit vector at v.

C Union closure and the scope of pairwise knowledge

We first prove Proposition C.1, then give two examples showing why its assumptions and reference state matter. For the classical correspondence between quasi ordinal spaces and preorders, we use the exposition of Doignon & Falmagne (2015, Definition 9 and Theorem 10), which attributes the correspondence to Birkhoff (1937). Our use of that correspondence concerns the certificate constraints for a particular testing question.

Write $s \preceq_{\mathcal{K}} v$ when every state mastering v also masters s, and let $\mathcal{K}^{\mathrm{pair}}$ contain all sets obeying these pairwise implications, the quasi ordinal envelope of the knowledge space. Fix a nonempty

Setting	Known	Here
Noisy search with adaptively chosen bipartitions (Karp & Kleinberg, 2007; Ben-Or & Hassidim, 2008)	order $\log_2 t$ queries locate one of t items	a fixed table of $h = \log_2 t$ pairs needs order $h \log(h+1)$ even to hit, unrecognized
Noisy computation (Feige et al., 1994; Gu et al., 2026)	$\Theta(n \log n)$ for parity and for reading n bits; $\Omega(I \log I)$ for every function of total influence I	a converse for hitting rather than computing, for every table, with multiple adaptive proposals, duplicate rows and unused signatures
Twenty questions with a restricted question set, noiseless (Dagan et al., 2017), and noisy twenty questions with coding (Chung et al., 2017)	restricted-query identification without noise; nonadaptive block coding for noisy estimation	surplus b pairs give $h \log(2 + h/(b+1))$; order h needs a constant factor surplus
Fixed confidence testing (Chernoff, 1992; Garivier & Kaufmann, 2016)	KL allocation gives the confidence cost	the optimal rate is the smaller of direct testing ka and discovery plus a (Theorem 4.1)

Table 1: The resource law next to its closest known results. The first two rows compare rates at fixed noise and fixed constant error.

Symbol	Meaning
V, $x \in \{0,1\}^V$	capabilities; mastery state, $x = \mathbf{1}_I$ for $I \subseteq V$
S, $k = \lvert S \rvert$	target set; number of targets
$G = (V, \mathcal{E})$, $\mathcal{K}_0(G)$	supplied DAG of original edges; its order ideals (exact legal states)
$B_G(x)$, $\mathcal{X}_r(G)$	number of failed original edges; validity class with at most r failures
η, ℓ_η, κ	flip probability; $\log \frac{1-\eta}{\eta}$; $\mathrm{kl}(1-\eta, \eta) = (1-2\eta)\ell_\eta$
δ, a, a_δ	error level; $\log(1/\delta)$; $\log(2/\delta)$
τ, $N_v(\tau)$	stopping time; number of queries at node v
$\mathfrak{T}_r(G, S, \delta; I)$	optimal cost at a supplied legal ideal I (5)
$\mathfrak{D}_r(G, S; \mathcal{F}, \delta)$	optimal common evaluator cost over a favorable family (6)
$\partial^+_G(D)$, $\mathcal{D}_r(G, S)$	outgoing edge boundary of D; low cut missing sets meeting S
$C_r(G, S)$	fractional cover of $\mathcal{D}_r(G, S)$ (9); $C_r(G[I], S)$ on the induced graph
$\mathrm{Desc}_G(s)$	descendants of s, including s
B, $\mathcal{F}$, D_I, Q, W	baseline ideal; favorable family; departures $I \setminus B$; their union; common descendants of S
H_{diff}, H_{cert}, C_B	difference and certifying covers (13); $C_0(G[B], S)$
$E : [t] \to \{0,1\}^m$, $t = 2^h$	diagnostic table; witnesses; pairs; address length
u_i, $d_{j,b}$, c_v, I_v	target roots; diagnostic roots; witness leaves; favorable profiles (8)
$\mathfrak{D}_{E,k}(\delta)$	family cost of the private witness network at $r = 0$
$\mathcal{H}_\eta(E)$	unrecognized hitting cost (Definition 2.1)
$\Psi(h, m)$	diagnostic resource law (Theorem 5.1)
w, C_v, F_v, R	leaves per witness group; group; redundant favorable profile; $w/(w-r)$
$\mathcal{X}_r(G_{E,w})$, $\mathcal{Y}_r$, $\mathfrak{D}_\mathcal{X}$, $\mathfrak{D}_\mathcal{Y}$	global and per group failure classes; their family costs
$\bar{\eta}$, $\mathcal{G}(\bar{\eta})$	known margin; response laws with rates outside $(\bar{\eta}, 1-\bar{\eta})$
$\mathcal{H}^{\mathrm{gap}}_{\bar{\eta}}(E)$, $\mathfrak{D}^{\mathrm{gap}}_{E,k}$	hitting cost and family cost required under every member
p_v, $g_v(\omega)$	pass rate of contract v; its grade on output ω
Small_j	sum of the j smallest entries of a vector

target set S; a fractional certificate at the reference V assigns nonnegative weights to coordinates so that every opposite answer difference $V \setminus J$, $J \in \mathcal{K}$, $S \not\subseteq J$, has weight at least one.

Proposition C.1 (Full mastery certificates survive the pairwise envelope)**.** *Let $\mathcal{K} \subseteq 2^V$ contain $\varnothing$ and V and be closed under unions. For every nonempty $S \subseteq V$, $\mathcal{K}$ and $\mathcal{K}^{\mathrm{pair}}$ have identical logical certificates and identical fractional certificate feasible regions at the reference V. With*

$$U_s = \bigcup_{\substack{J \in \mathcal{K} \\ s \notin J}} J, \qquad D_s = V \setminus U_s = \{v : s \preceq_\mathcal{K} v\},$$

their common fractional value is

$$C_{\mathcal{K}}(S) = C_{\mathcal{K}^{\mathrm{pair}}}(S) = \min_{\lambda \geq 0} \left\{ \sum_{v \in V} \lambda_v : \lambda(D_s) \geq 1 \quad (s \in S) \right\}. \tag{20}$$

Union closure makes U_s a genuine legal state missing s that contains every other such state, so the sets D_s are necessary and sufficient certificate constraints. Collapsing mutually implying capabilities gives a partial order and hence a DAG, without changing query costs. The reference state matters: for $V = S \cup \{b, c\}$, let mastery of c require either b or all of S; at $I^\star = S \cup \{c\}$ the observations $c = 1, b = 0$ certify every target, so for $k \geq 3$ both certificate costs are two, while the envelope is the full cube and both costs are k (Example C.2).

C.1 Certificate constraints for a general state family

Let $\mathcal{F} \subseteq 2^V$ be a finite family containing a positive reference state I, so that $S \subseteq I$. Define its opposite answer difference family by

$$\mathcal{B}_{\mathcal{F}}(I, S) = \{I \triangle J : J \in \mathcal{F},\ S \not\subseteq J\}. \tag{21}$$

A logical certificate at I is a set of queried coordinates $Q \subseteq V$ whose values agree with I and exclude every opposite answer state. Thus it is precisely a hitting set for $\mathcal{B}_{\mathcal{F}}(I, S)$. Denote its minimum size by $\mathsf{h}_{\mathcal{F}}(I, S)$, and define

$$\mathsf{C}_{\mathcal{F}}(I, S) = \min_{\lambda \geq 0} \left\{ \sum_{v \in V} \lambda_v : \sum_{v \in B} \lambda_v \geq 1 \text{ for every } B \in \mathcal{B}_{\mathcal{F}}(I, S) \right\}. \tag{22}$$

These definitions permit both positive and negative coordinate values in a certificate. At $I = V$, they reduce to the all positive certificates used in the main text. In particular, $\mathsf{C}_{\mathcal{K}}(V, S) = C_{\mathcal{K}}(S)$.

C.2 Proof of the envelope invariance

Suppose that $\mathcal{K}$ contains $\varnothing, V$ and is closed under unions. Define

$$s \preceq_{\mathcal{K}} v \quad \Longleftrightarrow \quad \big[v \in I \Longrightarrow s \in I \text{ for every } I \in \mathcal{K}\big],$$

and let

$$\overline{\mathcal{K}} = \mathcal{K}^{\mathrm{pair}} = \{J \subseteq V : v \in J,\ s \preceq_{\mathcal{K}} v \Longrightarrow s \in J\}.$$

The relation is reflexive and transitive. The envelope $\overline{\mathcal{K}}$ contains $\mathcal{K}$ and is closed under both unions and intersections.

Proof of Proposition C.1. For each $s \in V$, define

$$U_s = \bigcup \{I \in \mathcal{K} : s \notin I\}, \qquad D_s = V \setminus U_s.$$

The union is over a nonempty finite family, since it includes $\varnothing$. Therefore $U_s \in \mathcal{K}$, and $s \notin U_s$. For $s \in S$, D_s is itself the difference set of a genuine opposite answer state. Moreover,

$$v \in D_s \quad \Longleftrightarrow \quad \text{no } I \in \mathcal{K} \text{ contains } v \text{ but omits } s \quad \Longleftrightarrow \quad s \preceq_{\mathcal{K}} v.$$

If $J \in \mathcal{K}$ omits a target s, then $J \subseteq U_s$. Consequently $V \setminus J \supseteq D_s$. It follows that the constraints generated by the sets D_s, $s \in S$, imply every other opposite answer constraint, both for hitting sets and for nonnegative fractional weights. Conversely, each of those constraints is necessary because U_s belongs to the original family.

The same sets generate the constraints for $\overline{\mathcal{K}}$. Indeed, $U_s \in \overline{\mathcal{K}}$ because $\mathcal{K} \subseteq \overline{\mathcal{K}}$. If $J \in \overline{\mathcal{K}}$ omits s, it cannot contain any $v \in D_s$: such a v would entail s. Hence $J \subseteq U_s$, exactly as before.

Thus, in both families, the logical certificates are precisely

$$\{Q \subseteq V : Q \cap D_s \neq \varnothing \text{ for every } s \in S\},$$

and the fractional feasible region is precisely

$$\{\lambda \geq 0 : \lambda(D_s) \geq 1 \text{ for every } s \in S\}.$$

This proves equality of the logical certificate families, of their minimum sizes, and of the fractional optima. In fact, equality of the feasible regions also preserves any nonnegative weighted query cost objective. □

The envelope is the intersection closure of $\mathcal{K}$. For completeness, if $J \in \overline{\mathcal{K}}$, the argument above gives

$$J = \bigcap_{s \notin J} U_s,$$

where the empty intersection is V. Thus every envelope state is an intersection of original states. Conversely, intersections preserve all the pairwise implications. Since $\mathcal{K}$ is union closed and V is finite, its intersection closure is also union closed by distributivity. This identifies $\overline{\mathcal{K}}$ as the smallest quasi ordinal space containing $\mathcal{K}$.

The preorder may have mutually implying items. Collapsing the equivalence classes gives a partial order and hence a DAG. For the identical unit cost BSC channels considered here, queries to items in one class can all be performed at a single representative, with fresh noise at every query. Fractional weights within a class may likewise be summed. This observation justifies the DAG interpretation at zero rule error budget without assuming that every original knowledge space is discriminative.

The proposition equates certificate constraints. The information lower bound of Appendix D.3 applies to the genuine alternatives U_s in $\mathcal{K}$. Conversely, a verifier valid on the pairwise envelope is valid on its subset $\mathcal{K}$; after collapsing mutually implying items, Theorem 3.1 supplies that verifier. The two classes therefore have the same nonasymptotic cost order and the same fixed family high confidence constant at complete mastery. This does not by itself equate the exact finite confidence optimal stopping time distributions or expected costs for the two different promise classes.

C.3 Higher order knowledge at a partial reference

The full mastery reference is essential. A union closed knowledge space can contain useful information that is absent from its pairwise envelope when the reference also contains unmastered capabilities.

Example C.2 (Negative evidence can make alternative prerequisites useful). Let $V = S \cup \{b, c\}$, with $|S| = k \geq 3$, and define

$$\mathcal{K} = \{J \subseteq V : c \notin J \text{ or } b \in J \text{ or } S \subseteq J\}. \tag{23}$$

In words, mastery of c requires either b or all the target capabilities. At the legal positive reference $I^\star = S \cup \{c\}$, one has

$$\mathsf{h}_{\mathcal{K}}(I^\star, S) = \mathsf{C}_{\mathcal{K}}(I^\star, S) = 2, \qquad \mathsf{h}_{\overline{\mathcal{K}}}(I^\star, S) = \mathsf{C}_{\overline{\mathcal{K}}}(I^\star, S) = k. \tag{24}$$

Proof. The family contains $\varnothing, V$. If a union of its states contains c, then one of those states contains c, and that state already contains either b or all of S. The union retains that witness. Hence $\mathcal{K}$ is union closed.

There are no nontrivial pairwise implications. Each singleton $\{s\}$, $s \in S$, and the singleton $\{b\}$ is legal. For c, the legal state $\{b, c\}$ omits every target, while the legal state $S \cup \{c\}$ omits b. Therefore $\overline{\mathcal{K}} = 2^V$.

At $I^\star$, the observed values $c = 1$ and $b = 0$ force all of S to be present by (23). Thus $\{b, c\}$ is a two coordinate logical certificate. To compute the fractional optimum, let J be a bad state and choose $s \in S \setminus J$. If $c \notin J$, its difference set from $I^\star$ contains $\{s, c\}$. If $c \in J$, legality and $S \not\subseteq J$ force $b \in J$, so the difference set contains $\{s, b\}$. Every such pair is itself realized:

$$J = S \setminus \{s\} \quad \text{or} \quad J = (S \setminus \{s\}) \cup \{b, c\},$$

respectively. The fractional constraints therefore reduce exactly to

$$\lambda_s + \lambda_c \geq 1, \qquad \lambda_s + \lambda_b \geq 1, \qquad s \in S.$$

Writing $\alpha = \min\{\lambda_b, \lambda_c\}$, every feasible allocation has total weight at least

$$2\alpha + k \max\{0, 1 - \alpha\} \geq 2.$$

Equality is achieved by $\lambda_b = \lambda_c = 1$ and $\lambda_s = 0$ for $s \in S$. Thus the fractional optimum is two, and the exhibited logical certificate shows that the integer optimum is two as well.

In the full cube envelope, flipping only a single target from $I^\star$ yields a legal bad state. The fractional constraints consequently require $\lambda_s \geq 1$ for every $s \in S$, and the logical constraints require every target to be queried. Direct target evidence attains both lower bounds, proving (24). □

At complete mastery, the same example has cost parameter k in both families: a state omitting just one target is allowed because b is present. Thus it is the combination of a positive observation at c and a negative observation at b, absent from a full mastery certificate, that makes the higher order rule useful.

C.4 Why union closure cannot be dropped

Example C.3 (Joint positive evidence beyond the envelope). For $n = |V| \geq 3$, let

$$\mathcal{F} = \{\varnothing\} \cup \{\{v\} : v \in V\} \cup \{V\}, \qquad S = V.$$

Then $\mathcal{F}$ is not union closed and has no nontrivial pairwise implications. Its envelope is 2^V, but

$$\mathsf{h}_{\mathcal{F}}(V, V) = 2, \qquad \mathsf{C}_{\mathcal{F}}(V, V) = \frac{n}{n-1},$$

whereas both quantities for the envelope equal n.

Proof. Every singleton is a state, so it witnesses the failure of each proposed nontrivial pairwise implication. On the other hand, two distinct mastered coordinates exclude every proper state of $\mathcal{F}$, and one coordinate cannot do so. This gives the integer optimum. The inclusion minimal bad difference sets are $V \setminus \{v\}$, $v \in V$. Summing their fractional constraints gives

$$(n-1) \sum_{v \in V} \lambda_v \geq n.$$

The constant allocation $\lambda_v = 1/(n-1)$ attains this bound. In the full cube envelope, the singleton missing sets require all n unit weights and all n queried coordinates. □

The examples separate two boundaries: families that are not union closed may gain from joint positive evidence even at complete mastery, while union closed families may gain from higher order rules at references containing negative coordinates. Neither possibility is covered by the full mastery, zero rule error envelope reduction.

D Statistical proofs

Throughout this appendix, $p_1 = 1 - \eta$, $p_0 = \eta$, $\ell_\eta = \log(p_1/p_0)$, and $\kappa = \mathrm{kl}(p_1, p_0) = (1 - 2\eta)\ell_\eta$, where $0 < \eta < 1/2$. All logarithms are natural. Write $d = |V|$, $k = |S| \geq 1$, and $a_\delta = \log(2/\delta)$, with $0 < \delta < 1/2$. Each new observation has independent channel noise, conditionally on the queried node and the entire past. The random choices made by an algorithm are independent of this noise. In particular, revisiting a node produces a new observation. These assumptions also apply whenever a new stage of an algorithm is started.

We first prove Theorem 3.1 at the all one reference $I = V$. Appendix E.4 then transfers both bounds and the exact constant to a supplied legal ideal.

D.1 A one node sequential test

We first record the sequential testing facts used by the constructions. Their asymmetric use in the verification algorithm is related to asymmetric bit checks in noisy query complexity Gu et al. (2025).

Lemma D.1 (One node test). *Fix $u, b > 0$. Observe independent $Y_i \sim \text{Bernoulli}(p_x)$, where $x \in \{0, 1\}$, and define*

$$L_n = \sum_{i=1}^{n} \log \frac{p_1^{Y_i}(1-p_1)^{1-Y_i}}{p_0^{Y_i}(1-p_0)^{1-Y_i}} = \ell_\eta \sum_{i=1}^{n} (2Y_i - 1), \qquad T = \inf\{n \geq 1 : L_n \notin (-b, u)\}.$$

Return 1 on an upper exit and 0 on a lower exit. Under both values of x, T has finite expectation. Moreover,

$$\mathbb{P}_0(L_T \geq u) \leq e^{-u}, \qquad \mathbb{P}_1(L_T \leq -b) \leq e^{-b}, \tag{25}$$

$$\mathbb{E}_1 T \leq \frac{u + \ell_\eta}{\kappa}, \qquad \mathbb{E}_0 T \leq \frac{b + \ell_\eta}{\kappa}. \tag{26}$$

Proof. Choose an integer h such that $h\ell_\eta > u + b$. From any point in $(-b, u)$, h successive observations equal to 1 force an upper exit within these h observations. Each observation equals 1 with probability at least η, under either state. Independent increments therefore give

$$\mathbb{P}_x(T > jh) \leq (1 - \eta^h)^j, \qquad j = 0, 1, \ldots.$$

Consequently $\mathbb{E}_x T \leq h/\eta^h < \infty$. This bound is used only to justify integrability, not to establish the claimed query complexity.

Under $x = 0$, $\exp(L_n)$ is a nonnegative martingale starting at 1. For every integer N, stopping at the first crossing of e^u or at N shows that the probability of an upper crossing by time N is at most e^{-u}. Letting $N \to \infty$ gives the first error bound. Under $x = 1$, the same argument applied to the nonnegative martingale $\exp(-L_n)$ gives the second error bound.

The increments of L_n have mean κ under $x = 1$ and $-\kappa$ under $x = 0$. The bounded stopping identity at $T \wedge N$ gives

$$\mathbb{E}_1 L_{T \wedge N} = \kappa \mathbb{E}_1 (T \wedge N), \qquad \mathbb{E}_0 L_{T \wedge N} = -\kappa \mathbb{E}_0 (T \wedge N).$$

At every stopped or unstopped time in these displays, the likelihood sum lies in $[-b - \ell_\eta, u + \ell_\eta]$. Dominated convergence on the left and monotone convergence on the right therefore give Wald's identities

$$\mathbb{E}_1 L_T = \kappa \mathbb{E}_1 T, \qquad -\mathbb{E}_0 L_T = \kappa \mathbb{E}_0 T.$$

Finally, $L_T \leq u + \ell_\eta$ and $-L_T \leq b + \ell_\eta$ pointwise. These inequalities imply (26). □

D.2 A DIRECT TEST FOR THE TARGET CONJUNCTION

The following construction specifies the fallback used below. It is valid on the entire binary cube and requires no graph assumptions.

Lemma D.2 (Direct target fallback). *Fix any deterministic ordering of S. Test the targets in that order, using the test of Lemma D.1 at each node with thresholds*

$$u = a_\delta = \log(2/\delta), \qquad b = \log(2k/\delta).$$

Return 0 as soon as any target test returns 0; return 1 if all k tests return 1. Under every $x \in \{0, 1\}^d$, the procedure has finite expected duration and each type of error has probability at most $\delta/2$. If all target coordinates equal 1, its expected duration T_{dir} satisfies

$$\mathbb{E}_x T_{\text{dir}} \leq \frac{k(a_\delta + \ell_\eta)}{\kappa}. \tag{27}$$

The same error and expectation bounds hold conditionally on an arbitrary earlier history if the fallback uses fresh observations.

