CAN MORE SAMPLES FIX THE VOTE? ANYTIME-VALID AUDITS OF LLM SELF-CONSISTENCY

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-consistency selects the most frequent answer among repeated language model responses without a verifier. More samples stabilize voting but leave the response distribution unchanged, so the winner need not be acceptable. We propose Response Evaluation with Sequential Outcome-Level Verification via E-processes (**RESOLVE**), which audits whether population plurality favors PASS or FAIL under a fixed evaluator, leaving generation and selection unchanged. For a fixed policy, deterministic evaluator and i.i.d. responses, false reports remain controlled during continuous inspection. We bound response cost for an uncapped idealization and show when identifying a label can require less evidence than identifying an exact answer. Experiments distinguish answer availability, evaluator acceptance and plurality selection. An unfavorable diagnosis can motivate policy revision, which requires a fresh audit.

1 INTRODUCTION

Self-consistency (Wang et al., 2023) returns the most frequent answer among repeated language model responses without a verifier. Prior work asks when to stop without unnecessary calls (Aggarwal et al., 2023; Li et al., 2024). Yet more samples leave a fixed policy's response distribution unchanged. They can help some queries while hurting others (Chen et al., 2024). If the most likely answer is unacceptable, further voting can stabilize it rather than repair it.

A fixed evaluator labels each answer category PASS or FAIL. PASS may mean matching an answer key, naming a reference entity with retained supporting evidence, or giving a correct verdict with a complete evidence set. FAIL means that this criterion is not fully met. With the policy, categories and evaluator fixed, we ask

Does continued voting favor a population winner that passes the evaluator?

Finding a PASS establishes availability, not which category dominates voting. We compare the strongest PASS and FAIL category probabilities. Their strict ordering determines the population plurality label. Exact identity further asks which single category is most probable. CITE (Ota et al., 2026) certifies a supplied answer as this unique population mode.

Repeated fixed-sample tests can accumulate false reports. Anytime-valid inference controls error even under continuous inspection (Ramdas and Wang, 2025). RESOLVE discovers categories online and tests whether a candidate beats every opposite-label category. It reports favorable for PASS or unfavorable for FAIL after a confirmed threshold crossing, and otherwise remains unresolved at the budget.

Diagnosis does not prescribe a repair. An unfavorable report can motivate changing a prompt or retrieval, but the new response distribution requires a fresh audit. Using the evaluator to select responses also changes the policy (Cobbe et al., 2021). Here it serves only the audit.

Figure 1 illustrates why the label can be easier to identify than the exact answer. Two leading FAIL categories may be nearly tied while both are more probable than any PASS category. Identifying the mode requires resolving their close contest. An unfavorable label audit does not, because either winner fails the evaluator. This does not pool categories with the same label.

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

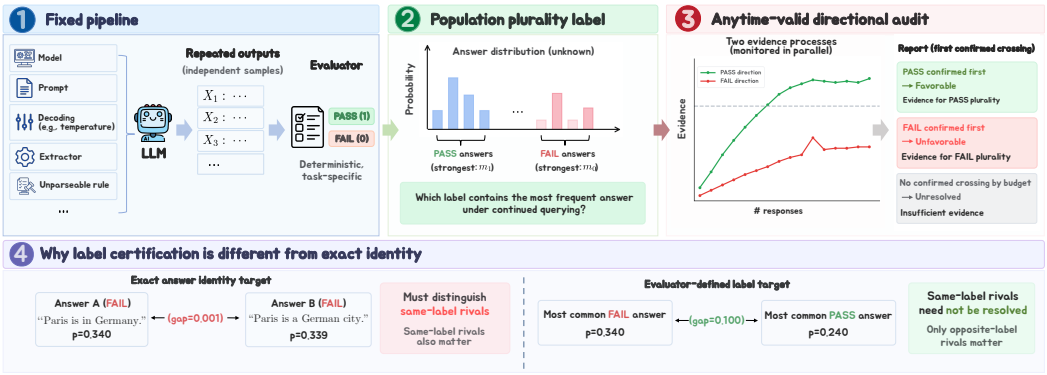


Figure 1: **RESOLVE overview**. RESOLVE sequentially audits whether the population plurality favors PASS or FAIL under a fixed policy and evaluator. It reports after a confirmed threshold crossing and otherwise remains unresolved. Unlike exact identity certification, it need not resolve same-label rivals.

With all categories and labels known in advance, confidence sequences (CS) can test the same label question under continuous monitoring (Howard et al., 2021; Lindon and Malek, 2022). RESOLVE instead discovers and evaluates categories as they appear. Appendix D discusses broader connections to voting, mode certification, verifier selection and sequential inference.

The main contributions of this work are summarized as follows.

- We define the population plurality label, distinct from PASS availability and the current vote.
- We construct an anytime-valid audit with online discovery on a countable answer space.
- We bound ideal response cost and explain when label certification avoids harder identity comparisons.
- We test these distinctions experimentally and re-audit a revised policy using fresh responses.

2 Population Label of Repeated Plurality

We define the frozen pipeline, its population plurality label and the corresponding audit objective.

2.1 Fixed Policy and Evaluator

Fix one query, model, prompt, decoder, token budget, unparseable-output rule and extractor mapping each response separately to a category. The extracted categories $X_1, X_2, \ldots$ are assumed i.i.d. from an unknown distribution P on countable $\mathcal{A}$, with masses $p_a = \mathbb{P}_P(X_t = a)$, omitting distribution subscripts when clear. A deterministic evaluator $h : \mathcal{A} \to \{0, 1\}$ assigns PASS (1) or FAIL (0), with $\mathcal{A}_c = \{a \in \mathcal{A} : h(a) = c\}$. Categories retain everything needed for evaluation, such as evidence IDs, so responses in one category share a label. The category map, evaluator and plurality tie rule are fixed independently of monitored responses. The target depends on this frozen generation, extraction and evaluation pipeline, not the model alone. Changing the map changes the vote.

2.2 Population Target of Repeated Plurality

Plurality compares individual category probabilities, not total label mass $\pi_c := \mathbb{P}(h(X_1) = c)$. Several PASS categories can jointly outweigh FAIL while one FAIL category is largest. Let m_1 and m_0 be the largest PASS and FAIL category probabilities, zero for an empty class. For $c \in \{0, 1\}$, these maxima and their gap are

$$m_c := \sup_{a \in \mathcal{A}_c} p_a, \qquad \Gamma := m_1 - m_0.$$

Let $\widehat{a}_n$ be the plurality of the first n responses. The following proposition shows that a strict gap fixes the eventual label, allowing ties among top categories with that same label.

Proposition 2.1. *For fixed P, h, and policy, if $\Gamma \neq 0$, then*

$$h(\widehat{a}_n) \to \mathbb{1}\{\Gamma > 0\} \quad \textit{almost surely as } n \to \infty.$$

Thus $\Gamma > 0$ is favorable to PASS and $\Gamma < 0$ is unfavorable. If $\pi_1 > 0$, the probability of finding a pass, $1 - (1 - \pi_1)^n$, approaches one as $n \to \infty$. The vote nevertheless fails with probability approaching one when $\Gamma < 0$. Votes at finite budgets may fluctuate, and the proposition makes no claim at $\Gamma = 0$. Section 5.1 examines availability versus selection. Appendix C.1 gives the proof.

2.3 Directional Auditing Objective

For $c \in \{0, 1\}$, $1 - c$ denotes the opposite label. Direction c asks whether its strongest category beats every opposite-label category,

$$\begin{aligned} H_{0,1}\colon m_1 \leq m_0 \quad &\text{versus} \quad H_{1,1}\colon m_1 > m_0, \\ H_{0,0}\colon m_0 \leq m_1 \quad &\text{versus} \quad H_{1,0}\colon m_0 > m_1. \end{aligned}$$

At $\Gamma = 0$, both nulls hold, so either report is an error. Unresolved means insufficient evidence, not a population tie. For prespecified $\alpha_c \in (0, 1)$, we require the probability of ever reporting direction c under $H_{0,c}$ to be at most α_c. For the fixed audit rule, this bound must hold for every null distribution and allow stopping based on observed evidence. Allocating $\alpha_1 + \alpha_0 \leq \alpha_{\text{query}}$ protects one standalone query without requiring independent directions.

Auditing continued voting thus means determining the sign of Γ. The maxima are fixed but their maximizing categories are unknown, so evidence must remain valid as candidates are discovered.