Proof. One can pregenerate an independent observation sequence for the test of each target; sequences associated with unreached targets are never used. This representation is equivalent in law to querying fresh observations as the procedure runs. Its duration is bounded by the sum of the k potential one node durations, so Lemma D.1 proves finite expectation at every state.

If at least one target is missing, fix such a target s. Returning 1 requires that the potential test of s return 1. Its probability is at most $e^{-a_\delta} = \delta/2$. If all targets are mastered, returning 0 requires a lower exit in at least one of the k potential tests. A union bound gives probability at most $k \exp[-\log(2k/\delta)] = \delta/2$.

When all targets are mastered, the expected duration of each potential test is at most $(a_\delta + \ell_\eta)/\kappa$ by Lemma D.1. Summing proves (27). Finally, conditionally on an earlier history, the fresh observation sequences have the same product distribution at the fixed state x. Every preceding argument therefore remains valid under this conditioning. □

D.3 The information lower bound

Proof of the lower bound in Theorem 3.1. Let $\mathcal{A}$ be any uniformly δ-correct procedure on $\mathcal{X}_r(G)$, with stopping time τ and output $\widehat{h} \in \{0, 1\}$. If $\mathbb{E}_{\mathbf{1}}\tau = \infty$, the lower bound is immediate; suppose this expectation is finite. For any $D \in \mathcal{D}_r(G, S)$, set $x^D = \mathbf{1} - \mathbf{1}_D$. The low cut representation gives $x^D \in \mathcal{X}_r(G) \cap H_0$, whereas $\mathbf{1} \in \mathcal{X}_r(G) \cap H_1$.

Let $N_v(t)$ count queries of node v by time t. We spell out the stopping argument to accommodate adaptive query selection and arbitrary internal randomization. For an integer n, retain the transcript through $\tau \wedge n$, its stopping indicators, and its output if stopping has occurred. The conditional laws of the query choices and the internal randomization are the same functions of the observed past under both states. They therefore contribute zero to the conditional relative entropy. An observation at $v \notin D$ also contributes zero, whereas an observation at $v \in D$ contributes $\mathrm{kl}(p_1, p_0) = \kappa$. The chain rule for relative entropy gives

$$\mathrm{KL}\big(\mathbb{P}_{\mathbf{1}}^{(n)}, \mathbb{P}_{x^D}^{(n)}\big) = \kappa \sum_{v \in D} \mathbb{E}_{\mathbf{1}} N_v(\tau \wedge n). \tag{28}$$

Here the superscript denotes the law of this finite stopped transcript. This is the usual sequential change of measure argument underlying fixed confidence information bounds Garivier & Kaufmann (2016); Degenne et al. (2019).

The event $E_n = \{\tau \le n, \widehat{h} = 1\}$ is measurable with respect to this transcript. Data processing applied to its indicator yields

$$\kappa \sum_{v \in D} \mathbb{E}_{\mathbf{1}} N_v(\tau \wedge n) \ge \mathrm{kl}\big(\mathbb{P}_{\mathbf{1}}(E_n), \mathbb{P}_{x^D}(E_n)\big).$$

Both states have almost surely finite stopping time by the definition of a valid procedure. Thus, as $n \to \infty$, the two probabilities tend to $p = \mathbb{P}_{\mathbf{1}}(\widehat{h} = 1) \ge 1 - \delta$ and $q = \mathbb{P}_{x^D}(\widehat{h} = 1) \le \delta$, respectively. Monotone convergence for the counts and lower semicontinuity of binary relative entropy imply

$$\kappa \sum_{v \in D} \mathbb{E}_{\mathbf{1}} N_v(\tau) \ge \mathrm{kl}(p, q) \ge \mathrm{kl}(1 - \delta, \delta).$$

For the last inequality, binary relative entropy is increasing in its first argument and decreasing in its second whenever the first exceeds the second; the boundary cases follow by continuity or give infinite relative entropy.

Consequently

$$\lambda_v = \frac{\kappa \mathbb{E}_{\mathbf{1}} N_v(\tau)}{\mathrm{kl}(1 - \delta, \delta)}, \qquad v \in V,$$

is feasible in the fractional certificate program. Since $\sum_v N_v(\tau) = \tau$, its objective value is at least $C_r(G, S)$. Rearranging and then taking the infimum over $\mathcal{A}$ proves

$$\mathfrak{T}_r(G, S, \delta) \ge \frac{C_r(G, S)}{\kappa} \mathrm{kl}(1 - \delta, \delta).$$

□

D.4 A FINITE CONFIDENCE UPPER BOUND FROM A FRACTIONAL CERTIFICATE

Algorithm D.3 (Fractional certificate verification)**.** Given a feasible allocation λ for (9), set

$$C' = \sum_v \lambda_v, \qquad \pi_v = \lambda_v / C', \qquad M = \lceil 2C' \log(2/\delta) \rceil, \qquad b_0 = \log(2kM/\delta).$$

For each of at most M blocks, draw a node independently from π and run the one node test with fresh observations and boundaries $(-b_0, \log 2)$. An upper exit advances to the next block. A lower exit ends the block stage and returns the output of the fresh direct target test in Lemma D.2. If all M blocks exit upward, return one. For a supplied legal ideal, apply this construction on the induced graph as justified in Appendix E.4.

Proof of the upper bound in Theorem 3.1. Write $C = C_r(G, S)$ and let λ^* be an optimal fractional certificate. Such an optimum exists: the finite linear program is feasible by putting weight 1 on every target, and it may be restricted to the compact set of nonnegative allocations of total mass at most k. Moreover $1 \le C \le k$, since $V \in \mathcal{D}_r(G, S)$ forces total mass at least 1.

Algorithm D.3 uses

$$\pi_v = \lambda^*_v / C, \qquad M = \lceil 2Ca_\delta \rceil, \qquad u_0 = \log 2, \qquad b_0 = \log(2kM/\delta).$$

Each block chooses a node independently according to π and applies Lemma D.1 to fresh observations at that node with thresholds (u_0, b_0). An upper exit advances to the next block. A lower exit ends this stage and invokes the fresh data procedure of Lemma D.2. Completing all M blocks with upper exits returns 1 directly.

Fix $x \in \mathcal{X}_r(G) \cap H_0$ and put $D = \{v : x_v = 0\}$. The certificate constraints imply

$$\pi(D) = \frac{\sum_{v \in D} \lambda^*_v}{C} \ge \frac{1}{C}.$$

Condition on reaching any block and on its entire earlier history. The new node still has law π. If it belongs to D, its test has upper exit probability at most $e^{-u_0} = 1/2$. Because that test terminates almost surely, its lower exit probability is at least $1/2$. It follows that the conditional probability of invoking the fallback in this block is at least $1/(2C)$. Iterating these conditional bounds,

$$\mathbb{P}_x(\text{all } M \text{ blocks have upper exits}) \le \left(1 - \frac{1}{2C}\right)^M \le \exp\left(-\frac{M}{2C}\right) \le \frac{\delta}{2}. \tag{29}$$

If the fallback is invoked, its conditional false certification probability is at most $\delta/2$. Adding the two disjoint ways to return an incorrect certificate proves a total probability at most δ under H_0.

For any $x \in \mathcal{X}_r(G) \cap H_1$, returning 1 in the main stage is correct. An incorrect return of 0 can only come from the fallback and has probability at most $\delta/2$. In particular, a lower exit at a nontarget node is not itself used to reject the target conjunction. This distinction ensures validity when targets are mastered but some other queried abilities are missing.

There are at most M one node tests and at most one fallback, and each has finite expectation at every fixed state. Thus the full procedure has finite expected stopping time everywhere in $\mathcal{X}_r(G)$.

It remains to bound its cost at $\mathbf{1}$. Consider M independently generated potential blocks, including the unused ones after any lower exit. Every block is at a mastered node and has expected length at most $(\log 2 + \ell_\eta)/\kappa$. Early termination only decreases the sum of their lengths. By the lower exit bound and a union bound, the event F of invoking the fallback satisfies

$$\mathbb{P}_{\mathbf{1}}(F) \le M e^{-b_0} = \frac{\delta}{2k}.$$

The fresh data property and Lemma D.2 give

$$\mathbb{E}_{\mathbf{1}}\left[\mathbf{1}_F T_{\mathrm{dir}}\right] \le \mathbb{P}_{\mathbf{1}}(F) \frac{k(a_\delta + \ell_\eta)}{\kappa} \le \frac{\delta(a_\delta + \ell_\eta)}{2\kappa}.$$

Consequently

$$\mathbb{E}_{\mathbf{1}}\tau \le \frac{\lceil 2Ca_\delta \rceil (\log 2 + \ell_\eta)}{\kappa} + \frac{\delta(a_\delta + \ell_\eta)}{2\kappa}, \tag{30}$$

as claimed. The proof only requires a feasible allocation: substituting any such λ and its total mass for λ^* and C proves the same guarantee with that total mass.

To establish the uniform rate statement, let $L = \log(1/\delta)$ and suppose $\delta \le 1/4$. Then $a_\delta = L + \log 2 \le \frac{3}{2}L$ and

$$\mathrm{kl}(1-\delta, \delta) = (1-2\delta)\log\frac{1-\delta}{\delta} \ge \frac{1}{4}L.$$

Indeed $1-2\delta \ge 1/2$ and $1-\delta \ge \sqrt{\delta}$ on this interval. Since $C \ge 1$ and $L \ge \log 4$, the ceiling in (30) and its additive fallback term are bounded by constants depending only on η times CL. Together with the lower bound, this proves

$$\mathfrak{T}_r(G, S, \delta) = \Theta_\eta\big(C_r(G, S)\log(1/\delta)\big), \qquad 0 < \delta \le 1/4,$$

with constants independent of G, d, S, r, and δ. □

D.5 The exact high confidence constant

Proof of the exact constant clause in Theorem 3.1. Keep G, S, r, and η fixed. Let λ^* be an optimal fractional certificate, $C = C_r(G, S)$, $\alpha = a_\delta = \log(2/\delta)$, and $\varepsilon = \alpha^{-1/4}$. In a deterministic first stage, query node v exactly

$$n_v = \left\lceil \frac{(1+\varepsilon)\alpha\lambda^*_v}{\kappa} \right\rceil$$

times, with $n_v = 0$ when $\lambda^*_v = 0$, and compute

$$L_v = \ell_\eta \sum_{j=1}^{n_v} (2Y_{vj} - 1).$$

Return 1 if

$$\sum_{v \in D} L_v \ge \alpha \quad \text{for every } D \in \mathcal{D}_r(G, S). \tag{31}$$

Otherwise invoke the fresh data direct target fallback. The first stage uses finitely many observations and the fallback has finite expected duration at every state, so the full procedure also has this property.

To prove validity, fix any $x \in \mathcal{X}_r(G) \cap H_0$ and write $D_x = \{v : x_v = 0\}$. Then $D_x \in \mathcal{D}_r(G, S)$. The likelihood ratio of the entire first stage observations under $\mathbf{1}$ versus x equals

$$R_x = \exp\left(\sum_{v \in D_x} L_v\right), \qquad \mathbb{E}_x R_x = 1.$$

The direct certification event (31) implies $R_x \ge e^\alpha$. Markov's inequality therefore bounds its probability by $e^{-\alpha} = \delta/2$. The fallback adds at most $\delta/2$ to the false certification probability. Under H_1, only the fallback can produce an incorrect answer and its error is at most $\delta/2$. Thus the procedure is uniformly δ-correct. No union bound over competing states is used for this validity argument: only the fixed true state x is needed.

We next control fallback probability at $\mathbf{1}$. For a fixed $D \in \mathcal{D}_r(G, S)$ let $N_D = \sum_{v \in D} n_v$. Certificate feasibility implies

$$N_D \ge \frac{(1+\varepsilon)\alpha}{\kappa}.$$

Under $\mathbf{1}$, the variable $\sum_{v \in D} L_v$ is a sum of N_D independent variables in $[-\ell_\eta, \ell_\eta]$ with mean κ each. Hoeffding's inequality gives

$$\mathbb{P}_{\mathbf{1}}\left(\sum_{v \in D} L_v < \alpha\right) \le \exp\left[-\frac{(\kappa N_D - \alpha)^2}{2\ell_\eta^2 N_D}\right]. \tag{32}$$

To handle the integer rounding explicitly, the function

$$f(N) = \frac{(\kappa N - \alpha)^2}{N} = \kappa^2 N - 2\kappa\alpha + \frac{\alpha^2}{N}$$

has derivative $\kappa^2 - \alpha^2/N^2 \geq 0$ for $N \geq \alpha/\kappa$. Substituting the lower bound on N_D into (32) is therefore valid and yields

$$\mathbb{P}_{\mathbf{1}}\left(\sum_{v\in D} L_v < \alpha\right) \leq \exp\left[-\frac{\kappa\varepsilon^2\alpha}{2\ell_\eta^2(1+\varepsilon)}\right] = \exp\left[-\frac{\kappa\sqrt{\alpha}}{2\ell_\eta^2(1+\alpha^{-1/4})}\right].$$

Let $M_{\mathrm{bad}} = |\mathcal{D}_r(G,S)| \leq 2^d$. A union bound, now used solely for the expected cost at the reference instance, gives

$$\mathbb{P}_{\mathbf{1}}(\text{fallback}) \leq \beta_\alpha, \qquad \beta_\alpha = \min\left\{1, M_{\mathrm{bad}} \exp\left[-\frac{\kappa\sqrt{\alpha}}{2\ell_\eta^2(1+\alpha^{-1/4})}\right]\right\}. \tag{33}$$

The deterministic first stage budget is at most $\sum_v n_v \leq (1+\alpha^{-1/4})\alpha C/\kappa + d$. Independence of the fallback observations and (27) consequently imply

$$\mathbb{E}_{\mathbf{1}}\tau \leq \frac{(1+\alpha^{-1/4})\alpha C}{\kappa} + d + \beta_\alpha \frac{k(\alpha+\ell_\eta)}{\kappa}. \tag{34}$$

For fixed G, S, r, the numbers d, k, and M_{bad} are fixed, so $\beta_\alpha \to 0$ as $\delta \downarrow 0$. Since $\alpha/\log(1/\delta) \to 1$, dividing (34) by $\log(1/\delta)$ gives the upper limit C/κ. The information lower bound and

$$\frac{\mathrm{kl}(1-\delta,\delta)}{\log(1/\delta)} \longrightarrow 1$$

give the matching lower limit. Hence

$$\lim_{\delta\downarrow 0} \frac{\mathfrak{T}_r(G,S,\delta)}{\log(1/\delta)} = \frac{C_r(G,S)}{\kappa}.$$

The fixed graph qualification is essential for this particular upper bound construction: its remainder explicitly contains d and M_{bad}. The finite confidence construction in Theorem 3.1 supplies the separate bound that is uniform in graph size. □

E Structural and robustness proofs

This appendix proves the graph reductions and robustness statements used in the main text. All cuts and degrees refer to the given edge set of G; no transitive edges are added. For a set $D \subseteq V$, write $\lambda(D) = \sum_{v\in D} \lambda_v$.

E.1 Low cuts, certificates, and disjoint paths

Proposition E.1 (Certificates and connectivity)**.** *A set Q of mastered nodes logically certifies all targets throughout $\mathcal{X}_r(G)$ if and only if $Q \cap D \neq \varnothing$ for every $D \in \mathcal{D}_r(G,S)$. Equivalently, for each $s \in S \setminus Q$, the minimum directed edge cut from s to Q exceeds r; by Menger's theorem, there are then at least $r+1$ edge disjoint directed paths from s to Q.*

Proof of Proposition E.1. For a state x, put $D = \{v : x_v = 0\}$. An edge (u,v) contributes one to $B_G(x)$ exactly when $u \in D$ and $v \notin D$. Consequently,

$$B_G(x) = |\partial_G^+(D)|.$$

Moreover, the target claim fails exactly when $D \cap S \neq \varnothing$. The correspondence $x \leftrightarrow D$ is a bijection, so the competing states are exactly the members of $\mathcal{D}_r(G,S)$.

Suppose the coordinates in Q have been observed without noise to be one. A competing state is consistent with these observations precisely when its missing set is disjoint from Q. Thus Q certifies the claim if and only if

$$Q \cap D \neq \varnothing \qquad \text{for every } D \in \mathcal{D}_r(G,S).$$

For each $s \in S \setminus Q$, define

$$c(s,Q) = \min\big\{|\partial_G^+(D)| : s \in D,\ D \cap Q = \varnothing\big\}.$$

It follows directly that Q is a certificate if and only if $c(s, Q) > r$ for every $s \in S \setminus Q$. Targets already in Q require no additional condition.

To identify $c(s, Q)$, attach a new sink to every vertex of Q with infinite capacity edges and give each original edge unit capacity. A finite s–sink cut has its source side disjoint from Q, and its capacity is exactly the outgoing boundary of that source side in G. Conversely, each set in the definition of $c(s, Q)$ gives such a cut. Hence $c(s, Q)$ is the minimum s–Q edge cut size. The directed edge version of Menger's theorem identifies this number with the maximum number of paths from s to Q that are disjoint in their original edges. Paths may share vertices, including their terminal vertex in Q. Therefore the certificate condition is equivalent to the existence of at least $r + 1$ such paths for every $s \in S \setminus Q$. If $Q = \varnothing$, the cut $D = V$ has size zero and the condition fails, as it must for a nonempty target set. □

The fractional certificate linear program is feasible: assigning weight one to every target and zero elsewhere satisfies all its constraints. Its optimum is at most $|S|$. Since $V \in \mathcal{D}_r(G, S)$, every feasible allocation has total mass at least one. The optimum is attained, for example by restricting the search to the closed bounded set of allocations with total mass at most $|S|$. Ordinary finite dimensional linear program duality gives the fractional packing dual. Increasing r enlarges the constraint family and therefore cannot decrease $C_r(G, S)$.

E.2 An exact allocation program for arbitrary two layer graphs

For a two layer graph $V = U \dot\cup A$, with all edges from basic nodes U to advanced nodes A, take $S = U$. Let $N(u)$ be the advanced neighbors of u, and $d_u = |N(u)|$. Write Small_j for the sum of the j smallest entries, with $\mathrm{Small}_0 = 0$.

Theorem E.2 (Exact two layer allocation)**.** *For every such graph,*

$$C_r(G, U) = \min_{\lambda \geq 0} \left\{ \sum_{v \in V} \lambda_v : \lambda_u + \mathrm{Small}_{(d_u - r)_+}\big((\lambda_c)_{c \in N(u)}\big) \geq 1 \quad (u \in U) \right\}. \tag{35}$$

It has an equivalent LP with $O(|V| + |\mathcal{E}|)$ variables and constraints.

Proof of Theorem E.2. Write $V = U \dot\cup A$, with every edge directed from U to A, and let $S = U$. For $u \in U$, let $N(u) \subseteq A$ be its neighborhood, $d_u = |N(u)|$, and $m_u = (d_u - r)_+$. Consider the finite collection of constraints

$$\lambda_u + \sum_{c \in B} \lambda_c \geq 1, \qquad u \in U, \quad B \subseteq N(u), \quad |B| = m_u. \tag{36}$$

Each set $\{u\} \cup B$ intersects the target set and has outgoing cut size $d_u - m_u = \min\{d_u, r\} \leq r$. Thus every feasible fractional certificate satisfies (36).

For the converse, let $\lambda \geq 0$ satisfy these constraints and take any $D \in \mathcal{D}_r(G, U)$. Write $D = U' \cup A'$, where $\varnothing \neq U' \subseteq U$ and $A' \subseteq A$. Since edges have distinct tails,

$$|\partial_G^+(D)| = \sum_{u \in U'} |N(u) \setminus A'| \leq r.$$

Fix any $u \in U'$. It follows that $|N(u) \cap A'| \geq m_u$. Choose $B \subseteq N(u) \cap A'$ of size m_u. Nonnegativity and (36) imply

$$\lambda(D) \geq \lambda_u + \sum_{c \in B} \lambda_c \geq 1.$$

All original cut constraints therefore hold. Minimizing the left side of (36) over B gives exactly the sum of the m_u smallest weights in $N(u)$. This proves the stated reduction, including $m_u = 0$, when the constraint is simply $\lambda_u \geq 1$. □

For completeness, the reduction has the following explicit linear program formulation; an exponential list of subset constraints is unnecessary. For each u with $r < d_u$, introduce $a_u \geq 0$ and $z_{uc} \geq 0$ for

$c \in N(u)$. In addition to $\lambda_v \geq 0$, impose

$$z_{uc} \geq \lambda_c - a_u, \qquad c \in N(u),\ r < d_u, \tag{37}$$

$$\lambda_u + \sum_{c \in N(u)} \lambda_c - r a_u - \sum_{c \in N(u)} z_{uc} \geq 1, \qquad r < d_u, \tag{38}$$

$$\lambda_u \geq 1, \qquad r \geq d_u. \tag{39}$$

The objective remains $\min \sum_v \lambda_v$.

To verify exactness, let $w_1 \geq \cdots \geq w_d \geq 0$ be sorted weights and let $0 \leq r < d$ be an integer. Then

$$\sum_{j=1}^{r} w_j = \min_{a \geq 0} \left\{ ra + \sum_{j=1}^{d} (w_j - a)_+ \right\}. \tag{40}$$

Indeed, for every a, the expression on the right is at least $\sum_{j=1}^{r}(a + (w_j - a)_+) \geq \sum_{j=1}^{r} w_j$. For $r \geq 1$, equality holds at any $a \in [w_{r+1}, w_r]$; for $r = 0$, take $a \geq w_1$. Subtracting (40) from the total weight expresses the sum of the $d - r$ smallest entries as

$$\max_{a \geq 0} \left\{ \sum_{j=1}^{d} w_j - ra - \sum_{j=1}^{d} (w_j - a)_+ \right\}.$$

For fixed a_u, the least admissible slack in (37) is $(\lambda_c - a_u)_+$. Existence of a_u, z_{uc} satisfying (38) is consequently equivalent to the smallest sum constraint. There are $O(|V| + |\mathcal{E}|)$ variables and inequalities, with rational coefficients of polynomial encoding length. This supplies an ordinary polynomial size linear program for the allocation problem. It concerns nonnegative allocation weights; it does not assert a corresponding formula for arbitrary signed likelihood statistics.

E.3 The biregular law and the integer certificate gap

Corollary E.3 (Biregular networks)**.** *If $|U| = k$, $|A| = t$, every basic node has degree $q \geq 1$, and every advanced node has indegree $b \geq 1$, then $kq = tb$ and*

$$C_r(G, U) = \begin{cases} \min\{k, t/(q-r)\}, & 0 \leq r < q, \\ k, & r \geq q. \end{cases} \tag{41}$$

Unit weights on the targets or weights $1/(q - r)$ on the advanced nodes attain the two branches; averaging the constraints of (35) gives the matching lower bound.

Proof of Corollary E.3. Let $|U| = k$, $|A| = t$, let every root have degree $q \geq 1$, and let every vertex of A have indegree $b \geq 1$. Counting edges gives $kq = tb$. First suppose $r < q$ and put $m = q - r$. The allocation $\lambda_u = 1$ on U, zero elsewhere, has cost k. The allocation $\lambda_c = 1/m$ on A, zero elsewhere, has cost t/m and satisfies every constraint in Theorem E.2. Thus $C_r(G, U) \leq \min\{k, t/m\}$.

For any feasible allocation and any root u, average (36) over all m-element subsets of $N(u)$. Each neighbor appears in a fraction m/q of these subsets, giving

$$\lambda_u + \frac{m}{q} \sum_{c \in N(u)} \lambda_c \geq 1.$$

Summing over the roots and using the common indegree b yields

$$\sum_{u \in U} \lambda_u + \frac{bm}{q} \sum_{c \in A} \lambda_c \geq k. \tag{42}$$

Set $\beta = bm/q = km/t$. For nonnegative a, z, the implication $a + \beta z \geq k \Rightarrow a + z \geq k/\max\{1, \beta\}$ follows by comparing coefficients. Applying it to (42) gives $\sum_v \lambda_v \geq \min\{k, t/m\}$, proving equality.

If $r \geq q$, each singleton $\{u\}$ is an admissible missing set. Hence every root must have weight at least one. The direct root allocation attains the resulting lower bound k. □

Proposition E.4 (Logical size and noisy allocation can differ)**.** *For the complete bipartite prerequisite graph $K_{k,t}$, the smallest noiseless full mastery certificate has size*

$$h_r(K_{k,t}, U) = \begin{cases} \min\{k, r+1\}, & r < t, \\ k, & r \geq t, \end{cases}$$

whereas $C_r = \min\{k, t/(t-r)\}$ for $r < t$, and $C_r = k$ otherwise.

For example, $k = 64, t = 32, r = 16$ gives $h_r = 17$ but $C_r = 2$. This is a distinction between integer and fractional certificates, not a separation from all adaptive algorithms.

Proof of Proposition E.4. Consider the complete two layer graph $K_{k,t}$, and decompose a candidate noiseless certificate as $Q = Q_U \cup Q_A$. If $Q_U = U$, then Q certifies the claim. If $|Q_A| \geq r + 1$, it also certifies the claim: any missing root would have an outgoing violated edge to each of the queried, mastered vertices in Q_A, exceeding the rule budget.

Conversely, suppose a root u is not in Q_U and $|Q_A| \leq r$. The missing set

$$D = \{u\} \cup (A \setminus Q_A)$$

avoids Q, intersects U, and has outgoing cut size $|Q_A| \leq r$. It is therefore a counterexample to certification. Thus a certificate must query either every root or at least $r + 1$ upper layer vertices. When $r < t$, either option is available and the smallest size is $\min\{k, r+1\}$. When $r \geq t$, the second option is unavailable, and every root is necessary; the optimum is k. □

E.4 A PRESCRIBED LEGAL REFERENCE STATE

Proof of Theorem 3.1. Let I be a legal ideal with $\varnothing \neq S \subseteq I$, set $x^* = \mathbf{1}_I$, and write $G_I = G[I]$. Because I is an ideal, no edge has its tail in $V \setminus I$ and its head in I.

For the lower bound, choose any $D \in \mathcal{D}_r(G_I, S)$ and define $x' = \mathbf{1}_{I \setminus D}$. On edges internal to I, the violations of x' are exactly the edges of $\partial^+_{G_I}(D)$. Edges with head outside I cannot be violated, because their heads have state zero. Edges from outside I into I do not exist. Thus

$$B_G(x') = |\partial^+_{G_I}(D)| \leq r. \tag{43}$$

The state x' makes the target claim false and differs from x^* only on D. For every uniformly δ-correct procedure, the information bound therefore gives

$$\kappa \sum_{v \in D} \mathbb{E}_{x^*} N_v(\tau) \geq \mathrm{kl}(1-\delta, \delta), \qquad D \in \mathcal{D}_r(G_I, S).$$

Normalizing these expected counts gives a feasible allocation on I. Consequently,

$$\mathbb{E}_{x^*} \tau \geq \frac{C_r(G_I, S)}{\kappa} \mathrm{kl}(1-\delta, \delta). \tag{44}$$

Observations outside I cannot weaken these constraints because all the constructed alternatives agree there with x^*.

For the upper bound, run a procedure for the induced graph G_I, using only coordinates in I. For every $x \in \mathcal{X}_r(G)$, restriction to I satisfies

$$B_{G_I}(x|_I) \leq B_G(x) \leq r,$$

because every violated edge internal to I is also a violated edge of G. Hence $x|_I \in \mathcal{X}_r(G_I)$, and the target claim has the same truth value before and after restriction. Any procedure uniformly valid on $\mathcal{X}_r(G_I)$ is consequently uniformly valid on the full domain $\mathcal{X}_r(G)$ when implemented in this way.

At the reference x^*, all queried coordinates are one. Their complete observation law equals the all one observation law on G_I. Set $C = C_r(G_I, S)$, $k = |S|$, and $M = \lceil 2C \log(2/\delta) \rceil$. Applying the finite confidence construction of Theorem 3.1 to G_I gives

$$\mathbb{E}_{x^*} \tau \leq \frac{M(\log 2 + \ell_\eta)}{\kappa} + \frac{\delta}{2\kappa}\big(\log(2/\delta) + \ell_\eta\big). \tag{45}$$

The exact constant construction in Appendix D.5 supplies the matching first order upper bound on G_I. Together with (44), it gives

$$\lim_{\delta \downarrow 0} \frac{\inf_{\mathcal{A}} \mathbb{E}_{x^*} \tau_{\mathcal{A}}}{\log(1/\delta)} = \frac{C_r(G[I], S)}{\kappa},$$

where the infimum is over procedures uniformly δ-correct on $\mathcal{X}_r(G)$. All graphs, the ideal, and the noise level are fixed in this limit.

The same expected upper bound holds for any admissible state that is one on all of I, because the procedure queries only I. No assumption that the proposed ideal is true was used to establish validity. The construction selects a procedure for a prescribed I. It does not supply one evaluator with a uniform discovery guarantee when the favorable ideal is unknown. □

F Discovery on general DAGs

Use the notation of Theorem 4.2; all rules here are exact. Descendant sets include their starting node. The baseline coefficient has the finite formulation

$$C_B = \min_{\gamma \in \mathbb{R}_+^B} \left\{ \sum_{v \in B} \gamma_v : \gamma(\mathrm{Desc}_{G[B]}(s)) \geq 1 \quad (s \in S) \right\}. \tag{46}$$

Indeed, omitting s forces omission of all its descendants, and deleting exactly those descendants from B gives a legal null ideal. Together with the two covers in the main text, this gives polynomial size LPs when the family $\mathcal{F}$ is explicitly listed. No claim is made for a succinct representation of exponentially many profiles.

Proof of Theorem 4.2. Put $a = \log(1/\delta)$, $b_\delta = \mathrm{kl}(1-\delta, \delta)$, and $K_\delta = \mathrm{kl}(1/2, \delta)$. Both divergences are comparable to a uniformly for $0 < \delta \leq 1/4$.

Lower bound. The dual of the difference cover LP supplies a distribution μ on $\mathcal{F}$ such that

$$\sum_{I : v \in D_I} \mu_I \leq 1/H_{\mathrm{diff}} \qquad (v \in V). \tag{47}$$

To obtain it, divide an optimal dual packing by its total mass H_{diff}. Fix a uniformly correct test and write $z = \mathbf{1}_B$. For $s \geq 0$, set $p_s = \mathbb{P}_z\{\tau \leq s, \widehat{h} = 1\}$, and let N_v^s count queries through $\tau \wedge \lfloor s \rfloor$. For every $u \in S$, the ideal $B \setminus \mathrm{Desc}_{G[B]}(u)$ is a full graph null; write its indicator as $z^{(u)}$. Finite-horizon change of measure, including the decision whenever stopping occurs, gives

$$\kappa \sum_{v \in \mathrm{Desc}_{G[B]}(u)} \mathbb{E}_z N_v^s \geq \mathrm{kl}\Big(p_s, \mathbb{P}_{z^{(u)}}\{\tau \leq s, \widehat{h} = 1\}\Big).$$

If $p_s \geq 1/2$, correctness bounds the second argument by δ, so the right side is at least K_δ. The normalized counts are then feasible in (46). Consequently

$$p_s \geq 1/2 \implies \kappa s \geq C_B K_\delta, \qquad s_0 = \min\{C_B K_\delta/(2\kappa), H_{\mathrm{diff}}/8\} \implies p_{s_0} < 1/2.$$

Couple the experiments at z and $\mathbf{1}_I$ with common randomization and flips. They agree until the first query in D_I. Let $H_I(s)$ be the event that the baseline experiment queries D_I before stopping and within $\lfloor s \rfloor$ queries. For every realized baseline query sequence, a union bound and (47) imply

$$\sum_I \mu_I \mathbf{1}\{H_I(s)\} \leq \lfloor s \rfloor / H_{\mathrm{diff}}.$$

Since z is positive, its probability of rejecting is at most δ. Coupling and averaging therefore give

$$\sum_I \mu_I \mathbb{P}_{\mathbf{1}_I}\{\tau \leq s_0\} \leq p_{s_0} + \delta + s_0/H_{\mathrm{diff}} \leq 7/8.$$

Thus $\max_I \mathbb{E}_{\mathbf{1}_I} \tau \geq s_0/8$, without any mean duration assumption at z. Separately, comparison of each $\mathbf{1}_I$ with the empty ideal gives $\kappa \mathbb{E}_{\mathbf{1}_I} \tau \geq b_\delta$: every query contributes at most κ, and the stopped argument of Appendix D applies whenever the mean is finite. Hence

$$\max_I \mathbb{E}_{\mathbf{1}_I} \tau \geq \max\left\{ \frac{b_\delta}{\kappa}, \frac{1}{8} \min\left\{ \frac{C_B K_\delta}{2\kappa}, \frac{H_{\mathrm{diff}}}{8} \right\} \right\}. \tag{48}$$

Using $C_B \geq 1$ and

$$\max\{a, \min(C_B a, H)\} \leq \min\{C_B a, H + a\} \leq 2\max\{a, \min(C_B a, H)\}$$

proves the required lower rate.

Upper bound and full class stopping. Theorem 3.1, applied to the supplied ideal B at level $\delta/2$, gives a verifier querying only B, with finite mean everywhere and mean $O_\eta(C_B a)$ at every $\mathbf{1}_I$. If $H_{\text{cert}} = \infty$, this suffices. Otherwise take an optimal certifying cover λ. Independently in each trial l, draw $v \in Q \cap W$ with probability $\pi_v = \lambda_v / H_{\text{cert}}$ and run a fresh one node SPRT with boundaries

$$-\log 2, \qquad u_l = \log \frac{\pi^2 l^2}{3\delta}.$$

An upper exit accepts the target claim; a lower exit starts another trial. Every legal null has all of W zero, so the probability this search ever falsely accepts is at most $\sum_{l \geq 1} e^{-u_l} = \delta/2$.

At $\mathbf{1}_I$, a fresh selected node is mastered with probability $p = \pi(I \cap Q \cap W) \geq 1/H_{\text{cert}}$. Its check accepts with probability at least $1/2$. Trial l is consequently reached with probability at most ρ^{l-1}, where $\rho = 1 - p/2$. The one node duration bounds and Tonelli's theorem yield

$$\mathbb{E}_{\mathbf{1}_I} T_{\text{search}} \leq \frac{1}{\kappa} \sum_{l \geq 1} \rho^{l-1} \{(1-p)(\log 2 + \ell_\eta) + p(u_l + \ell_\eta)\}.$$

For a geometric L of parameter $p/2$, $\sum_l \rho^{l-1} = 2/p$ and $\sum_l \rho^{l-1} \log l = (2/p)\mathbb{E} \log L \leq (2/p)\log(2/p)$. Therefore

$$\mathbb{E}_{\mathbf{1}_I} T_{\text{search}} \leq \frac{2}{\kappa} \left[\frac{1-p}{p} (\log 2 + \ell_\eta) + \log \frac{\pi^2}{3\delta} + \ell_\eta + 2\log(2/p) \right] = O_\eta(H_{\text{cert}} + a).$$

Alternate individual queries between this search and the independent baseline verifier, stopping on the first decision. Each sees fresh observations in its own query clock. Null error is at most δ; on positive states only the baseline verifier can reject, with probability at most $\delta/2$. The race lasts at most twice either routine's duration, and the baseline verifier ensures finite mean on every legal state. Its favorable cost is thus $O_\eta(\min\{C_B a, H_{\text{cert}} + a\})$. If $Q \subseteq W$, the two cover programs coincide; equality of their values gives the stated matching conclusion. □

Overlap and complementary diagnostics. The matching class realizes any finite hypergraph $\mathcal{H}$ of nonempty sets: take targets S disjoint from $Q = \bigcup_{H \in \mathcal{H}} H$, every edge $s \to q$, baseline $B = S$, and profiles $S \cup H$. Both covers are exactly the fractional transversal number of $\mathcal{H}$, while $C_B = |S|$ and every supplied profile has verification coefficient one. For all q-subsets of t witnesses this number is t/q: weight $1/q$ at each witness is feasible, and uniform subset averaging gives the matching lower bound.

In a private witness graph, instead take B to contain the targets and all $2m$ diagnostics. For the profiles $B \cup \{c_v\}$, $C_B = k$ and both covers equal t, so their cost is $\Theta_\eta(\min\{ka, t + a\})$, irrespective of the table. The same rate holds for any larger favorable family in which every profile masters a witness: the subfamily gives the lower bound; using baseline S and unit cover weight on all witnesses gives the upper bound. Thus complementary diagnostic failures are an essential part of the informative favorable family.

F.1 The worst case over all positive ideals

Proposition F.1. *Let M be the maximal elements of S in the reachability order and $b = |M|$. For $0 < \delta < 1/2$,*

$$\frac{b}{\kappa} \operatorname{kl}(1-\delta, \delta) \leq \inf_{\mathcal{A}\ \delta\text{-correct on } \mathcal{X}_0(G)} \max_{\substack{I \text{ legal ideal} \\ S \subseteq I}} \mathbb{E}_{\mathbf{1}_I} \tau_{\mathcal{A}} \leq \frac{b}{\kappa} \big(\log(1/\delta) + \ell_\eta\big). \tag{49}$$

Proof. The smallest positive ideal is $B_{\min} = \bigcup_{s \in S} \operatorname{Anc}_G(s)$. Each $u \in M$ is maximal in $B_{\min}$: a strict descendant there would be an ancestor of a target strictly above u. Thus $B_{\min} \setminus \{u\}$ is a legal

null differing only at u, and change of measure forces $\kappa \mathbb{E}_{\mathbf{1}_{B_{\min}}} N_u(\tau) \geq \mathrm{kl}(1-\delta, \delta)$. Summing proves the lower bound.

Under exact rules, all targets are mastered if and only if all of M are. Test these nodes in order with fresh SPRTs having upper boundary $\log(1/\delta)$ and lower boundary $-\log(b/\delta)$. Reject at the first lower exit, and accept if all exit above. At a null, acceptance requires an upper exit at a fixed missing member; at a positive state, a union bound over lower exits gives error at most δ. All tests have finite mean, and at a positive state their summed mean is at most the upper bound in (49). In particular, the fixed graph leading constant is b/κ. A restricted favorable family is therefore needed to express a discovery gain. □

F.2 Equal cover values need not give equal discovery costs

Example F.2 (Three diagnostic tables). Let $h \geq 2$, $t = k = 2^h$, and $m = 2h$. On witnesses indexed by $u \in \{0,1\}^h$, compare

$$E_{\mathrm{rep}}(u) = (u_1, \ldots, u_1), \qquad E_{\mathrm{raw}}(u) = (u, u), \qquad E_{\mathrm{code}}(u) = (u, uP),$$

where $P \in \{0,1\}^{h \times h}$ has nonzero columns and the code $\{(u, uP)\}$ has the distance guaranteed by the surplus construction of Appendix H with $b = h$. All three private witness networks have the same degrees and profile sizes, and, with baseline $B = S$, all have

$$C_B = t, \qquad C_0(G[I_u], S) = 1, \qquad H_{\mathrm{diff}} = 2, \qquad H_{\mathrm{cert}} = t.$$

Their testing costs are respectively $\Theta_\eta(t + a)$, $\Theta_\eta(h \log(h+1) + a)$ and $\Theta_\eta(h + a)$, uniformly for $\delta \leq 1/4$.

Proof. Every column is balanced: the first two tables copy address bits, and a parity column $u \mapsto u \cdot p$ with $p \neq 0$ takes each value on half of the messages. Hence every diagnostic root has outdegree $t/2$, every target has outdegree t, and every witness has indegree $k + 2h$. Each profile contains $k + 2h + 1$ nodes. The baseline has isolated targets, giving $C_B = t$; its supplied profile is certified by its unique witness, giving coefficient one. Putting unit weight on both members of one diagnostic pair proves $H_{\mathrm{diff}} \leq 2$. Under a uniform hidden index, each diagnostic is present with probability $1/2$, and each witness with probability $1/t \leq 1/2$. Averaging any feasible difference cover constraints therefore forces total weight at least two. Only witness nodes in the departure certify the conjunction, and each profile contains just its own witness, so $H_{\mathrm{cert}} = t$.