3 RESOLVE

Neither the strongest candidate nor a suitable bet size is known in advance. At a category's first appearance, RESOLVE opens an account, a candidate-specific evidence tracker. Each account averages over fixed bets and uses only later responses, avoiding reuse of discovery data.

Evidence against the strongest single opponent. Fix direction c and category $z \in \mathcal{A}_c$, discovered at $T(z) = \inf\{t \geq 1 : X_t = z\}$, with $\inf \varnothing = \infty$. Discovery identifies z but contributes no evidence to its account. Write $N_t(a) = \sum_{k=1}^{t} \mathbb{1}\{X_k = a\}$ and $N_0(a) = 0$ for global counts. For $t \geq T(z)$, $C_t(z)$ counts later occurrences of z, while $M_t(z)$ is the largest individual opposite-label count after discovery,

$$\begin{aligned} C_t(z) &:= N_t(z) - N_{T(z)}(z), \\ M_t(z) &:= \max_{b \in \mathcal{A}_{1-c}} \big[N_t(b) - N_{T(z)}(b)\big]. \end{aligned} \tag{1}$$

An empty opposite class has $M_t(z) = 0$, and unseen categories have zero counts. For example, on the sequence A(PASS), B(FAIL), A(PASS), C(FAIL), $C_4(A) = 1$ and $M_4(A) = 1$, not 2, since B and C are separate competitors. Pooling their FAIL counts would test a different target.

For bet $0 < \lambda < 1$, we use CITE's pairwise betting form (Ota et al., 2026) but compare only opposite labels. The largest opposing count gives the weakest evidence because $1 - \lambda < 1$,

$$G_t(z; \lambda) := (1 + \lambda)^{C_t(z)} (1 - \lambda)^{M_t(z)}.$$

Accounts stay at one through discovery. Thereafter, each increase in $C_t(z)$ multiplies evidence by $1 + \lambda$, and each increase in $M_t(z)$ by $1 - \lambda$. Other responses are neutral. Larger bets amplify gains and losses, while smaller bets can grow at narrower gaps.

Discovery mixture and reporting. Let $R_{c,j} \in \mathcal{A}_c$ be direction c's jth discovery at $T_{c,j} = T(R_{c,j})$. Keep the first K_{cap} accounts per direction. Bars mark the capped monitor and its stored weights. Section 4.2 studies the uncapped idealization. Fix finite $\Lambda \subset (0, 1)$ and nonnegative

weights v_λ over bets and $\bar{w}_j$ over discovery indices, with $\sum_{\lambda\in\Lambda} v_\lambda = 1$ and $\sum_{j=1}^{K_{\mathrm{cap}}} \bar{w}_j \le 1$. Let $K_c(t) \le K_{\mathrm{cap}}$ count direction-c accounts opened by time t. Unused weight stays at evidence value one. We average opened accounts over candidates and bets,

$$\bar{E}_t^c = 1 - \sum_{j=1}^{K_c(t)} \bar{w}_j + \sum_{j=1}^{K_c(t)} \bar{w}_j \sum_{\lambda\in\Lambda} v_\lambda G_t(R_{c,j};\lambda). \tag{2}$$

Thus $\bar{E}_0^c = 1$, and merely discovering a category creates no evidence. Define the first mathematical threshold crossing by

$$\bar{\tau}_c(\alpha_c) := \inf\{t \ge 0 : \bar{E}_t^c \ge 1/\alpha_c\}.$$

The implementation's first exactly confirmed report time is ρ_c, infinite if no report occurs on the full stream. Floating-point screening may delay exact arithmetic confirmation, so $\rho_c \ge \bar{\tau}_c(\alpha_c)$. Both times use only past and current responses. Validity controls crossing at $\bar{\tau}_c$, while empirical cost measures confirmation at ρ_c. Every erroneous report is contained in

$$\mathcal{E}_{\mathrm{err}} := \{\Gamma \le 0,\ \rho_1 < \infty\} \cup \{\Gamma \ge 0,\ \rho_0 < \infty\}. \tag{3}$$

An audit is resolved by budget B if either direction has a confirmed report, otherwise unresolved. It returns the first confirmed report by B, with PASS priority if simultaneous. The budget limits reporting, while the cap limits candidates, not opponents or responses. Appendix A.3 gives the algorithm. This construction leaves open whether repeated inspection controls false reports and how quickly a report becomes possible.