Repeated columns preserve hitting cost exactly: a query to any copy can be replaced by a fresh query to its original coordinate, and conversely a representative copy simulates the original. This preserves every adaptive transcript and the hitting event on the favorable family, without imposing equal labels throughout the validity class. Hence E_{rep} has the hitting cost of its one column table, and E_{raw} that of the h-column identity table. The lower bound of Theorem 5.1 at budget one and witness scanning give $\mathcal{H}_\eta(E_{\mathrm{rep}}) = \Theta_\eta(t)$. At budget h, that theorem's lower bound and its grouped signature upper construction, which is precisely the identity table, give $\mathcal{H}_\eta(E_{\mathrm{raw}}) = \Theta_\eta(h \log(h+1))$. For E_{code}, the decoding policy of Appendix H with $b = h$ has horizon $O_\eta(h \log(2 + h/(h+1))) = O_\eta(h)$, and Lemma G.1 gives $\mathcal{H}_\eta(E_{\mathrm{code}}) \geq \log_2(1 + 3t/4) \geq h/2$. Theorem 4.1 now gives the stated testing rates, since $a \geq \log 4$ and each hitting cost is $O(t)$. The favorable profile vectors follow their respective tables; the comparison does not rewire a fixed population at unchanged states. □

G From witness discovery to a valid test

Use the private witness model and hitting quantity of Definition 2.1. The favorable family guides efficiency only; validity always covers every legal ideal.

Proof of Theorem 4.1. Extracting a bounded horizon search. Fix a uniformly correct procedure with $T = \max_j \mathbb{E}_{\mathbf{1}_{I_j}} \tau < \infty$; otherwise the lower bound is immediate. Before the first query its behavior is independent of the state. Validity at a null and a positive state bounds the probabilities of accepting and rejecting without a query by δ each. Thus $T \geq 1 - 2\delta \geq 1/2$, and

$$Q = \lceil 8T \rceil \leq 10T, \qquad K_\delta = \mathrm{kl}(1/2, \delta).$$

Delete the private witness to obtain the positive baseline $z_j = \mathbf{1}_{I_j \setminus \{c_j\}}$. All its mastered nodes are roots, so each $z_j - e_{u_i}$ is a legal null. If $p_j = \mathbb{P}_{z_j}\{\tau \le Q, \widehat{h} = 1\} \ge 1/2$, the finite transcript through $\tau \wedge Q$, including its decision, gives

$$\kappa \mathbb{E}_{z_j} N_{u_i}(\tau \wedge Q) \ge K_\delta \quad (i \in [k]), \qquad \kappa Q \ge k K_\delta.$$

Here data processing and correctness at $z_j - e_{u_i}$ supply the binary divergence bound. Therefore $Q < kK_\delta/\kappa$ implies $p_j < 1/2$ for every j.

At the favorable profile, Markov's inequality and correctness give

$$\mathbb{P}_{\mathbf{1}_{I_j}}\{\tau \le Q, \widehat{h} = 1\} \ge 1 - \delta - T/Q \ge 5/8.$$

Couple this experiment to z_j using identical randomization and flips; they coincide until the first query at c_j. In the case $Q < kK_\delta/\kappa$, the probability of hitting c_j by $\tau \wedge Q$ is consequently above $1/8$. Eleven independent repetitions hit with probability at least $1 - (7/8)^{11} > 3/4$. Target queries can be simulated internally as fresh BSC responses at a known one, so this search uses at most $11Q \le 110T$ actual diagnostic or witness queries. In the other case $T \ge kK_\delta/(10\kappa)$. Thus

$$T \ge \min\left\{\frac{kK_\delta}{10\kappa}, \frac{\mathcal{H}_\eta(E)}{110}\right\}.$$

Separately, comparing any favorable profile with a legal null by stopped change of measure gives $T \ge \mathrm{kl}(1-\delta, \delta)/\kappa$, since no observation has divergence exceeding κ. Both divergences are uniformly comparable to $a = \log(1/\delta)$. For $k \ge 1, H \ge 0$,

$$\max\{a, \min(ka, H)\} \le \min\{ka, H + a\} \le 2\max\{a, \min(ka, H)\},$$

which proves the lower bound. No expected duration assumption was used at any comparison baseline.

Checking unrecognized proposals. Put $H = \mathcal{H}_\eta(E)$ and choose a policy attaining that integer horizon. In each cycle $l \ge 1$, run a fresh copy for at most H search queries. Whenever it first queries a witness in that cycle, supply its ordinary response, pause the policy, and run an independent SPRT at that witness with boundaries

$$-\log 2, \qquad u_l = \log \frac{\pi^2 H l^2}{3\delta}.$$

An upper crossing accepts the conjunction. A lower crossing resumes the search. Further queries to a checked witness receive ordinary fresh search responses but trigger no new check. If the search ends without acceptance, start a new cycle. All checking data are withheld from the search policy.

Every legal null has all witnesses zero. The conditional error of each reached check is at most e^{-u_l}, and cycle l has at most H checks. Thus this certifier's probability of ever falsely accepting is at most

$$\sum_{l \ge 1} H e^{-u_l} = \sum_{l \ge 1} \frac{3\delta}{\pi^2 l^2} = \delta/2.$$

At I_j, consider the full potential search trajectory and independent potential checks, even if acceptance stops the cycle early. The search hits c_j with probability at least $3/4$, and its check accepts with probability at least $1/2$. An earlier acceptance is also correct. A reached cycle therefore accepts with probability at least $3/8$, so cycle l is reached with probability at most $(5/8)^{l-1}$.

Deduplication means that at most one checked witness is mastered. The one node bounds give reached cycle expected cost at most

$$H + \frac{H(\log 2 + \ell_\eta) + u_l + \ell_\eta}{\kappa}.$$

Potential observations are independent of the event that their test is reached, and early acceptance only decreases the cost. Tonelli's theorem then gives

$$\mathbb{E}_{\mathbf{1}_{I_j}} T_{\mathrm{cert}} \le \sum_{l \ge 1} (5/8)^{l-1} \left[H + \frac{H(\log 2 + \ell_\eta) + \log(\pi^2 H/(3\delta)) + 2\log l + \ell_\eta}{\kappa}\right] = O_\eta(H + a),$$

using $\log H \le H$.

Finally alternate individual queries between this certifier and the fresh direct target test of Lemma D.2, whose two errors are at most $\delta/2$. Stop at the first decision. Null error is at most δ; on a positive state only the direct test can reject. The race lasts at most twice either routine's duration in its own clock. The direct test therefore ensures finite mean on the full legal class and gives the alternative $O_\eta(ka)$ favorable bound, proving the theorem. □

G.1 A FINITE POLICY REPRESENTATION AND A UNIVERSAL LOWER BOUND

For rational η, the hitting quantity is exactly computable by finite rational LPs, but the representation is exponential. With $A = 2m + t$ actions and horizon $q \le t$, let h record past actions and responses. Introduce realization weights

$$w_\varnothing = 1, \qquad w_{h,v} \ge 0, \qquad \sum_v w_{h,v} = w_h, \qquad w_{hvy} = w_{h,v} \quad (y = 0, 1).$$

These weights account only for action randomization. At a terminal history h, let $p_j(h)$ be the product of its known response probabilities at I_j, and impose

$$\sum_{\substack{h \text{ terminal} \\ c_j \text{ occurs in } h}} w_h p_j(h) \ge 3/4 \qquad (j \in [t]).$$

Every randomized policy gives feasible realization weights. Conversely, at positive weight histories use action probability $w_{h,v}/w_h$, and choose arbitrarily at zero weight histories. This recovers the specified hit probabilities; early stopping can be padded. Feasibility is therefore equivalent to $\mathcal{H}_\eta(E) \le q$. Searching $q = 1, \dots, t$ computes the value, with a history tree of order $(2A)^q$. This policy representation is not the novelty of the reduction (Damaschke, 2016).

Lemma G.1 (Binary-tree hitting bound)**.** *For every table, $\mathcal{H}_\eta(E) \ge \log_2(1 + 3t/4)$.*

Proof. Fix the internal random seed and the response flip sequence. The policy becomes a binary tree of depth q, containing at most $2^q - 1$ queried witness labels. For a uniform hidden index, its own witness can occur on its path only if that label occurs in the tree. Its average hit probability is therefore at most $(2^q - 1)/t$. Integrate over the fixed random quantities and use the required success probability $3/4$. □

H PROOF OF THE DIAGNOSTIC BUDGET LAW

We prove Theorem 5.1. Fix $t = 2^h$, $0 < \eta < 1/2$, and write $\ell = \ell_\eta$, $\gamma = 1/2 - \eta$. All horizons below count repeated reads and witness proposals. Target responses have the same law at every favorable profile and at every baseline below, so a policy may simulate them internally; we assume it does, and every counted step is a diagnostic or witness query.

H.1 AN ENVELOPE FOR ARBITRARY TABLES

Fix a table with $m \ge 1$ columns and a policy of horizon q. For every $z \in \{0, 1\}^m$, including unused signatures, let x^z master the targets and the diagnostic nodes named by z, but no witnesses. These are legal root only baselines. The profile I_j differs from $x^{E(j)}$ only at c_j, so coupling with common flips and randomization gives equal probabilities of hitting c_j by the horizon in the two experiments.

Fix the policy's seed. Before query n, let $P_z(y)$ be the baseline likelihood of prefix history y, and $a_j(y)$ the indicator that its next query proposes c_j. Stopped histories can be padded with a terminal symbol and have all $a_j = 0$. Since $\sum_j a_j(y) \le 1$,

$$\frac{1}{t} \sum_j \mathbb{P}_{x^{E(j)}}\{A_n = c_j\} \le \frac{1}{t} \sum_y \max_z P_z(y) = \frac{2^m}{t} \mathbb{E}_{\overline{P}} \max_z \mathbb{P}(Z = z \mid y), \tag{50}$$

where Z is an auxiliary uniform full cube signature and $\overline{P} = 2^{-m} \sum_z P_z$. The equality is Bayes' rule. Duplicate rows do not change the inequality: selection indicators, not row multiplicities, sum to at most one. For this fixed seed and query index, maximal leakage is $\log\bigl(\sum_y \max_z P_z(y)\bigr)$ (Issa et al., 2019), before the normalization by t.

At these baselines all witness responses are independent of Z; a diagnostic response concerns a single bit, with the complementary member's response recoded if necessary. For a fixed history and seed, adaptive action and stopping choices add no state dependent likelihood factor. The auxiliary posterior thus factors over coordinates. If $n_i(y)$ counts reads of bit i and $L_i(y)$ is its posterior log odds, then $|L_i(y)| \le \ell n_i(y)$ and $\sum_i n_i(y) \le q$. Using $\log(1+x) \ge x/2$ for $0 \le x \le 1$, then Jensen,

$$\max_z \mathbb{P}(Z = z \mid y) = \prod_i (1 + e^{-|L_i(y)|})^{-1} \le \exp\left(-\frac{1}{2}\sum_i e^{-\ell n_i(y)}\right) \le \exp\Big(-\frac{m}{2} e^{-\ell q/m}\Big).$$

Sum (50) over at most q proposals and integrate the seed. With $L = t/2^m$ this proves

$$\frac{1}{t}\sum_j \mathbb{P}_{\mathbf{1}_{I_j}}\{\text{hit } c_j \text{ by } q\} \le \frac{q}{L}\exp\Big(-\frac{m}{2}e^{-\ell q/m}\Big). \tag{51}$$

The full cube prior bounds the envelope; it imposes no prior or independence assumption on the actual table's rows.

H.2 At most the full signature budget

Lower bound. Suppose $1 \le m \le h$, so $L \ge 1$. Required average success $3/4$ and (51) imply $\mathcal{H}_\eta(E) \ge 3L/4$. Put

$$c_\eta = \frac{1}{128(1+\ell)}, \qquad q_m = c_\eta m \log(m+1).$$

The envelope is increasing in real $q \ge 0$, and at q_m is below $1/8$. For $m = 1$, $q_m < 1/128$ suffices. For $m \ge 2$, $(m+1)^{\ell c_\eta} \le \sqrt{m}$. With $x = \sqrt{m}$ and $\log(m+1) \le 1 + x$, the envelope is at most

$$\frac{m\log(m+1)}{128} e^{-\sqrt{m}/2} \le \frac{(x^2+x^3)e^{-x/2}}{128} \le \frac{16/e^2 + 216/e^3}{128} < \frac{1}{8}.$$

Thus $\mathcal{H}_\eta(E) > q_m$, proving $\Omega_\eta(L + m\log(m+1))$. For $m = 0$, all baselines have the same response law and the one point envelope gives average hit probability at most q/t, hence the lower bound $3t/4$.

The same bound applies to tables using fewer than the allowed m pairs: pad with constant columns. Their responses, being constant label channels on the favorable family, can be simulated internally; conversely a policy can ignore them. Padding therefore preserves the hitting cost exactly.

Matching grouped signatures. Index witnesses by h-bit strings and use their first m bits as signatures. Each cell contains $L = 2^{h-m}$ witnesses. For $m \ge 1$, read each diagnostic coordinate

$$n_m = \left\lceil \frac{\log(8m)}{2\gamma^2} \right\rceil$$

times and take its majority. Hoeffding's inequality and a union bound give the full signature correctly with probability at least $7/8$. Query every witness in its cell, using horizon $mn_m + L = O_\eta(m\log(m+1) + L)$. This only needs to hit, not recognize, the true witness. For $m = 0$, scan all witnesses. Each nonconstant column is balanced.

H.3 Surplus diagnostic pairs

Lower bound. Write $m = h + b$, $b \ge 1$, and $x = h/(b+1)$. The envelope becomes

$$\frac{1}{t}\sum_j \mathbb{P}_{\mathbf{1}_{I_j}}\{\text{hit } c_j \text{ by } q\} \le q\, 2^b \exp\Big(-\frac{m}{2}e^{-\ell q/m}\Big).$$

If $x < 256$, Lemma G.1 gives

$$\mathcal{H}_\eta(E) \ge \log_2(1 + 3t/4) \ge h/2 \ge \frac{h\log(2+x)}{2\log 258}.$$

For $x \ge 256$, set

$$\alpha_\eta = \frac{1}{4(1+\ell)}, \qquad q_* = \alpha_\eta h \log(2+x).$$

Since $m \geq h$ and $\ell\alpha_\eta h/m \leq 1/4$, the log envelope at q_* is at most

$$\log h + \log\log(2+x) + b\log 2 - \frac{h}{2(2+x)^{1/4}}.$$

Divide by h, use $h \geq x \geq 256$, decreasing $(\log h)/h$, and $b/h \leq 1/x$, to bound this by

$$\frac{\log x + \log\log(2+x) + \log 2}{x} - \frac{1}{2(2+x)^{1/4}}.$$

The numerator is at most $2\log x$, since $2\log(2+x) \leq x$. Also

$$\frac{x}{2(2+x)^{1/4}} \geq \frac{x^{3/4}}{3} \geq 3\log x.$$

The final inequality holds at 256 and persists because $x^{3/4}/\log x$ increases for $x > e^{4/3}$. The log envelope is therefore at most $-(h/x)\log x \leq -\log 256$. Monotonicity in q rules out every integer $q \leq q_*$, proving $\Omega_\eta(h\log(2+x))$. This includes $b > h$; constant column padding again covers budgets of at most m pairs.

A balanced systematic code. For $1 \leq b \leq h$, put $m = h + b$ and

$$A = \log_2(em/b), \qquad s = \left\lfloor \frac{b}{8A} \right\rfloor, \qquad d = s + 1.$$

Choose a binary generator $[I_h \mid P]$, with the b columns of P drawn independently and uniformly from the nonzero h-bit vectors. Every column is balanced over all 2^h messages. For a fixed nonzero message of weight i, its parity bits are independent with success probability $2^{h-1}/(2^h - 1) \geq 1/2$. Their sum stochastically dominates $\mathrm{Bin}(b, 1/2)$. The expected number of nonzero codewords of weight below d is consequently at most

$$2^{-b} \sum_{i=1}^{\min\{h,s\}} \binom{h}{i} \sum_{v=0}^{s-i} \binom{b}{v} \leq 2^{-b} \sum_{v=0}^{s} \binom{m}{v}.$$

The last inequality is Vandermonde's identity after adding nonnegative terms. If $s = 0$, this bound is at most $1/2$. Otherwise the elementary volume bound $\sum_{v=0}^{s}\binom{m}{v} \leq (em/s)^s$ applies, as in Gilbert's counting argument (Gilbert, 1952). The function $z\log_2(em/z)$ increases for $z \leq m$. As $A \geq 2$ and $3 + \log_2 A \leq 3A$,

$$s\log_2(em/s) \leq \frac{b}{8A}(A + 3 + \log_2 A) \leq b/2.$$

Thus the expected deficient count is at most $2^{-b/2}$; $s \geq 1$ entails $b \geq 16$, making it at most $1/256$. With probability at least $1/2$, the code has distance

$$d \geq \max\{1, b/(8A)\}.$$

Its rows form the diagnostic table. Checking the $t - 1$ nonzero codeword weights and resampling on failure constructs such a table in expected polynomial time in the explicitly listed size t, h, m, not in h alone.

Read every diagnostic coordinate

$$n = \left\lceil \frac{\log(8m/d)}{2\gamma^2} \right\rceil$$

times and take majorities. Each coordinate is wrong with probability at most $d/(8m)$; the expected error count is at most $d/8$. Markov's inequality gives fewer than $d/2$ errors with probability at least $3/4$. Nearest-codeword decoding then recovers the message, and one further query hits its witness. For $x = h/(b+1) \geq 1/2$, use $h/b \leq 2x$, $m \leq 2h$, and

$$A \leq \log_2(4ex), \qquad m/d \leq 8Am/b \leq 32x\log_2(4ex).$$

It follows uniformly that $\log(8m/d) = O(\log(2+x))$; the horizon $mn + 1$ is $O_\eta(h\log(2+x))$. For $b > h$, use only $2h$ pairs: the case $b = h$ costs $O_\eta(h)$, comparable to the claimed expression throughout this range. This completes the hitting upper bound.

H.4 Conversion to testing and uniform constants

Theorem 4.1 holds with constants independent of the table. Taking its infimum over at most m pairs and substituting the matching hitting bounds proves the testing law, including finite mean stopping at every legal state. The calculations above give

$$\frac{\Psi(h,m)}{256(1+\ell_\eta)} \le \inf_E \mathcal{H}_\eta(E) \le \frac{K}{(1/2-\eta)^2}\Psi(h,m)$$

for an absolute K; they do not imply a noise uniform approximation ratio as $\eta \uparrow 1/2$. For $k \ge t$, $\Psi(h,m) = O(t)$ uniformly: if $m \le h$, use $m\log(m+1) \le 2^{m+1} \le 2t$; if $m > h$, use $\Psi(h,m) \le h\log(h+2) = O(2^h)$. Since $a \ge \log 4$, the direct branch is absorbed up to constants and the rate becomes $\Theta_\eta(\Psi + a)$.

I Redundant witnesses and failed rules

We prove Theorem 6.1. Observations use the known fresh BSC of (3); every response, including repetitions and screening responses, costs one query. Constants in O_η, Ω_η and Θ_η depend only on η.

I.1 The graph and two facts

Fix integers $k, t, w \ge 1$, $m \ge 0$, and a table $E : [t] \to \{0,1\}^m$, allowing repeated rows. In this appendix witnesses are indexed by j and diagnostic pairs by p. The vertices of $G_{E,w}$ are the targets $S = \{u_1, \dots, u_k\}$, the diagnostics $d_{p,b}$, and disjoint leaf groups $C_j = \{c_{j,s} : s \in [w]\}$, $j \in [t]$. Its original edges are $(u_i, c_{j,s})$ for all i, j, s and $(d_{p,E(j)_p}, c_{j,s})$ for all p, j, s. For a state x, put

$$B_j(x) = \sum_{(v,c)\in\mathcal{E}(G_{E,w}):\, c\in C_j} (1-x_v)x_c, \qquad \mathcal{Y}_r = \{x : B_j(x) \le r \text{ for all } j\}. \tag{52}$$

Since $B_{G_{E,w}} = \sum_j B_j$, we have $\mathcal{X}_r(G_{E,w}) \subseteq \mathcal{Y}_r$; the inclusion can be strict. The favorable profiles $F_j = S \cup C_j \cup \{d_{p,E(j)_p}\}$ are exact rule ideals, and $\mathfrak{D}_\mathcal{X}$, $\mathfrak{D}_\mathcal{Y}$ are the infima of $\max_j \mathbb{E}_{\mathbf{1}_{F_j}}\tau$ over procedures that do not know j, stop almost surely and have both errors at most δ on $\mathcal{X}_r(G_{E,w})$ and on $\mathcal{Y}_r$, respectively. Every $\mathcal{Y}_r$-valid procedure is $\mathcal{X}_r$-valid with the same favorable cost, so $\mathfrak{D}_\mathcal{X} \le \mathfrak{D}_\mathcal{Y}$.