4 Statistical Guarantees

We first establish validity for the capped monitor. An uncapped idealization then isolates response cost and the label versus identity comparison.

4.1 Anytime Validity under Continuous Inspection

Continuous inspection needs more than a fixed-time error bound. The following theorem controls false threshold crossings at any time (Ramdas and Wang, 2025).

Theorem 4.1. *For each $c \in \{0,1\}$ and prespecified $\alpha_c \in (0,1)$, under $H_{0,c}$,*

$$\mathbb{P}\big(\bar{\tau}_c(\alpha_c) < \infty\big) \le \alpha_c.$$

The bound holds uniformly over all distributions satisfying the fixed i.i.d. setup of Section 2, with the RESOLVE monitoring rule fixed.

The computed evidence need not be a supermartingale and is not a posterior probability. Under each null distribution, a fixed opposite-label category is at least as probable as every candidate. Betting against it after discovery gives a nonnegative supermartingale. Using the largest observed opposing count can only lower computed evidence. Fixed nonnegative mixtures preserve the dominating supermartingale (Ramdas et al., 2020), so Ville's inequality controls crossing at any inspection. The corollary protects both directions by a union bound.

Corollary 4.2. *Under Theorem 4.1, if $\alpha_0 + \alpha_1 \le \alpha_{\mathrm{query}} < 1$, then the event in (3) satisfies*

$$\mathbb{P}(\mathcal{E}_{\mathrm{err}}) \le \alpha_0 + \alpha_1 \le \alpha_{\mathrm{query}}.$$

Exact confirmation inherits this single-query guarantee, not protection for a collection. Appendix C.2 proves both results. Validity does not determine reporting speed.

4.2 Response Cost of an Ideal Uncapped RESOLVE Process

For response cost, remove the account cap, retain the construction and drop bars. For finite or countable $\Lambda \subset (0,1)$ and fixed nonnegative weights with $\sum_{j\ge1} w_j = \sum_{\lambda\in\Lambda} v_\lambda = 1$, ideal evidence and crossing at level $\alpha \in (0,1)$ are

$$E_t^c = \sum_{j\ge1} w_j \sum_{\lambda\in\Lambda} v_\lambda G_t(R_{c,j};\lambda),$$
$$\tau_c(\alpha) = \inf\{t \ge 0 : E_t^c \ge 1/\alpha\}. \tag{4}$$

Undiscovered accounts contribute one. The implemented and ideal evidence processes need not be ordered pathwise because their weights differ. For a strongest category $a^\star \in \mathcal{A}_c$ in the winning label, its mass is $p = p_{a^\star} = m_c$ and its strongest opposite-label rival has mass $q = m_{1-c} < p$. The terms below average log gains on candidate responses and log losses on strongest-rival responses. Others contribute zero. Their sum gives asymptotic log growth per response, maximized over bets below,

$$\ell_\lambda(p,q) = p\log(1+\lambda) + q\log(1-\lambda),$$
$$\mathcal{I}(p,q) = \sup_{\lambda\in(0,1)} \ell_\lambda(p,q) = p\log\frac{2p}{p+q} + q\log\frac{2q}{p+q}.$$

Logarithms are natural and $0\log 0 = 0$. Write $\mathcal{I}_\Lambda(p,q) = \sup_{\lambda\in\Lambda}\ell_\lambda(p,q)$ for the best asymptotic log-evidence growth per response available on the grid. The following theorem turns positive growth into an expected response bound for the ideal process.

Theorem 4.3. *Fix P, h, the uncapped weights $w_j = 6/(\pi^2 j^2)$, a finite or countable grid $\Lambda \subset (0,1)$, and positive normalized grid weights. If $p = m_c > q = m_{1-c}$ and $\mathcal{I}_\Lambda(p,q) > 0$, then, as $\alpha \downarrow 0$,*

$$\limsup_{\alpha\downarrow 0} \frac{\mathbb{E}[\tau_c(\alpha)]}{\log(1/\alpha)} \le \frac{1}{\mathcal{I}_\Lambda(p,q)}.$$

The leading bound divides the log threshold $\log(1/\alpha)$ by growth per response $\mathcal{I}_\Lambda(p,q)$. Stricter error control raises the threshold, and a closer opposite-label rival slows growth. A dense grid gives $\mathcal{I}_\Lambda = \mathcal{I}$. A finite grid may miss positive growth, a cap may omit every maximizer, and exact confirmation has no delay bound. These limits concern efficiency, not validity. Appendix C.3 gives the proofs.