Null mass. If s_j counts one leaves in C_j, z_S zero targets, and $z_{D,j}$ zero diagnostic parents named by row j, then $B_j(x) = s_j(z_S + z_{D,j})$. Hence at every null state of $\mathcal{Y}_r$, and therefore of $\mathcal{X}_r$,

$$s_j \le r \qquad (j \in [t]), \tag{53}$$

so every group has at least $w - r$ zero leaves.

Lemma I.1 (Group hitting)**.** *The smallest horizon at which a randomized diagnostic or leaf policy queries some member of C_j with probability at least $3/4$ at each $\mathbf{1}_{F_j}$ equals $\mathcal{H}_\eta(E)$.*

Proof. On the favorable families, the label of every $c_{j,s}$ equals the label of the single witness c_j. To simulate a grouped policy on the original graph, replace each query to $c_{j,s}$ by one fresh query to c_j and keep the simulated member in the policy's history. Conversely, replace queries to c_j by queries to a fixed representative $c_{j,1}$. Conditional response laws agree at every step, even for repeated and adaptive queries, so transcripts and hitting probabilities agree at equal cost. The statement concerns the favorable family only; general states may have unequal labels within a group. □

I.2 Lower bounds on the global class

Let $r < w$, put $R = w/(w-r)$, $b_\delta = \mathrm{kl}(1-\delta, \delta)$ and $K_\delta = \mathrm{kl}(1/2, \delta)$, and fix a procedure valid on $\mathcal{X}_r$.

Confidence. For $i \in [k]$ and $D \subseteq C_j$ with $|D| = w - r$, the state $x^{i,D} = \mathbf{1}_{F_j\setminus(\{u_i\}\cup D)}$ is a null whose only failed edges run from u_i to the r leaves left in C_j. Stopped change of measure, justified as in Appendix D.3 through $\tau \wedge n$ without assuming a finite mean at the comparison state, gives

$$\kappa\Big(\mathbb{E}_{\mathbf{1}_{F_j}} N_{u_i}(\tau) + \sum_{c\in D} \mathbb{E}_{\mathbf{1}_{F_j}} N_c(\tau)\Big) \ge b_\delta.$$

With $N_S = \sum_i \mathbb{E}N_{u_i}$ and $N_C = \sum_{c \in C_j} \mathbb{E}N_c$, averaging over i and uniformly over D gives $N_S/k + N_C/R \geq b_\delta/\kappa$, hence $\mathbb{E}_{\mathbf{1}_{F_j}}\tau \geq \min\{k, R\}\, b_\delta/\kappa$. Unit mass on the targets, or mass $1/(w-r)$ on every leaf of C_j, is a feasible allocation by (53), so the supplied profile coefficient is exactly $\min\{k, R\}$.

Discovery. Let $T = \max_j \mathbb{E}_{\mathbf{1}_{F_j}}\tau < \infty$. Correctness before the first query gives $T \geq 1 - 2\delta \geq 1/2$, so $Q = \lceil 8T \rceil \leq 10T$. Remove the whole active group to get the positive root only state $z_j = \mathbf{1}_{F_j \setminus C_j}$; each $z_j - e_{u_i}$ is a legal null. If $\mathbb{P}_{z_j}\{\tau \leq Q, \widehat{h} = 1\} \geq 1/2$ for some j, the finite horizon change of measure against these nulls forces $\kappa Q \geq kK_\delta$. Otherwise, at every j, timely favorable acceptance is at least $1 - \delta - T/Q \geq 5/8$, and coupling F_j with z_j until the first query to C_j gives a hit probability above $1/8$. Eleven independent truncated runs hit with probability at least $1 - (7/8)^{11} > 3/4$; simulating target responses internally and using Lemma I.1 gives a discovery policy with at most $11Q \leq 110T$ queries. Hence $T \geq \min\{kK_\delta/(10\kappa), \mathcal{H}_\eta(E)/110\}$.

Both divergences are uniformly comparable to a for $0 < \delta \leq 1/4$. With $A' = ka$ and $B' = Ra$, $\max\{\min(A', \mathcal{H}), \min(A', B')\} \geq \frac{1}{2}\min\{A', \mathcal{H} + B'\}$, which proves the lower bound in (18) for $\mathfrak{D}_\mathcal{X}$, hence also for $\mathfrak{D}_\mathcal{Y}$. For $r \geq w$, removing only a target from F_j gives a legal null with w failures, and the k single coordinate constraints yield $\Omega_\eta(ka)$, matching direct testing.

I.3 Upper bound on the per group class

Write $H = \mathcal{H}_\eta(E)$ and fix a policy attaining this integer horizon; by Lemma I.1, a fixed representative $c_{j,1}$ per group simulates it on every favorable profile.

Selecting one group. Run a fresh copy of the policy for at most H search queries. After its first query in a group, supply the ordinary response, pause, and screen that representative with the one node test of Lemma D.1, using fresh data never supplied to the search and boundaries $-\log 4$ and $\log(16H)$. A lower exit resumes the search; later queries to that group trigger no further screen. The first upper exit selects its group. At F_j, build the complete potential search transcript and an independent potential screen for every group. The probability that some empty group on the transcript passes is at most $H/(16H) = 1/16$, and the search hits C_j and its screen passes with probability at least $\frac{3}{4} \cdot \frac{3}{4}$. Hence

$$\mathbb{P}_{\mathbf{1}_{F_j}}\{\text{the selected group is } C_j\} \geq \tfrac{9}{16} - \tfrac{1}{16} = \tfrac{1}{2}. \tag{54}$$

The search uses at most H queries, each empty screen costs $O_\eta(1)$ in expectation and the true group is screened once at cost $O_\eta(\log(16H))$, so selection costs $O_\eta(H)$ at F_j and has finite mean at every state.

Validating it with fresh observations. For $0 < \epsilon \leq 1/4$ put $N = \lceil 2R\log(1/\epsilon) \rceil$. In each of at most N blocks, draw a uniform leaf of the selected group and test it with fresh observations at boundaries $-\log(4N)$ and $\log 2$. A lower exit abandons the group without a decision; N upper exits certify the conjunction. At any null of $\mathcal{Y}_r$, condition on the complete selection history: the selected group is fixed and has at least $w - r$ zero leaves by (53), each block draws a zero with probability at least $1/R$, and that test exits upward with probability at most $1/2$. Therefore

$$\mathbb{P}_x\{\text{validation accepts} \mid \text{selection history}\} \leq \left(1 - \frac{1}{2R}\right)^N \leq \epsilon. \tag{55}$$

One validation runs per cycle, so no union over H or t groups is needed. On an all one group, a union bound over the blocks gives acceptance probability at least $3/4$ at expected cost $O_\eta(N)$; on an all zero group, validation ends after at most two expected blocks.

Error spending and the race. In cycle $l \geq 1$, run a fresh selection and validate with $\epsilon_l = 3\delta/(\pi^2 l^2)$. Conditional on any history at a null, cycle l accepts with probability at most ϵ_l, so this certifier ever accepts with probability at most $\delta/2$. At F_j, a reached cycle selects and validates the true group with probability at least $3/8$ at conditional cost $O_\eta(H + R[a + \log(l+1)])$, so Tonelli's theorem gives $\max_j \mathbb{E}_{\mathbf{1}_{F_j}} T_{\text{cert}} = O_\eta(H + Ra)$. Finally alternate individual queries between this certifier and the fresh direct target test of Lemma D.2, whose two errors are at most $\delta/2$, and stop at the first decision. Null error is at most δ; on positive states only the direct test rejects. The race uses at most twice either routine's queries, so it stops in finite expectation on $\mathcal{Y}_r$ and has favorable cost $O_\eta(\min\{ka, H + Ra\})$. This proves the upper bound for $\mathfrak{D}_\mathcal{Y}$, hence for $\mathfrak{D}_\mathcal{X}$.

Optimized tables. All constants are uniform in E, so taking the infimum over tables with $t = 2^h$ witnesses and at most m pairs and applying Theorem 5.1 replaces $\mathcal{H}_\eta(E)$ by $\Psi(h, m)$. The comparison of the two branches is an order comparison, not an exact policy threshold.

J CHANNELS WITH A KNOWN MARGIN

This appendix proves Theorem 6.2, an estimated margin corollary, and Proposition J.4, which covers sampled outputs. Fix $0 < \bar{\eta} < 1/2$ and write $\ell = \ell_{\bar{\eta}}$ and $\kappa = (1 - 2\bar{\eta})\ell$. Throughout, the direct test means the fresh direct target test of Lemma D.2, whose two errors are at most $\delta/2$.

The class. Let $\mathcal{F}_{n-1}$ record all nodes queried and responses received before query n, together with the internal randomness the procedure has drawn up to the choice of query n; randomness drawn later is not included. A member of $\mathcal{G}(\bar{\eta})$ is a conditional response law under which, at every state x, $Y_n \sim \mathrm{Bern}(p_n)$ conditionally on $\mathcal{F}_{n-1}$, where p_n is $\mathcal{F}_{n-1}$-measurable, may depend on x and on n, and satisfies $p_n \geq 1-\bar{\eta}$ if $x_{A_n} = 1$ and $p_n \leq \bar{\eta}$ if $x_{A_n} = 0$. The class is thus defined relative to when a procedure draws its randomness; the infima below range over implementations, and the designs of Theorem 5.1 are deterministic. The BSC with parameter $\bar{\eta}$ is the member with $p_n \in \{\bar{\eta}, 1 - \bar{\eta}\}$. Two closure properties are used. First, if a subprocedure chooses its own queries, using only its own randomness, and observes only its own responses, then given its own past the success probability of its next response is a conditional expectation of p_n; the event $\{x_{A_n} = 1\}$ is determined by that past, so the conditional expectation obeys the same bounds, and the transcript of the subprocedure is generated by a member. Second, membership is a pointwise condition on (p_n), so a law that obeys the bounds before a time σ such that $\{\sigma = n\}$ is determined before the nth response, and follows the BSC or any other member from query σ on, is a member.

The hitting cost $\mathcal{H}^{\mathrm{gap}}_{\bar{\eta}}(E)$ ranges over procedures that query only diagnostics and witnesses, at most q of them, as in Definition 2.1; target queries, which the symmetric channel lets a search simulate, are not available to it here. A procedure meeting the margin requirement meets it under the BSC member, and querying every witness once hits under every member, so $\mathcal{H}_{\bar{\eta}}(E) \leq \mathcal{H}^{\mathrm{gap}}_{\bar{\eta}}(E) \leq t$; forbidding target queries can only increase $\mathcal{H}^{\mathrm{gap}}_{\bar{\eta}}(E)$, and the lower bound below uses $\mathcal{H}_{\bar{\eta}}(E)$.

Lemma J.1 (One node test under the margin)**.** *Run the test of Lemma D.1 with thresholds $u, b > 0$ and increments $\pm\ell$ on the responses of any member of $\mathcal{G}(\bar{\eta})$ at a node with label β, starting at any stopping time and possibly interleaved with other queries. Conditionally on the past at its start, the error bounds* (25) *and the duration bounds* (26) *hold with $x = \beta$.*

Proof. Let $\Delta_i = \ell(2Y_i - 1)$ be the increments and p_i the conditional success probabilities. If $\beta = 0$, then $\mathbb{E}[e^{\Delta_i} \mid \text{past}] = p_i e^{\ell} + (1 - p_i)e^{-\ell}$ is increasing in p_i and equals 1 at $p_i = \bar{\eta}$, hence is at most 1, so $\exp(L_n)$ is a nonnegative supermartingale with initial value 1. Optional stopping at $T \wedge N$ followed by $N \to \infty$ gives $\mathbb{P}_0(T < \infty,\ L_T \geq u) \leq e^{-u}$, which is Ville's inequality. If $\beta = 1$, then $\mathbb{E}[e^{-\Delta_i} \mid \text{past}] = p_i e^{-\ell} + (1-p_i)e^{\ell}$ is decreasing in p_i and equals 1 at $p_i = 1-\bar{\eta}$, so $\exp(-L_n)$ is a supermartingale and $\mathbb{P}_1(L_T \leq -b) \leq e^{-b}$. For the durations, $\mathbb{E}[\Delta_i \mid \text{past}] = \ell(2p_i - 1)$ is at least κ if $\beta = 1$ and at most $-\kappa$ if $\beta = 0$. Since $|\Delta_i| \leq \ell$, the process $L_n - \sum_{i \leq n} \ell(2p_i - 1)$ is a martingale with bounded increments, and optional stopping at $T \wedge N$ gives $\mathbb{E}L_{T\wedge N} = \mathbb{E}\sum_{i \leq T\wedge N} \ell(2p_i - 1)$. If $\beta = 1$, the right side is at least $\kappa\mathbb{E}(T \wedge N)$ and the left side is at most $u + \ell$, so monotone convergence gives $\mathbb{E}_1 T \leq (u + \ell)/\kappa$; in particular $T < \infty$ almost surely. If $\beta = 0$, the same argument with $-L_{T\wedge N} \leq b + \ell$ gives $\mathbb{E}_0 T \leq (b + \ell)/\kappa$. Every step holds conditionally on the past at the start of the test, because the supermartingale and martingale properties hold with respect to the full filtration. □

The direct test keeps its error and duration bounds under every member. Its proof in Appendix D.2 pregenerates observation sequences, which a past dependent member does not permit; instead apply Lemma J.1 at the start of each target test. Returning 1 at a state with $x_s = 0$ requires the test at s, if reached, to exit upward, which has conditional probability at most $e^{-a_\delta} = \delta/2$. Returning 0 at a state with all targets one requires a lower exit at one of the k tests, of total probability at most $ke^{-\log(2k/\delta)} = \delta/2$, and the expected duration at such a state is at most $k(a_\delta + \ell)/\kappa$. At every state each test has conditional expected duration at most $(\log(2k/\delta) + \ell)/\kappa$, so the direct test stops in finite expectation everywhere.

Proof of Theorem 6.2. Lower bound. A procedure that is δ-correct on $\mathcal{X}_0(G_E)$ under every member is δ-correct under the BSC member, and its worst case over members of $\max_v \mathbb{E}_{\mathbf{1}_{I_v}} \tau$ is at least its cost under that member. Theorem 4.1 at $\eta = \bar{\eta}$ gives the left inequality of (19).

Upper bound. Put $H = \mathcal{H}^{\mathrm{gap}}_{\bar{\eta}}(E)$ and fix a procedure attaining it. Run the certifier of Appendix G with this procedure as the fresh search experiment of each cycle, increments $\pm\ell$, boundaries $-\log 2$ and $u_l = \log(\pi^2 H l^2/(3\delta))$ for the witness checks, and alternate its queries with those of the direct test. The analysis of Appendix G used three facts about the BSC, each of which holds conditionally on the past under every member. First, every legal null has all witnesses zero, and a check at a zero witness accepts with conditional probability at most e^{-u_l} by Lemma J.1; a cycle makes at most H checks, so the certifier ever accepts at a null with probability at most $\sum_l H e^{-u_l} = \delta/2$, and the direct test adds at most $\delta/2$. On positive states only the direct test can reject, with probability at most $\delta/2$. Second, fix $\mathbf{1}_{I_v}$ and the past at the start of a cycle, before the cycle draws its randomness, and let the cycle's search run to its end even if the certifier accepts first, with responses from any member after the acceptance; this continuation serves the proof only and makes no queries. By the two closure properties its transcript is generated by a member, so it queries c_v with conditional probability at least $3/4$. On that event the cycle either has already accepted, which is correct at this positive profile, or reaches its query at c_v, where the check exits downward with conditional probability at most $1/2$ and stops almost surely. A reached cycle therefore accepts with conditional probability at least $3/8$, and cycle l is reached with probability at most $(5/8)^{l-1}$. Third, a reached cycle makes at most H search queries, at most H checks at zero witnesses, each of conditional expected duration at most $(\log 2 + \ell)/\kappa$, and at most one check at c_v, of conditional expected duration at most $(u_l + \ell)/\kappa$. Summing against the reach probabilities gives $O_{\bar{\eta}}(H + a)$. Continue both procedures virtually after the race stops, with responses from any member; by the second closure property the extended law is a member, and by the first the stopping times τ_{C} and τ_{D} of the certifier and the direct test have the laws analysed above. The race makes at most $2\min\{\tau_{\mathrm{C}}, \tau_{\mathrm{D}}\}$ queries, and $\mathbb{E}\min\{\tau_{\mathrm{C}}, \tau_{\mathrm{D}}\} \le \min\{\mathbb{E}\tau_{\mathrm{C}}, \mathbb{E}\tau_{\mathrm{D}}\}$, so the procedure costs $O_{\bar{\eta}}(\min\{ka, H + a\})$ at every favorable profile under every member and stops in finite expectation at every legal state. This is the right inequality of (19).

Optimized tables. Since $\mathcal{H}_{\bar{\eta}} \le \mathcal{H}^{\mathrm{gap}}_{\bar{\eta}}$, the lower bound $\inf_E \mathcal{H}_{\bar{\eta}}(E) = \Omega_{\bar{\eta}}(\Psi(h, m))$ of Appendix H bounds $\inf_E \mathcal{H}^{\mathrm{gap}}_{\bar{\eta}}(E)$ from below. The designs of Appendix H run unchanged; only their majority votes need a new argument. Consider n responses $Y_1, \ldots, Y_n$ at one diagnostic node with conditional success probabilities $p_1, \ldots, p_n$, and put $D_r = Y_r - p_r$. These are martingale differences with $D_r \in [-p_r, 1 - p_r]$, an interval of length one determined by the past, so Hoeffding's lemma applied conditionally gives $\mathbb{E}\exp(\lambda \sum_r D_r) \le e^{n\lambda^2/8}$ for every real λ, and Markov's inequality with $\lambda = \mp 4s/n$ bounds both $\mathbb{P}\{\sum_r D_r \le -s\}$ and $\mathbb{P}\{\sum_r D_r \ge s\}$ by $\exp(-2s^2/n)$. A wrong or tied majority at a node with label one requires $\sum_r D_r \le -n(1/2 - \bar{\eta})$, and at a node with label zero it requires $\sum_r D_r \ge n(1/2 - \bar{\eta})$; either has probability at most $\exp(-2n(1/2 - \bar{\eta})^2)$, the bound that Hoeffding's inequality supplied under the BSC. The grouped design therefore reads the full signature correctly with probability at least $7/8$; for the systematic code, the expected number of wrong coordinates is at most $d/8$ by linearity, without any independence, so Markov's inequality and nearest codeword decoding proceed as before. Larger budgets use $2h$ pairs, and $m = 0$ scans all witnesses, which hits under every member. Hence $\inf_E \mathcal{H}^{\mathrm{gap}}_{\bar{\eta}}(E) = O_{\bar{\eta}}(\Psi(h, m))$ with the same constants, and substituting both bounds into (19) gives the optimized law. □

Proposition J.2 (Hitting under the margin)**.** *For every table, with $\gamma = 1/2 - \bar{\eta}$ and $H = \mathcal{H}_{\bar{\eta}}(E)$,*

$$\mathcal{H}^{\mathrm{gap}}_{\bar{\eta}}(E) \le 2H \left\lceil \frac{\log(16H)}{2\gamma^2} \right\rceil = O_{\bar{\eta}}\big(H(1 + \log H)\big),$$

so $\mathfrak{D}^{\mathrm{gap}}_{E,k}(\delta) \le C_{\bar{\eta}} \min\{ka, H(1 + \log H) + a\}$.

Read every node a symmetric channel policy would query $\lceil \log(16H)/(2\gamma^2) \rceil$ times, and feed the policy the majority vote flipped by a fresh $\mathrm{Bern}(\bar{\eta})$ bit. Conditionally on the past, the fed bit is within total variation $1/(16H)$ of the symmetric response, so a stepwise coupling keeps the policy's view on its symmetric channel trajectory with probability at least $15/16$; two successive runs, the second bounded conditionally on the first, hit with probability above $3/4$ (proof below). Whether the logarithm can be removed is open.

Proof of Proposition J.2. Let π attain $H = \mathcal{H}_{\bar{\eta}}(E)$ on the symmetric channel with parameter $\bar{\eta}$. Target responses have a known law on the favorable family, so π may be taken to query only diagnostics and witnesses. Put $\rho = \lceil \log(16H)/(2\gamma^2) \rceil$ and define π': whenever π queries a node v, query v exactly ρ times, let M be the majority of the responses with ties broken arbitrarily, draw an independent $Z \sim \mathrm{Bern}(\bar{\eta})$, and return $M \oplus Z$ to π. Fix a member of $\mathcal{G}(\bar{\eta})$ and a favorable profile $\mathbf{1}_{I_v}$, and condition on the entire past at the start of the ρ reads. If v has label β, the ρ responses have conditional success probabilities on the side of β, so the martingale form of Hoeffding's inequality used in the proof of Theorem 6.2, which already covers a wrong or tied majority, gives $\mathbb{P}(M \neq \beta \mid \text{past}) \leq \exp(-2\rho\gamma^2) \leq 1/(16H)$. On $\{M = \beta\}$ the returned bit has conditional law $\beta \oplus \mathrm{Bern}(\bar{\eta})$, the symmetric channel response, because Z is fresh. Hence the conditional law of the bit returned to π is within total variation $1/(16H)$ of the symmetric channel law at v, whatever the member and the past. Couple the run of π' with a run of π on the symmetric channel, using the same internal seed, step by step: at each of the at most H steps, the two runs agree so far and their next returned bits can be coupled to agree except on an event of conditional probability at most $1/(16H)$. The two transcripts therefore coincide with probability at least $15/16$, and on that event π' queries c_v whenever π does. Thus π' hits c_v with conditional probability at least $3/4 - 1/16 = 11/16$ given any past, under every member, using at most $H\rho$ queries. Run π' twice with fresh internal randomness; the second run's bound holds conditionally on the first, so the two runs miss together with probability at most $(5/16)^2$, and some run hits with probability at least $1 - (5/16)^2 > 3/4$ using at most $2H\rho$ queries. This proves the bound on $\mathcal{H}^{\mathrm{gap}}_{\bar{\eta}}(E)$. The members allow p_n to depend on randomness drawn before query n; the coupling uses only conditional laws given that past, and Z is drawn after the ρ reads, so the argument covers them. Substituting into the right side of (19) gives the testing bound. □

Corollary J.3 (Estimated margin)**.** *Suppose an uncounted calibration phase, whose transcript generates $\mathcal{F}_{\mathrm{cal}}$, returns $0 < \widehat{\eta} < 1/2$. Let Γ be the event that every later response has p_n on the side of the $\widehat{\eta}$ margin that its label requires, and suppose that, for the procedure being run, $\mathbb{P}_x(\Gamma) \geq 1 - \epsilon_{\mathrm{cal}}$ at every $x \in \mathcal{X}_0(G_E)$. Then the procedures of Theorem 6.2 designed for $\widehat{\eta}$ err or fail to stop with probability at most $\delta + \epsilon_{\mathrm{cal}}$ at every legal state. Their truncated cost $\mathbb{E}_x[\tau \mathbf{1}_\Gamma \mid \mathcal{F}_{\mathrm{cal}}]$ is finite at every legal state and at most $C_{\widehat{\eta}} \min\{ka, \mathcal{H}^{\mathrm{gap}}_{\widehat{\eta}}(E) + a\}$ at every favorable profile. Bounding the conditional mean $\mathbb{E}_x[\tau \mid \Gamma, \mathcal{F}_{\mathrm{cal}}]$ would also need a lower bound on $\mathbb{P}_x(\Gamma \mid \mathcal{F}_{\mathrm{cal}})$.*

Proof. Fix the calibration transcript, so that $\widehat{\eta}$ and the procedure are fixed. Let σ be the first later query at which $p_n < 1 - \widehat{\eta}$ at a node with label one or $p_n > \widehat{\eta}$ at a node with label zero. Since p_n is determined by the past, the event $\{\sigma = n\}$ is determined before the nth response, and $\Gamma = \{\sigma = \infty\}$. The law that follows the actual channel before σ and the BSC with parameter $\widehat{\eta}$ from query σ on is a member of $\mathcal{G}(\widehat{\eta})$ by the second closure property. Generating both experiments from the same seed and the same uniform variables makes their transcripts, stopping times and decisions coincide on Γ. Theorem 6.2 for the switched law therefore bounds the probability of error or nonstopping jointly with Γ by δ, and bounds $\mathbb{E}_x[\tau \mathbf{1}_\Gamma \mid \mathcal{F}_{\mathrm{cal}}]$ by the switched expectation. Adding $\mathbb{P}_x(\Gamma^c) \leq \epsilon_{\mathrm{cal}}$ and integrating over the calibration transcript proves the claim. □

J.1 Sampled outputs graded by exact checkers

Let the system's output on a fixed probe be a sample ω from a fixed law, drawn afresh at every query, and let $g_v(\omega) \in \{0, 1\}$ be the grade of contract v. Contract design imposes three pointwise inclusions: $g_{c_v} \leq g_{u_i}$ for every target, $g_{c_v} \leq g_{d_{j,E(v)_j}}$ for every diagnostic named by row $E(v)$, and $g_{d_{j,0}} + g_{d_{j,1}} \leq 1$ for every pair, because the two members check incompatible properties of the same output. With $p_v = \mathbb{E} g_v(\omega)$ these give $p_{c_v} \leq p_{u_i}$, $p_{c_v} \leq p_{d_{j,E(v)_j}}$ and $p_{d_{j,0}} + p_{d_{j,1}} \leq 1$. Suppose $p_v \notin (\bar{\eta}, 1 - \bar{\eta})$ at every node and put $x_v = \mathbf{1}\{p_v \geq 1 - \bar{\eta}\}$. If $x_{c_v} = 1$, every contract that c_v contains has pass rate at least $1 - \bar{\eta}$, so $x \in \mathcal{K}_0(G)$. If $x_{d_{j,1}} = 1$, then $p_{d_{j,0}} \leq \bar{\eta}$, so at most one member of each pair is mastered, a state with $x_{c_v} = 1$ masters exactly the diagnostics of row $E(v)$, and any other mastered witness has the same row. A query at A_n returns $Y_n = g_{A_n}(\omega_n)$ with ω_n independent of the past, so $\mathbb{P}(Y_n = 1 \mid \text{past}) = p_{A_n}$: a member of $\mathcal{G}(\bar{\eta})$ with $p_n = p_{A_n}$. Richness requires every grade vector satisfying the witness inclusions and pair incompatibility to be produced by some output; the next construction realizes these vectors. This condition matters: if all target checkers are identical pointwise, $g_{u_1} = \cdots = g_{u_k}$, their labels agree in every sampled system, and a single test at u_1 decides the claim with $O_{\bar{\eta}}(a)$ queries, whereas the law grows with k and t when

$m = 0$. Grading one sample against several contracts at once returns correlated grades of one ω; the query model charges one grade per fresh sample and does not cover that batch observation. Persistent noise is a different obstruction: with one node and one stored corrupted bit $Y = x \oplus Z$, rereading reveals only Y; under an equal prior on x the optimal error is η, and the decision Y attains it at both states, so the minimax error stays η at any reread budget.

J.2 A PAPER INSTANCE: A PROGRAM SYNTHESIS PROBE

The following contract set is written out to show where each assumption of Section 6.2 enters; nothing is measured. Fix a probe: a program synthesis task whose specification asks for a source file with four parts, a header line naming the function, a documentation block, a metadata record with keys $f_1, \dots, f_m$ and a list L, and the function body. Each query samples one fresh output of the system on this probe and grades one contract. Every checker extracts its own part leniently, by pattern matching on the raw text, and fails only when that part is wrong or absent; the test harness locates the function body itself, so a malformed header fails neither a key check nor a test case. The graded parts can thus be set separately.

- *Targets* ($k = 3$): u_1, the header names the required function; u_2, the documentation block is present; u_3, the list L has the required length.
- *Diagnostic pairs* ($j \in [m]$): $d_{j,0}$ passes when key f_j has the value α_j, $d_{j,1}$ when it has the value $\beta_j \neq \alpha_j$. The two checks are incompatible on every output, so $g_{d_{j,0}} + g_{d_{j,1}} \leq 1$ pointwise; both fail when f_j is absent or takes a third value, which the legal class permits.
- *Witnesses* ($v \in [t]$): c_v passes when all three targets pass, key f_j has the value named by $E(v)_j$ for every j, and the function body returns 0 on input v. The t test inputs are distinct, and the body may implement a finite lookup table returning 0 or 1 at each input. Thus $g_{c_v} \leq g_{u_i}$ and $g_{c_v} \leq g_{d_{j,E(v)_j}}$ pointwise. The target fields, keys and individual test outcomes can be set separately; the realization argument below shows that there are no constraints beyond these inclusions and pair incompatibility.

Write p_v for the pass rate of contract v over the system's output distribution on the probe. The gap promise, $p_v \notin (\bar{\eta}, 1 - \bar{\eta})$ for every contract, is a statement about the system. It holds, for example, when the system is consistent on the probe: on all but a fraction at most $\bar{\eta}$ of its samples, jointly, it produces the same header, documentation and list status, the same key values, and the same test outcomes; every contract then takes one value on those samples. It fails when some key is split between two values with frequencies inside the margin; that node then has no state and the theory says nothing about it. Under the promise the state $x_v = \mathbf{1}\{p_v \geq 1 - \bar{\eta}\}$ is legal and masters at most one member of each pair, by Appendix J.1. It equals the favorable profile I_v exactly when c_v passes at rate at least $1 - \bar{\eta}$ and every other witness v' with the same row $E(v') = E(v)$ passes at rate at most $\bar{\eta}$. Indeed the first condition forces the targets and the keys of row $E(v)$ to pass at rate at least $1 - \bar{\eta}$, hence the complementary keys to pass at rate at most $\bar{\eta}$; a witness of another row then fails at rate at least $1 - \bar{\eta}$ through a key mismatch, whatever its test case does; and the second condition removes the remaining witnesses. The complementary failures thus come from the key values, not from an assumption on the population. Richness, the hypothesis of Proposition J.4, holds by construction: given any grade vector obeying the two inclusion rules and the pair incompatibility, set the header, documentation and list to match the target grades, set each key to α_j, β_j or a third value according to its pair, and let the body return 0 at input v when $g_{c_v} = 1$, and 1 otherwise. For those v with $g_{c_v} = 1$, the included parts pass by the inclusions, so the conjunction defining c_v equals the prescribed grade.

Sharing a component or parser can restrict realizable grade patterns, but need not do so; the actual predicates determine whether richness holds. Without richness, the upper bound still follows from the checker constraints and the gap, while a converse needs separate justification. For example, pointwise-identical target checkers reduce the claim to one test costing $O_{\bar{\eta}}(a)$. Grading one sample against all contracts at once returns correlated grades of one output, which the query model does not cover; each query is one fresh sample and one grade.

Proposition J.4 (Contract sampled systems)**.** *Call a response law a sampled system if it arises as above with every pass rate outside $(\bar{\eta}, 1 - \bar{\eta})$, and assume that every grade vector obeying the three inclusions is the grade vector of some output. Let $\mathfrak{D}^{\mathrm{samp}}_{E,k}(\delta)$ be defined as $\mathfrak{D}^{\mathrm{gap}}_{E,k}(\delta)$ with sampled systems in place of members. Then (19) holds with $\mathfrak{D}^{\mathrm{samp}}_{E,k}(\delta)$ in place of $\mathfrak{D}^{\mathrm{gap}}_{E,k}(\delta)$,*

under the same quantifiers, and over tables with $t = 2^h$ witnesses, $h \geq 1$, and at most m pairs, $\inf_E \mathfrak{D}^{\mathrm{samp}}_{E,k}(\delta) = \Theta_{\bar{\eta}}(\min\{ka, \Psi(h,m) + a\})$.

Proof. A sampled system is a member of $\mathcal{G}(\bar{\eta})$ at a legal state, so the procedure of the upper bound is valid under every sampled system and $\mathfrak{D}^{\mathrm{samp}}_{E,k}(\delta) \leq \mathfrak{D}^{\mathrm{gap}}_{E,k}(\delta)$. The lower bound of Theorem 6.2 used the BSC member, which is not a sampled system as soon as two witnesses have different rows: for such witnesses c_v and $c_{v'}$ the inclusions give $g_{c_v} + g_{c_{v'}} \leq g_{u_1}$ for every output, whereas the BSC at $\mathbf{1}_{I_v}$ has $p_{c_v} + p_{c_{v'}} = 1 > 1 - \bar{\eta} = p_{u_1}$. We therefore run the argument of Appendix G on explicit sampled systems. Write $z_v = \mathbf{1}_{I_v \setminus \{c_v\}}$, and let o_v be the grade vector that passes exactly the diagnostics $d_{j,1-E(v)_j}$, $j \in [m]$. For $s \in \{\mathbf{1}_{I_v}, z_v, z_v - e_{u_i}\}$, where $v \in [t]$ and $i \in [k]$, let the system $P_v[s]$ produce an output with grade vector s with probability $1 - 2\bar{\eta}$, $\mathbf{1}_{I_v}$ with probability $\bar{\eta}$ and o_v with probability $\bar{\eta}$; all three vectors obey the inclusions. Under $P_v[s]$, the targets and the diagnostics of row $E(v)$ pass with probability $1 - \bar{\eta}$, except that u_i passes with probability $\bar{\eta}$ when $s = z_v - e_{u_i}$; the other diagnostics pass with probability $\bar{\eta}$; c_v passes with probability $1 - \bar{\eta}$ if $s = \mathbf{1}_{I_v}$ and $\bar{\eta}$ otherwise; and the other witnesses never pass. Hence the state of $P_v[s]$ is s. Fix a procedure that is δ-correct under every sampled system and put $T = \max_v \mathbb{E}\tau$ under $P_v[\mathbf{1}_{I_v}]$. The three properties of the BSC that the argument of Appendix G uses hold for these systems: $P_v[\mathbf{1}_{I_v}]$ and $P_v[z_v]$ differ only at c_v, which gives the coupling up to the first query at c_v; $P_v[z_v]$ and $P_v[z_v - e_{u_i}]$ differ only at u_i, with divergence κ per query, which gives the change of measure against each target deletion; and $P_v[\mathbf{1}_{I_v}]$ and $P_v[z_v - e_{u_i}]$ differ only at u_i and c_v, which gives $T \geq \mathrm{kl}(1-\delta, \delta)/\kappa$. Validity is invoked only under $P_v[\mathbf{1}_{I_v}]$ and $P_v[z_v - e_{u_i}]$. When two rows coincide, one state carries several of these systems, which is allowed because validity is required under every sampled system.

In its branch $Q < kK_\delta/\kappa$ the argument yields a procedure that queries only diagnostics and witnesses, at most $110T$ of them, and hits c_v with probability at least $3/4$ under $P_v[\mathbf{1}_{I_v}]$ for every v; its repetitions are independent and its simulated $\mathrm{Bern}(1-\bar{\eta})$ target responses are exact, because every response is a fresh draw. Under $P_v[\mathbf{1}_{I_v}]$ the diagnostics pass with the BSC probabilities and the witnesses other than c_v never pass. Reporting 0 to the procedure at every witness query therefore gives its transcript, up to its first query at c_v, the same law under the BSC with parameter $\bar{\eta}$ at $\mathbf{1}_{I_v}$, so it hits with the same probability there, and $\mathcal{H}_{\bar{\eta}}(E) \leq 110T$. The lower bound of (19) for $\mathfrak{D}^{\mathrm{samp}}_{E,k}(\delta)$ follows as in Appendix G, and the table law follows from the two bounds. □

K Summary of the experiments

The main text reports these measurements next to the laws they test; this section collects them with their settings, and Appendix L gives every design. The theory makes three kinds of prediction that data can check: how the cost of each procedure scales, whether sampled outputs of language models can realize the private witness architecture, and what happens when an assumption fails. Every design, prediction and script below was fixed, and its SHA-256 hash recorded, before its data were collected or analyzed; predictions that failed are reported as failed (Appendix L). Costs count grades: one checker invocation on one fresh sample, the query of Section 2. A replay runs a procedure on the empirical channel of a system, where a query returns one of the stored grades of the queried node, drawn uniformly and independently of the past, as in resampling studies of sampled outputs (Ota et al., 2026).

K.1 The laws on the model's channel

The first question is whether the procedures behind the upper bounds scale as the laws say. We simulate the symmetric channel with $\eta = 1/4$ at favorable profiles. The direct test is the fallback of Lemma D.2. The certifier is the construction behind Theorem 4.1, raced query by query against the direct test at level $\delta/2$; it either reads each diagnostic once per unknown address bit or scans the witnesses. An oracle that knows the mastered witness tests only that witness. Three baselines complete the comparison: the LCB sampling and Box stopping rules for the lowest of the target means (Kaufmann et al., 2018), which do not use the margin; witness search with the LUCB_G sampling rule (Kano et al., 2019), a standard baseline for returning one arm above a threshold (LI & Cheung, 2025), run with the same confidence bounds and raced against the direct test; and a fixed number of samples per target with a Bonferroni correction.

Discovery and confidence add, as Theorem 4.1 states. In Figure 2(a), the cost of direct testing grows by 30.5 grades per unit of a, and that of the certifier with diagnostics by 3.7. The ratio is close to $k/2$ rather than k because the race lets the direct test spend about as many grades as the certifier itself; the oracle's slope of 1.8 is the confidence term alone. The certifier is cheaper than direct testing, scanning, LCB + Box and LUCB_G at every δ from 10^{-1} to 10^{-8}. The baselines are valid but cost about 3 to 30 times more, because LCB + Box tests each target against $1/2$ without the margin and LUCB_G searches the witnesses without diagnostics. The same order holds on the stored grades of Appendix K.2, where the two slopes are 15.7 and 3.0. The advantage over scanning comes from the diagnostics, not from the certifier's smaller accepting boundary: given scanning's boundary, the certifier with diagnostics costs 3.2 to 10.9 grades more, at most 21% of its advantage, and stays cheaper in all nine settings of a registered ablation (Appendix L.1).

The better procedure switches where ka meets $\mathcal{H} + a$. Figure 2(b) shows direct testing cheapest for $k \leq 2$ at every h, and the certifier cheapest once k exceeds a threshold that grows from 4 at $h = 1$ to 64 at $h = 8$. Scanning wins only when there are at most four candidates; beyond that, reading the address is cheaper, so the threshold follows the cost of reading the address rather than t.

The diagnostic resource law contrasts $m = h$ with a constant factor surplus of pairs (Theorem 5.1). Figure 2(c) shows the hitting budgets of two policies, the grades with which each hits the mastered witness with probability $3/4$ at every favorable profile. Majority reads of the identity table need 5 reads per address bit at $h = 2$ and 15 from $h = 12$ on, because every bit must be read more often as their number grows. Decoding a $[2h, h]$ code with minimum distance at most 5 keeps about 8 reads per address bit from $h = 6$ to $h = 16$, and at $h = 16$ it hits with 129 grades against 241. The identity budget per bit rises in steps because reads come in odd numbers. These are budgets of two particular policies; $\mathcal{H}$ is the smallest budget over all policies, and up to $h = 6$ scanning $3t/4$ witnesses has a smaller one (Table 5). Panel (b) compares complete procedures, which also pay to check each proposal.

Validity holds wherever the model's assumptions hold. On symmetric channel positives and nulls, and on the stored grades of every population below, each procedure with a guarantee errs in at most a fraction δ of runs for $\delta \in \{0.1, 0.05, 0.02, 0.01\}$. Dropping the factor k from the direct test's lower threshold is not harmless: that shortcut rejects true claims in 880 of 2000 runs at $\delta = 0.1$ and 115 of 2000 at $\delta = 0.01$. The robust law's factor R is needed in the same sense. With r of w witnesses allowed to break their rules, validating a group without the factor accepts false claims in 627 of 1000 runs at $R = 4$; with it, in at most 18 (Appendix L.2).

K.2 Sampled outputs of language models

Program synthesis supplies exact checkers but not private witnesses. On 20 HumanEval tasks (Chen et al., 2021) with 200 samples each from gpt-4o-mini, every assert is a target and each witness requires every assert and one HumanEval+ test (Liu et al., 2023) on the same sample. At $\bar{\eta} = 1/4$, 14 of the 20 tasks have every contract outside the gap (95% interval 0.46 to 0.88) at temperatures 1.0 and 0.8. In 19 of the 20 tasks, however, the four witnesses agree on every sample, so discovery costs one query: the certifier uses 14.1 grades against 37.1 for direct testing at $\delta = 0.05$, without error in 14000 runs. These tasks may also be memorized (Riddell et al., 2024); the measurement concerns grades, not how they arise.