4.3 Label versus Identity

Label and identity differ in which rival must be separated. Recall the strongest opposite-label mass q. For a unique mode $a^\star$ with another category, let $s = \sup_{a\in\mathcal{A}\setminus\{a^\star\}} p_a$ be the strongest other mass of either label. Label certification compares p with q, whereas exact identity compares p with s. A closer rival shares the winner's label exactly when $q < s$, as in Figure 1.

For comparison, CITE (Ota et al., 2026) receives $a^\star$ before sampling to certify it as the unique population mode. Let $\tau_{\mathrm{CITE}}(\alpha)$ be its first mathematical certification time, infinite for no crossing. A level-α certifier reports with probability at most α when the target is not the unique mode. Its rule and the distribution of response-independent auxiliary input stay fixed across distributions. The next corollary combines CITE's Corollary C.11 lower bound $\log(1/\alpha)/\mathcal{I}(p,s)$ with our upper bound to isolate same-label competition.

Corollary 4.4. *Under Theorem 4.3, use a countable dense grid and the specified fixed-target level-α CITE certifier. If $p = p_{a^\star} = m_c > s > q = m_{1-c}$, then*

$$\limsup_{\alpha\downarrow 0} \frac{\mathbb{E}[\tau_c(\alpha)]}{\log(1/\alpha)} \le \frac{1}{\mathcal{I}(p,q)} < \frac{1}{\mathcal{I}(p,s)} \le \liminf_{\alpha\downarrow 0} \frac{\mathbb{E}[\tau_{\mathrm{CITE}}(\alpha)]}{\log(1/\alpha)},$$

with all configuration choices other than α held fixed.

The leading response bounds separate as $\alpha \downarrow 0$, not necessarily costs at every finite budget. Same-label rivals can affect discovery and weights, but not the leading opposite-label rate. With only one PASS category, the strict regime $s > q$ can occur for a FAIL winner but not a PASS winner. Appendix C.4 gives the proof.

We next examine these target and evidence distinctions at finite budgets.

5 Experiments

The experiments examine PASS availability versus plurality selection, continuous-monitoring validity and difficulty, and the cost of same-label competition. We then compare procedures for the same label target and illustrate fresh auditing after policy revision.

Shared protocol. The primary prospective study has 140 queries. Each reasoning task has 20. AIME 2026 asks for an integer in mathematics, GPQA-Diamond a science option, and MMLU-Pro an option across subjects. The extracted integer or letter is the category. Matching the answer key is PASS (MathArena, 2026; Rein et al., 2024; Wang et al., 2024b).

Each evidence task has 40 queries and allows several PASS categories. QAMPARI-based QA asks for one entity from ten retrieved passages, such as a museum matching a location. PASS requires a reference entity with retained annotated support in those passages (Amouyal et al., 2023; Gao et al., 2023). FEVEROUS-based verification asks for a verdict and evidence-ID set, which together define the category (Aly et al., 2021). PASS requires the reference verdict and a complete annotated evidence set. A correct verdict about a mountain's height and location can fail with location evidence alone. Neither task equates FAIL with factual error.

Rules and queries are frozen before generation. Audits use $B = 160$ responses, $K_{\text{cap}} = 50$ accounts per direction, five bets and $\alpha_0 = \alpha_1 = 0.025$. The primary policy is Qwen2.5-7B-Instruct at temperature 0.700, top-p 0.900 and top-k 20. Evidence tasks also use Qwen3-30B-A3B-Instruct-2507 and gpt-oss-20b. We lock diagnoses before collecting 64 fresh responses per query under a frozen follow-up protocol. Appendix A.2 gives details. The separate 398-query archived comparisons are retrospective.