A population built from differences in knowledge realizes the architecture. Each of 16 systems is gpt-4o-mini with a knowledge base for one office: a random project codename, which is its witness, a coded fact for each of four binary attributes, which are its diagnostic pairs, and three common facts. One probe of 40 items is graded by exact contracts; the inclusions hold on every reply, and the richness condition of Proposition J.4 holds because each item is answered separately. At $\bar{\eta} = 1/4$, 14 of the 16 systems realize their favorable profile at temperature 1.0 and all 16 at 0.8, and the diagnostics are read without error. Table 2 replays the 14 systems. Scanning costs about as much as testing the 16 targets directly and grows linearly in t, while reading the four diagnostic answers first halves the cost at $t = 16$ and grows slowly with t. The diagnostics here are noiseless, so this population supports the additive law and the address reading of Theorem 4.1, not the logarithmic factor of Theorem 5.1.

The favorable population is a property of the model. Rerun unchanged with gpt-4.1-mini and gpt-4.1-nano, the same design yields no favorable profile among the 16 systems of either model. The

Candidate witnesses t	2	4	8	16
Certifier, scanning	21.8	31.6	43.8	67.8
Certifier, reading diagnostics	21.3	24.3	29.7	32.2
Direct testing (does not depend on t)	66.4	66.4	66.4	66.4

Table 2: Mean grades to certify $k = 16$ targets at $\delta = 0.05$ on the 16 knowledge configured gpt-4o-mini systems, averaged over the 14 with a favorable profile, 1000 replays each. The evaluator knows all but $\log_2 t$ address bits. Errors: at most 5 of 14000 runs in any cell.

knowledge part is reproduced exactly, but each model misanswers one or two basic items; gpt-4.1-mini, for instance, answers "Wednesday" to "What day of the week is 10 days after a Monday?" in 2459 of 3200 replies. Their claims are therefore false, and in an exploratory replay on the 13 systems whose contracts are all outside the gap, no procedure accepts a false claim in more than 0.18% of runs.

K.3 Where the assumptions stop

Fresh noise is an assumption, and computed answers violate it. When a diagnostic must be computed from the knowledge base, here by checking whether a four digit code is divisible by 3, gpt-4o-mini gives nearly the same answer on every sample: 59 of 64 own diagnostic pass rates are at most 0.1 or at least 0.9, and 31 of the 64 are misread. A system with a misread bit masters its witness but not a diagnostic of that witness's row, a state the rules forbid, so 14 of the 16 systems lie outside the class the laws cover. On them the diagnostics point to a witness the system does not master; the certifier stays valid but costs 129.8 grades, against 65.4 for direct testing and 59.5 for scanning, because only the raced direct test ends the run. Fresh noise is different: flipping each diagnostic grade of the knowledge configured systems with probability 0.2 raises the certifier's cost from 32.0 to 56.7 grades, still below scanning at 67.6.

Codes correct the errors that repetition cannot. We made diagnostic reads of gpt-4o-mini mostly fresh by randomizing them: each read selects one of two new random numbers by a flag of the knowledge base and adds a third, with the clause order random. Over 30720 reads the error is 0.185, with an invalid answer (6.6% of reads) read as a fair coin fixed in advance; it is 0.219 if invalid answers count as errors. The correlation between reads of the same flag is 0.06 rather than 0, because a few flags stay misread. We stored 24 reads of every flag of 16 systems for each $h \in \{4, 8, 12, 16\}$ and estimate success by resampling them. Storing the address as a systematic $[2h, h]$ code and decoding by maximum likelihood reaches success $3/4$ on average over the 16 addresses with 4 to 6 reads per address bit, and at every address with 8 to 12. Repeating the raw bits needs 9 to 11 reads per bit on average up to $h = 12$, does not reach $3/4$ on average at $h = 16$, and leaves 3 to 7 addresses below $3/4$ at every h within 11 reads per bit, because rereading a persistently misread bit does not repair it (Appendix L.6). At $h = 16$ and 96 reads, the code succeeds with probability 0.821 and repetition with 0.414, a difference of 0.41 with 95% interval 0.32 to 0.50 over resamples of the systems. Coding still beats repetition when invalid answers count as errors. The surplus of pairs in Theorem 5.1 pays on the reads of a language model.

Rules must come from contract design, not from difficulty labels. On the public grades of 13 open models on 128 MATH problems (Hendrycks et al., 2021), with 10000 samples per model and problem (Brown et al., 2024), we take the Level 1 problems as targets and each subject's Level 4 and 5 problems as a witness group, and supply the rule that every hard problem requires every easy one. The rule fails wherever it matters: all 12 pairs of a model and a mastered hard problem belong to a model that fails some Level 1 problem. A certifier that trusts the rule accepts the false claim of the strongest model in 1.4% of runs at $\delta = 0.05$, below δ but not zero. A failure budget of three per group removed these errors, although this model breaks more rules per group than that budget covers, so the theory does not guarantee it.

The cost unit is the grade. If one sample can be graded against every target at once, direct testing at $k = 16$ needs only 6.4 samples on the knowledge configured systems, but still 66.4 grades, as many as the sequential test. The query model therefore describes evaluations in which each grade is the expensive step or needs its own sample, as with a separate prompt per contract, slow test suites, or model or human judges.

L EXPERIMENTAL DETAILS

The supplement holds, for every experiment, a design file with its predictions, the SHA-256 hash of that file and of its scripts together with the time the hash was recorded, the scripts, every sampled reply and grade, and one script that recomputes every number below from the stored data. Design files are named T1 and E1 to E11; X1 to X3 are checks that were not in a design file and are labeled exploratory. The language models are OpenAI gpt-4o-mini (snapshot 2024-07-18), gpt-4.1-mini and gpt-4.1-nano (snapshots 2025-04-14), queried through the chat completions API with top_p 1; the open models enter through the public grades of Brown et al. (2024). Intervals for the pass rate of one node are Clopper–Pearson intervals at level 95%, exact when the node's grades are independent draws; the reads of E10 are correlated within a flag, and their intervals resample the systems instead.

Procedures and replays. All procedures are designed for the margin $\bar{\eta} = 1/4$, so each response moves the reference BSC log score by $\pm \log 3$. The direct test is Lemma D.2, with thresholds $\log(2/\delta)$ and $-\log(2k/\delta)$. The certifier is the construction of Theorem 4.1: a hitting policy proposes witnesses, each proposal costs one query and is checked by the one node test with boundaries $-\log 2$ and $\log(\pi^2 H_b l^2/(3\delta))$ in cycle l, where H_b is the number of candidates for scanning and the number of reads plus one for the diagnostic policy, and the certifier is raced query by query against the direct test at level $\delta/2$. The scanning policy proposes the candidate witnesses in random order; the diagnostic policy reads $d_{j,1}$ once, or r times with a majority vote, for each unknown address bit and proposes the witness with those bits. A replay runs these procedures on the empirical channel of a system, whose response at a node is one of that node's stored grades, drawn uniformly and independently of the past. Unless stated otherwise, each configuration is run 1000 times.

Batch grading (X1, exploratory). Every cost in this paper counts grades. Table 3 shows what changes when one sample is graded against all targets at once, with the same test per target and a target no longer graded once its test accepts. Batch grading saves samples, not grades: the batch direct test needs a few samples but as many target grades as the sequential test, which uses 66.4, 83.1, 37.1 and 46.4 grades in the four cells. One batch graded sample also reveals the passing witness of a knowledge configured system with probability equal to that witness's pass rate, 0.675 to 0.890, so the discovery problem disappears when grades are free.

	$\delta = 0.05$		$\delta = 0.01$	
Data	samples	grades	samples	grades
Knowledge configured systems, 14, $k = 16$	6.4	66.4	8.0	83.0
Program synthesis, 12 positive tasks	4.0	37.1	5.0	46.4

Table 3: Direct testing when every sample is graded against all targets at once (exploratory; 1000 runs per system or task). Mean samples and mean target grades per run.

L.1 THE LAWS ON THE MODEL'S CHANNEL (E8)

E8 simulates the paper's procedures on the symmetric channel with $\eta = 1/4$ at favorable profiles (a uniformly random witness per run), and replays them on the stored grades of the knowledge configured systems of Appendix L.4; no model is queried. Four references complete the comparison. LCB + Box tests whether every target mean exceeds $1/2$ with the Bernoulli KL confidence bounds of Kaufmann et al. (2018), whose threshold is $3\log(1 + \log N) + T(x)$ with their function T: it accepts when every lower bound exceeds $1/2$ ($x = \log(2/\delta)$), rejects when some upper bound falls below $1/2$ ($x = \log(2k/\delta)$), and samples the target with the smallest lower bound. LUCB_G search samples the witness with the largest upper bound, as in Kano et al. (2019), but with the same KL bounds in place of their Hoeffding bounds; it certifies the claim when some witness's lower bound exceeds $1/2$, with a union bound over the witnesses at level $\delta/2$, withdraws when every upper bound falls below $1/2$, and is raced against the direct test at $\delta/2$. The fixed design draws n samples per target and accepts when every target passes at least c times, with the smallest n and c for which $\mathbb{P}(\mathrm{Bin}(n, 3/4) < c) \le \delta/(2k)$ and $\mathbb{P}(\mathrm{Bin}(n, 1/4) \ge c) \le \delta/2$. The oracle tests the known witness alone, with thresholds $\pm \log(1/\delta)$. The invalid shortcut is the direct test with its lower threshold $-\log(2/\delta)$, without the factor k.

The design file registered twelve predictions after a quick run of the scripts on a few dozen replications, which is kept in the supplement and not reported. After that run only the speed of the two KL baselines changed, without changing their decisions, and the number of their replications was lowered as stated in the captions. Ten predictions hold. Two fail as stated. At $k = 64$ scanning, not reading diagnostics, is cheapest when $h \leq 2$, where there are at most four candidates. The identity table's budget per address bit is not strictly increasing: with a fixed number r of reads per bit it equals $r + 1/h$, which dips slightly between the values of h at which r stays the same. The design file writes $\mathcal{H}$ for the budgets of Table 5; they are budgets of the named policies, and $\mathcal{H}$ is at most the smallest of them. Tables 4 to 7 give the results behind Figure 2.

δ	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-6}	10^{-8}	slope in a
Symmetric channel, $\eta = 1/4$							
Direct testing	93.1	159.7	223.6	321.4	449.2	577.8	30.49
Certifier, diagnostics	86.9	93.0	102.9	109.6	128.5	145.9	3.72
Certifier, scanning	104.2	126.0	131.6	147.8	163.5	180.3	4.49
Oracle (known witness)	5.4	9.9	14.2	18.0	26.0	34.1	1.76
Fixed n, Bonferroni	272.0	528.0	768.0	1008.0	1504.0	2000.0	106.89
LCB + Box	1879.7	2290.5	2659.3	3008.6	3677.9	4336.7	151.02
LUCB_G search	256.2	376.0	508.4	634.6	885.7	1165.0	56.24
Stored grades of the 14 knowledge configured systems							
Direct testing	49.9	82.9	116.3	166.2	232.5	298.9	15.70
Certifier, diagnostics	29.3	35.8	42.9	49.0	64.1	76.7	2.97
Certifier, scanning	65.7	76.7	83.1	91.4	107.5	117.4	3.21
Oracle (known witness)	4.6	7.9	11.3	14.4	20.8	27.1	1.39
Fixed n, Bonferroni	272.0	528.0	768.0	1008.0	1504.0	2000.0	106.89
LCB + Box	406.1	494.8	569.7	647.7	787.7	937.8	32.56
LUCB_G search	131.0	195.7	259.2	322.4	440.0	583.1	27.70

Table 4: Mean grades at $k = t = 16$ against δ, with the least squares slope in $a = \log(1/\delta)$. Symmetric channel: 1000 runs per cell (200 for LCB + Box and LUCB_G search). Stored grades: 100 runs per system (20 for the two KL baselines).

h	2	4	6	8	10	12	14	16
Identity table: reads per bit	5	9	11	11	13	15	15	15
Identity table: budget$/h$	5.50	9.25	11.17	11.12	13.10	15.08	15.07	15.06
Code: reads per code bit	3	3	4	4	4	4	4	4
Code: budget$/h$	6.50	6.25	8.17	8.12	8.10	8.08	8.07	8.06
Code: minimum distance	2	3	4	4	4	5	5	5
Scanning: budget	3	12	48	192	768	3072	12288	49152

Table 5: Hitting budgets of three policies on the symmetric channel with $\eta = 1/4$: the grades, reads plus one proposal, with which the policy hits the mastered witness with probability $3/4$ at every favorable profile. The identity table reads each address bit the same number of times and takes the majority (exact). The systematic $[2h, h]$ code is the best of 50 random parity matrices by minimum distance, read r' times per code bit and decoded by maximum likelihood (4000 Monte Carlo trials). Scanning proposes $3t/4$ witnesses in random order. The hitting cost $\mathcal{H}$, the smallest budget over all policies, is at most the smallest entry of each column.

A common boundary (E11). In cycle l the certifier with diagnostics accepts at a lower boundary than scanning, since its $H_b = h + 1$ is smaller than scanning's $H_b = t$. A design file registered after the E8 results gives the certifier with diagnostics scanning's boundary. Table 8 shows that this adds 3.2 to 10.9 grades, 3% to 21% of the gap to scanning, and leaves the certifier with diagnostics cheaper than scanning in all nine settings. The gap comes from search: at $h = 4$ and $\delta = 0.05$ scanning spends 15.2 grades on search queries and 28.3 on failed checks, against 19.4 reads, 4.8 search queries and 7.8 on failed checks with diagnostics, while the accepting checks cost 11.3 and 13.5. In 2000 null runs at $\delta = 0.05$, the procedures the proof covers err in at most 37; scanning with the smaller boundary, which the proof does not cover, errs in 104 at $h = 8$, more than a fraction δ. The three predictions of this part hold.

Cell	Direct	Diag.	Scan	Fixed	Box	LUCB	No k
Error fraction at $\delta = 0.1$							
Symmetric channel, positive	0.0230	0.0075	0.0090	0.0545	0.0000	0.0220	0.4400
Symmetric channel, null	0.0360	0.0175	0.0290	0.0335	0.0000	0.0120	0.0205
Knowledge systems, positive	0.0005	0.0000	0.0005	0.0019	0.0000	0.0000	0.0152
Knowledge systems, derived null	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Other models, natural null	0.0052	0.0005	0.0005	0.0019	0.0000	0.0024	0.0043
Misread diagnostics, positive	0.0000	0.0000	0.0000	–	–	–	–
Error fraction at $\delta = 0.01$							
Symmetric channel, positive	0.0015	0.0005	0.0005	0.0030	0.0000	0.0040	0.0575
Symmetric channel, null	0.0050	0.0025	0.0015	0.0045	0.0000	0.0010	0.0050
Knowledge systems, positive	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0005
Knowledge systems, derived null	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Other models, natural null	0.0005	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Misread diagnostics, positive	0.0000	0.0000	0.0000	–	–	–	–

Table 6: Error fractions of every procedure (errors over runs). Symmetric channel: 2000 runs per cell, 1000 for LCB + Box and LUCB_G search. Stored grades: about 2100 runs per cell, 420 for the two KL baselines. The misread systems were replayed with direct testing and both certifiers only. The last column is the invalid shortcut that drops the factor k from the direct test's lower threshold.

Setting	reads	search	failed checks	accepting check	direct share	oracle
Symmetric, $h = 4$, $\delta = 0.05$	18.2	4.5	7.2	13.5	43.7	5.7
Symmetric, $h = 4$, $\delta = 10^{-6}$	19.7	4.9	8.1	33.6	66.3	26.6
Symmetric, $h = 8$, $\delta = 10^{-6}$	113.9	14.2	26.4	34.5	189.5	25.9
Knowledge systems, $\delta = 0.05$	5.3	1.3	0.5	9.1	16.3	4.7
Knowledge systems, $\delta = 10^{-6}$	5.1	1.3	0.5	25.1	31.9	20.9

Table 7: Where the certifier with diagnostics (read once per bit) spends its grades, per run, at $k = 16$: diagnostic reads, search queries, failed witness checks and the accepting check, together with the grades the raced direct test spends meanwhile, and the cost of the oracle that tests the known witness alone. Symmetric channel: 1000 runs; knowledge systems: 100 runs per system.

L.2 Redundant witnesses and failed rules (E7)

E7 simulates the upper bound of Theorem 6.1 with one group of $w \in \{4, 8\}$ witnesses, $k = 64$ targets and no diagnostics, so that $\mathcal{H} = 1$ and only the confidence term varies. The certifier screens the group's representative with boundaries $-\log 4$ and $\log 16$ and validates it with $N = \lceil 2R \log(1/\epsilon_l) \rceil$ blocks at uniformly drawn witnesses, each a one node test with boundaries $-\log(4N)$ and $\log 2$, with $\epsilon_l = 3\delta/(\pi^2 l^2)$ in cycle l; it is raced against the direct test at $\delta/2$. The comparison validates with $R = 1$ whatever r is. The worst null of the per group class deletes the last target and masters exactly r of the w witnesses, the representative among them. All three predictions hold (Table 9): with the factor R every null error fraction is at most δ; without it the error exceeds δ whenever $R \geq 2.6$, and, not predicted, already at $R = 2$; and the favorable cost grows about linearly in R, by a factor 1.89 from $R = 4$ to $R = 8$ at $\delta = 0.05$.

L.3 Program synthesis (T1)

Twenty HumanEval tasks (Chen et al., 2021) were drawn at random (seed 20260925) from the 153 tasks whose original test is a list of at least three asserts and that have at least four short HumanEval+ inputs (Liu et al., 2023). Each original assert is a target (k from 3 to 25, 172 in all; HumanEval/31 repeats one assert, so 171 are distinct), and each task has $t = 4$ witnesses, each requiring every assert and one HumanEval+ test on the same sample, and no diagnostics. The system message is "You are an expert Python programmer", the request asks for the complete function in one code block, at most 768 tokens are generated, and the program is the task prompt followed by the first code block of the reply; temperatures 1.0 and 0.8 give 8000 samples and no rule violation. At $\bar{\eta} = 1/4$, 14 of 20 tasks have every contract outside the gap at both temperatures (interval 0.46 to 0.88); 12 are positive and 2 null. The same six tasks lie inside the gap at both temperatures, with in gap pass rates from 0.275 to

		Diagnostics			
Setting	δ	own boundary	scanning's boundary	Scanning	S
Symmetric, $h = 4$	0.05	91.4	95.5	110.5	0.21
Symmetric, $h = 4$	10^{-3}	104.2	108.1	135.2	0.13
Symmetric, $h = 4$	10^{-6}	130.0	133.9	160.7	0.13
Symmetric, $h = 8$	0.05	186.7	192.9	244.9	0.11
Symmetric, $h = 8$	10^{-3}	290.2	299.5	464.0	0.05
Symmetric, $h = 8$	10^{-6}	370.9	381.8	740.1	0.03
Knowledge systems	0.05	32.2	35.4	68.9	0.09
Knowledge systems	10^{-3}	41.8	45.0	83.5	0.08
Knowledge systems	10^{-6}	63.7	66.9	103.9	0.08

Table 8: Mean grades at $k = 16$ of the certifier with diagnostics under its own boundary ($H_b = h+1$) and under scanning's ($H_b = t$), and of scanning, each raced against the direct test; S is the share of the gap between the first and the third column that the boundary explains. Symmetric channel: 1000 runs per cell; knowledge systems: 100 runs per system.

			null errors, $\delta = 0.05$		null errors, $\delta = 0.01$		favorable grades
w	r	R	with R	without	with R	without	$\delta = 0.05$
4	0	1.00	11	7	1	3	52.0
4	1	1.33	12	19	3	0	66.3
4	2	2.00	7	65	3	15	92.4
4	3	4.00	15	627	1	489	166.8
8	0	1.00	17	12	2	4	52.7
8	1	1.14	15	16	4	1	57.8
8	2	1.33	13	23	1	1	66.4
8	3	1.60	9	21	2	2	76.5
8	4	2.00	10	71	2	19	89.8
8	5	2.67	14	183	1	80	115.4
8	6	4.00	11	619	3	449	164.1
8	7	8.00	18	980	3	983	310.5

Table 9: The robust law on the symmetric channel ($\eta = 1/4$, $k = 64$, one group of w witnesses): errors out of 1000 runs at the worst null of the per group class (the last target deleted and r of the w witnesses mastered), for the validation with the factor $R = w/(w - r)$ and without it, and the mean grades of the certifier with the factor at the favorable profile.

0.74. On the 14 tasks outside the gap neither procedure errs in 14000 runs (upper bound 0.00026). On the positive tasks direct testing uses 37.1 grades and the certifier 14.1 at $\delta = 0.05$ (46.4 and 16.1 at 0.01); the direct cost grows from 12 at $k = 3$ to 100 at $k = 25$ while the certifier stays between 14.0 and 16.7, and per task standard deviations are at most 1.76. Two reporting checks: pass rates are homogeneous across the four requests per task (7 of 66 contract columns reject at 0.05, none after Bonferroni), and removing the duplicate assert changes no primary or secondary outcome.

T1 tests the gap promise and the upper bound only. In every positive task all four witnesses are mastered, and in 19 of 20 tasks they agree on every sample, so there is no single witness profile, discovery costs one query, and the richness condition is not exercised. HumanEval solutions occur in open pretraining corpora (Riddell et al., 2024) and the training data of the model are not public, so high pass rates may reflect memorization; the measurement concerns the grades a system returns, not how it produces them. Sampled code from one prompt varies across requests (Ouyang et al., 2025); here that variation is the fresh noise of the query model.

L.4 Knowledge configured systems (E1)

System $v \in \{0, \dots, 15\}$ is gpt-4o-mini with a system message that tells it to answer internal questions only from its knowledge base and otherwise to answer N/A, followed by a knowledge base for Office $v + 1$: a random project codename, one coded fact for each of four binary attributes whose values are

the bits of v, and three facts common to all offices. One probe of 40 items asks 16 basic questions (the targets, $k = 16$), the codes of both values of each attribute (the pairs $d_{j,0}, d_{j,1}$, $h = m = 4$), and the codename of every office. The pair member $d_{j,b}$ passes when the code of value b is given and the other value's code is not reproduced; the witness c_v passes when every target passes, the two codes of the row of v pass, and the codename of Office $v + 1$ is given. The inclusions of Appendix J.1 hold on every reply, and because each item is answered separately every grade vector obeying them is realizable. Each system is sampled 200 times at temperature 1.0 and 200 times at 0.8.

At $\bar{\eta} = 1/4$, 14 of 16 systems realize their favorable profile at temperature 1.0 (interval 0.62 to 0.98), 15 at $\bar{\eta} = 0.30$ and 16 at 0.35; at 0.8 all 16 do at every margin. The two exceptions have their own witness inside the gap (0.705 and 0.675). Own diagnostics pass at 1.000, the other value's at 0.000, and every other witness at 0.000; the only basic item that fails is the reversal of "kappa", wrong in 20.0% of replies at temperature 1.0 (exploratory). Three of four predictions hold. The fourth, that direct testing costs $4k$ grades within 10% at $\delta = 0.05$ and $5k$ at 0.01, fails at $k = 2$ and 4 (10.3 and 18.4 at $\delta = 0.05$), because the reversal item passes at about 0.8 rather than 1. With $t = 16$ and $\delta = 0.05$, direct testing uses 10.3, 18.4, 34.4 and 66.4 grades for $k = 2, 4, 8, 16$; the certifier with diagnostics uses 19.0, 28.4, 31.9 and 32.2, and with scanning 19.6, 33.9, 53.7 and 67.8. The largest error fraction in any positive cell is $39/14000$, for direct testing at $k = 2$, and no procedure errs on the derived nulls, which fail u_1 and every witness.

The same design, grader and analysis were rerun unchanged with gpt-4.1-mini and gpt-4.1-nano at temperature 1.0 (E5). Neither model realizes any favorable profile (0 of 16, interval 0 to 0.21), so the predicted count for gpt-4.1-mini, at least 12, fails and the replay predictions cannot be evaluated. Diagnostics and codenames reach the states of gpt-4o-mini; basic targets do not. gpt-4.1-mini answers "Wednesday" to item 4 in 2459 of 3200 replies, and gpt-4.1-nano misses both item 4 and the reversal item. In an exploratory replay on the 13 systems whose contracts are all outside the gap and whose claim is false (X2), no procedure accepts a false claim in more than 0.18% of runs at $\delta = 0.05$ or 0.034% at $\delta = 0.01$.

L.5 Computed and noisy diagnostics (E6)

The plan was a variant whose diagnostic answers need a small computation, so that reads fail with a fresh probability of 10 to 20%. A pilot on four offices of three designs, asking for an attribute value that follows from a code by divisibility by 3 or by the parity of a digit sum, showed nearly deterministic answers per code: 14, 14 and 15 of 16 own diagnostic pass rates were at most 0.1 or at least 0.9. The design was therefore split before any main sample was drawn. In E6a (divisibility by 3 of four digit codes, 16 new systems, 200 samples each), 31 of 64 own diagnostics are misread (pass rate at most 0.25), 26 of them on attributes of value 1, and only 2 systems realize their favorable profile. The witness contract of E6a does not include the diagnostics, so a misread system masters its witness but not a diagnostic of its row, which the rules forbid, so it lies outside the class the laws cover. On the 14 such systems the certifier with diagnostics uses 129.8 grades at $\delta = 0.05$, against 65.4 for direct testing and 59.5 for scanning, and three reads per bit do not help (129.7); on the 2 clean systems it uses 27.8, against 57.7 for scanning and 65.3 for direct testing. No run errs. E6b flips each diagnostic grade of the knowledge configured systems with probability e: at $k = t = 16$ and $\delta = 0.05$ the certifier reading once costs 32.0, 41.9 and 56.7 grades for $e = 0$, 0.1 and 0.2, reading three times costs 56.0, 60.2 and 73.2, and scanning costs about 67.6, with every error fraction at most $9/14000$. All ten predictions of E6 hold.

L.6 Repetition versus coding on reads of a language model (E10)

E10 asks whether the coded regime of Theorem 5.1 appears on diagnostic reads of a language model. For each $h \in \{4, 8, 12, 16\}$, 16 systems are gpt-4o-mini with a knowledge base of $2h$ flags: the codeword of a random address under a systematic $[2h, h]$ code, the best of 50 random parity matrices by minimum distance (3, 4, 5 and 5). A read asks for $y + c$, where y is one of two new random numbers selected by one flag and the clause order is random; the address bit or code bit is recovered from the answer. Every flag is read 24 times, 32 reads per request at temperature 1.0, for 30720 reads in all.

The read was chosen in a pilot of three rounds and nine variants, by a rule fixed before the first pilot request. Parity and residue reads were at chance, additive reads erred on one side, and the chosen

selection read had error 0.170 with a persistence correlation of 0.32. In the main bank the error is 0.249, 0.227, 0.215 and 0.223 when the flag is 0 and 0.170, 0.128, 0.135 and 0.161 when it is 1, for $h = 4, 8, 12, 16$. An invalid answer (6.6% of reads) is read as a fair coin fixed by its read id, as the design file registers. The correlation between reads of the same flag is 0.061, against a 95% range of -0.004 to 0.003 under independent reads. Majorities of 24 reads are wrong for 2, 2, 9 and 12 of the 128, 256, 384 and 512 flags. Table 10 resamples this bank: each of the 2000 draws per cell takes the next of the 16 systems and a random order of every flag's 24 reads, so a cell averages over the 16 addresses, and all draws reuse the same stored reads.

Total reads N at $h = 16$	32	64	96	128	160	192
Repetition, N/h reads per address bit	0.060	0.239	0.414	0.532	0.615	0.668
Coding, $N/(2h)$ reads per code bit	0.238	0.584	0.821	0.926	0.965	0.981
Budget for average success $3/4$	$h = 4$	$h = 8$	$h = 12$	$h = 16$		
Repetition	45	73	109	–		
Coding	25	33	73	97		

Table 10: Diagnostic reads of gpt-4o-mini, resampled from the stored bank and averaged over the 16 systems of each h: probability of recovering the exact address at equal total reads (top), and the budget, reads plus one proposal, at which this average reaches $3/4$ (bottom; repetition does not reach $3/4$ at $h = 16$ within 12 reads per bit).

Coding tracks the calculation for a symmetric channel with the measured error, falling short of it by at most 0.094. Repetition falls further behind as reads grow, to 0.668 against 0.862 at $h = 16$ and $N = 192$, because rereading does not repair a persistently misread bit, while the code corrects such bits. Two predictions hold: at $h = 16$ coding reaches $3/4$ with fewer reads than repetition, and it wins at all six equal read totals. The comparison of growth rates in h is not evaluable, since repetition's cost at $h = 16$ is undefined. The prediction that every cell agrees with the symmetric channel was conditioned on a correlation of at most 0.01; that premise fails, as the design file expected, and 13 of 48 cells agree.

Every address, intervals and invalid answers (E11). A second design file, registered after these results, reanalyzes the same reads. The laws ask for success $3/4$ at every favorable profile, while Table 10 averages over 16 addresses. Resampling each system's reads on its own (1000 draws per system), the code reaches $3/4$ at every one of the 16 addresses with 41, 65, 97 and 193 grades for $h = 4, 8, 12, 16$, that is 10, 8, 8 and 12 reads per address bit; repetition leaves 4, 3, 4 and 7 addresses below $3/4$ within 11 reads per bit. The independent units are the systems, so intervals resample them (2000 resamples): at $h = 16$ and $N = 96$ the code's advantage is 0.41 with 95% interval 0.32 to 0.50, and the code's average budget of 97 has interval 97 to 129. Counting every invalid answer as an error raises the error from 0.185 to 0.219; repetition then reaches $3/4$ on average at no h, and the code does with 25, 49, 73 and 129 grades. The four predictions of this part hold.

L.7 Rules taken from difficulty labels (E9)

E9 uses the public MATH grades of Brown et al. (2024): 13 models, the same 128 test problems of Hendrycks et al. (2021), and 10000 samples per model and problem at temperature 0.6, graded by a symbolic answer check. Levels and subjects come from an exact match to the MATH test set. The network was fixed from these metadata alone, before any grade was read: the 9 Level 1 problems are the targets, each subject's Level 4 and 5 problems form a witness group (7 groups of 3 to 16), and the supplied rule says that every hard problem requires every easy one. Table 11 gives the states at $\bar{\eta} = 1/4$. Every claim is false: two Level 1 problems, both stated through a diagram, are unmastered by all 13 models, with highest pass rates 0.030 and 0.083. All 12 pairs of a model and a mastered hard problem belong to a model with an unmastered target, so the rule fails wherever it would be used.

The replay covers the 9 models whose targets are all outside the gap. On Llama-3-70B-Instruct, which masters nine hard problems, direct testing never errs and uses 22.1 grades at $\delta = 0.05$; the certifier that trusts the rule accepts the false claim in 14 of 1000 runs (8 at $\delta = 0.01$), and with failure budgets of 1, 2 and 3 per group in 5, 3 and 0 runs. This model breaks more rules per group than any of these

	Level 1 targets			nodes	hard problems	mastered hard problems
Model	mastered	unmastered	in gap	in gap	mastered	by group
Llama-3-8B	3	2	4	8	0	0 0 0 0 0 0 0
Llama-3-8B-Instruct	4	2	3	12	2	1 1 0 0 0 0 0
Llama-3-70B-Instruct	7	2	0	18	9	4 0 1 0 0 3 1
Gemma-2B	1	2	6	8	0	0 0 0 0 0 0 0
Gemma-7B	6	2	1	4	1	1 0 0 0 0 0 0
Pythia-70M	0	9	0	0	0	0 0 0 0 0 0 0
Pythia-160M	0	9	0	0	0	0 0 0 0 0 0 0
Pythia-410M	0	9	0	0	0	0 0 0 0 0 0 0
Pythia-1B	0	9	0	0	0	0 0 0 0 0 0 0
Pythia-1.4B	0	9	0	0	0	0 0 0 0 0 0 0
Pythia-2.8B	0	9	0	0	0	0 0 0 0 0 0 0
Pythia-6.9B	0	9	0	0	0	0 0 0 0 0 0 0
Pythia-12B	0	9	0	0	0	0 0 0 0 0 0 0

Table 11: Public MATH grades of 13 models (10000 samples per model and problem) under the supplied rule that every Level 4 or 5 problem requires every Level 1 problem, at $\bar{\eta} = 1/4$. Every claim is false. Groups, in order: algebra, counting and probability, geometry, intermediate algebra, number theory, prealgebra, precalculus.

budgets covers, so the theory makes no promise for it; the prediction that the plain certifier would err in more than a fraction δ of runs fails at both levels. The eight Pythia models master nothing and every procedure is correct in all of their runs. Difficulty labels are therefore not prerequisite rules: the rules of the theory must come from contract design, as in Appendix L.4, or be trusted only with a failure budget.

L.8 Outside the promise (E2)

Stress tests on the program synthesis grades (temperature 1.0, 1000 runs per cell) check the stated limits of Theorem 6.2. On the six tasks inside the gap, the decision follows the weakest in gap target at $\delta = 0.05$: acceptance is at most 0.082 when that target passes at 0.275 or 0.325, between 0.37 and 0.68 at 0.46 to 0.51, and at least 0.937 at 0.63, at far higher cost; a prediction that five of these tasks would be accepted rested on a wrong premise in the design file and fails. When one target of each positive task is thinned to pass rate p, the pooled acceptance is at most 0.019 for $p \le 1/4$ and at least 0.996 for $p \ge 3/4$ at $\delta = 0.05$, and it increases in p. A procedure designed for margin 0.10 accepts at $p = 0.2$ more often than δ, while one designed for 0.40 stays valid at $p = 0.3$ and costs 2.1 to 2.5 times as much at $p = 1$. With a persistently corrupted target the error equals the corruption rate at both δ, which fresh noise would not allow. Drawing grades in the stored order instead of independently gives no error and changes mean costs by at most 0.9%.

M Computing the low cut coefficient

The structural characterization of a verification cost and the computational cost of finding its allocation are separate questions. This appendix locates the boundary for the low cut constraints of (9). All cuts refer to the given edge set; no transitive edges are added.

M.1 Separating low cut constraints

For a designated vertex s, nonnegative weights w, and integer budget r, define

$$L(G, s, r, w) = \min \left\{ \sum_{v \in D} w_v : s \in D,\ |\partial_G^+(D)| \le r \right\}.$$

When S contains several targets, the smallest left hand side of a fractional certificate constraint is $\min_{s \in S} L(G, s, r, w)$.

Proposition M.1 (Hardness of low cut separation)**.** *Deciding whether $L(G, s, r, w)$ is at most a supplied integer threshold is NP-complete, even for a three layer DAG and nonnegative integer weights. Here the robustness budget r is part of the input. Consequently, finding a violated low cut covering inequality for a supplied nonnegative rational allocation is NP-hard on this class.*

Proof. Membership in NP follows by presenting the set D. For hardness, consider an instance of CLIQUE: an undirected simple graph $\Gamma = (V_\Gamma, E_\Gamma)$, $n = |V_\Gamma|$, and an integer k. Instances with $k \le 1$, $k > n$, or maximum degree less than $k - 1$ can be decided immediately. It therefore suffices to consider $2 \le k \le n$ and $d = \max_v \deg_\Gamma(v) \ge k - 1$.

Construct a DAG with a source s, a second layer node for every vertex $v \in V_\Gamma$, and a third layer node for every edge $e \in E_\Gamma$. Include $s \to v$ for every second layer node, and $v \to e$ whenever v is an endpoint of e in Γ. Set

$$r = n - k, \qquad w_s = 0, \qquad w_v = d - \deg_\Gamma(v), \qquad w_e = 1.$$

The graph and weights have polynomial encoding size.

For a set D containing s, let $T = D \cap V_\Gamma$ and $Q = D \cap E_\Gamma$, identifying third layer nodes with their original edges. Its outgoing cut has size

$$|\partial_G^+(D)| = n - |T| + b(T, Q), \qquad b(T, Q) = |\{(v, e) : v \in T,\ e \ni v,\ e \notin Q\}|.$$

Feasibility implies

$$|T| \ge k, \qquad b(T, Q) \le |T| - k.$$

We show that any feasible set can be reduced to one with exactly k second layer nodes without increasing its weight.

Suppose $|T| > k$. If $b(T, Q) > 0$, choose a vertex $v \in T$ incident to at least one edge outside Q, and let $b_v \ge 1$ count those incidences. Removing v from D changes its outgoing cut by $1 - b_v \le 0$, and removes nonnegative vertex weight. If $b(T, Q) = 0$, the cut is $n - |T| \le n - k - 1$. Removing any vertex of T increases the cut by exactly one and therefore preserves feasibility, again without increasing weight. Repeating gives $|T| = k$; feasibility then forces $b(T, Q) = 0$. Thus Q must contain every edge incident to T. Delete any additional edge nodes, which does not affect the cut.

Writing $E_\Gamma(T)$ for the edges with both endpoints in T, the resulting weight is

$$\begin{aligned} w(D) &= \sum_{v \in T} (d - \deg_\Gamma(v)) + |\{e \in E_\Gamma : e \cap T \ne \varnothing\}| \\ &= kd - |E_\Gamma(T)|. \end{aligned}$$

Conversely, every k-element set $T \subseteq V_\Gamma$, together with s and its incident edge nodes, produces such a feasible D. Consequently,

$$L(G, s, r, w) = kd - \max_{|T| = k} |E_\Gamma(T)|.$$

The original graph has a k-clique if and only if this value is at most $L_0 = kd - \binom{k}{2}$.

For the covering inequality statement, give the same DAG the rational allocation $\lambda_v = w_v / (L_0 + 1/2)$. The denominator is positive under the conditions above. Since every unscaled weight is integral, there exists an eligible D with $\lambda(D) < 1$ exactly when the original graph has a k-clique. □

This result does not assert that the manuscript's unweighted scalar $C_r(G, S)$ is NP-hard to compute. In the reduction just given, $S = \{s\}$ and $C_r(G, \{s\}) = 1$, because directly sampling s is feasible and every fractional certificate has total mass at least one. Hard separation for arbitrary trial allocations does not imply hard optimization of one particular objective.

Proposition M.2 (Separation at a fixed robustness budget)**.** *For nonnegative weights,*

$$L(G, s, r, w) = \min_{\substack{F \subseteq \mathcal{E} \\ |F| \le r}} w\big(\mathrm{Reach}_{G - F}(s)\big),$$

where the reachable set includes s. Consequently, separation for the low cut covering program can be performed in

$$O\left(|S|(|V|+|\mathcal{E}|)\sum_{j=0}^{\min\{r,|\mathcal{E}|\}}\binom{|\mathcal{E}|}{j}\right)$$

graph operations, in addition to rational arithmetic on the input weights. In particular, it is polynomial for every fixed integer r.

Proof. For any F, put $R = \mathrm{Reach}_{G-F}(s)$. Every original edge leaving R must belong to F; otherwise its head would be reachable. Hence $|\partial_G^+(R)| \le |F|$, and the right hand side minimizes over feasible sets.

Conversely, let D attain the defining minimum and set $F = \partial_G^+(D)$. No path from s in $G - F$ can leave D, so $R = \mathrm{Reach}_{G-F}(s) \subseteq D$. Nonnegative weights imply $w(R) \le w(D)$. This proves the equality. Enumerate the indicated edge sets and compute each reachable set by a graph traversal, then repeat over $s \in S$. □

The algorithm has an exponent depending on r, so the result is an XP bound in the robustness budget; it is not a fixed parameter tractability claim. Nor does it contradict Proposition M.1, where r grows with the input. Together with the standard separation based solution (Grötschel et al., 1981) of rational linear programs, it makes the fractional allocation computable in polynomial time for each fixed r, including arbitrary graph depth.

An explicit program for fixed r. The proof of Proposition M.2 shows more: for every $D \in \mathcal{D}_r(G,S)$ containing a target s, the set $R = \mathrm{Reach}_{G-F}(s)$ with $F = \partial_G^+(D)$ satisfies $R \subseteq D$ and $R \in \mathcal{D}_r(G,S)$. For nonnegative allocations, $\lambda(D) \ge \lambda(R)$, so the constraints $\lambda(\mathrm{Reach}_{G-F}(s)) \ge 1$, $s \in S$, $|F| \le r$, generate the whole feasible region of (9). For fixed r this is an explicit linear program with at most $|S| \sum_{j \le r} \binom{|\mathcal{E}|}{j}$ constraints, and no separation oracle is needed.