Availability and selection measure different events in a uniformly ordered prefix of n responses from a retained block. Their probabilities are $A(n)$ for finding a PASS and $V(n)$ for plurality selecting one, averaged equally over queries. A subscript i denotes one query. Report rate is the fraction returning either direction by B, not accuracy on real data. Restricted mean stopping time averages response use with nonreturns charged the budget. For M runs or orderings with first confirmed report times ρ_j in either direction, it is

$$\widehat{\text{RMST}}_B = \frac{1}{M}\sum_{j=1}^{M} \min\{\rho_j, B\}.$$

Nonreturns have $\rho_j = \infty$ and pay B. Reported uncertainties use $\pm$ one Monte Carlo standard error (SE) over simulated streams or one ordering SE within fixed blocks, as specified in Appendix A.5.

5.1 PASS Availability versus Plurality Selection

Proposition 2.1 distinguishes PASS availability from plurality selection. Figure 2(a) examines whether availability rises while PASS plurality stays low. Of the 60 reasoning queries, 40 receive unfavorable reports. Their fresh blocks contain passes on 26 queries, but no full-block plurality passes. Availability is exact within each block. Plurality uses 2000 orders per query with one ordering SE. These finite blocks do not identify population signs or prove convergence. One unfavorable AIME 2026 report is driven by extraction failure, part of the frozen policy.

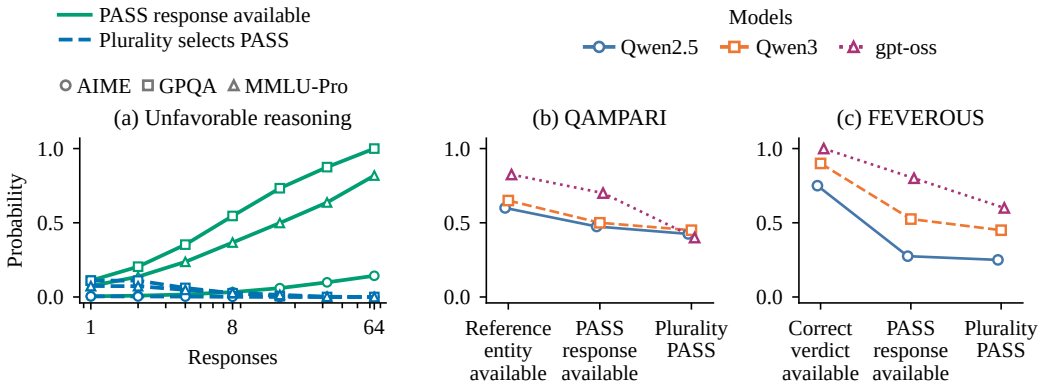


Figure 2: Availability, evaluator acceptance and plurality selection. (a) Unfavorable reasoning queries as responses accumulate. (b) and (c) show all 40 entity QA and verification queries at 64 responses under three fixed policies. Lines connect stages within each policy.

Figures 2(b) and 2(c) separate loose task success, full evaluator acceptance and plurality selection. Loose coverage counts a reference entity or correct verdict. The drop to $A(64)$ counts blocks with that loose success but no response meeting the full evidence criterion. The drop from $A(64)$ to $V(64)$ counts blocks where a PASS appears but does not win. Both losses occur for gpt-oss on FEVEROUS-based verification, although every block contains the correct verdict. The first concerns evaluation. Only the second concerns plurality under Proposition 2.1. This distinction matters when evidence IDs, not just the verdict, define a category.

Finding a PASS is therefore not enough to assess continued voting. The primary audit reports on 132 of 140 queries, using 39.1 responses on average with nonreturns charged 160. This excludes follow-up costs and implies neither token nor latency savings. These results motivate the label target, but neither availability nor report rate establishes validity under repeated monitoring.

5.2 Continuous-Monitoring Validity and Audit Difficulty

False-report control and reporting difficulty are distinct. Figure 3(a) shows how repeated inspection accumulates fixed-sample error. At the binary tie $P = (1/2, 1/2)$, every report is false. One prespecified direction uses level 0.050. Exact binomial calculations give once-only and repeated-test errors, compared with 30000 RESOLVE streams. By 1024 responses, repeated testing reaches 34.0% false reports versus $(1.82 \pm 0.08)\%$ for RESOLVE, within Theorem 4.1's anytime budget. This bounds unconditional error, not error conditional on returning. The 5% level is an upper bound, not a target rate.

Low false-report rates do not ensure quick answers. Near the population boundary, p and q are close and the growth rate in Section 4.2 becomes small. Many audits then remain unresolved at moderate budgets (Appendices B.2.1 and B.2.2).

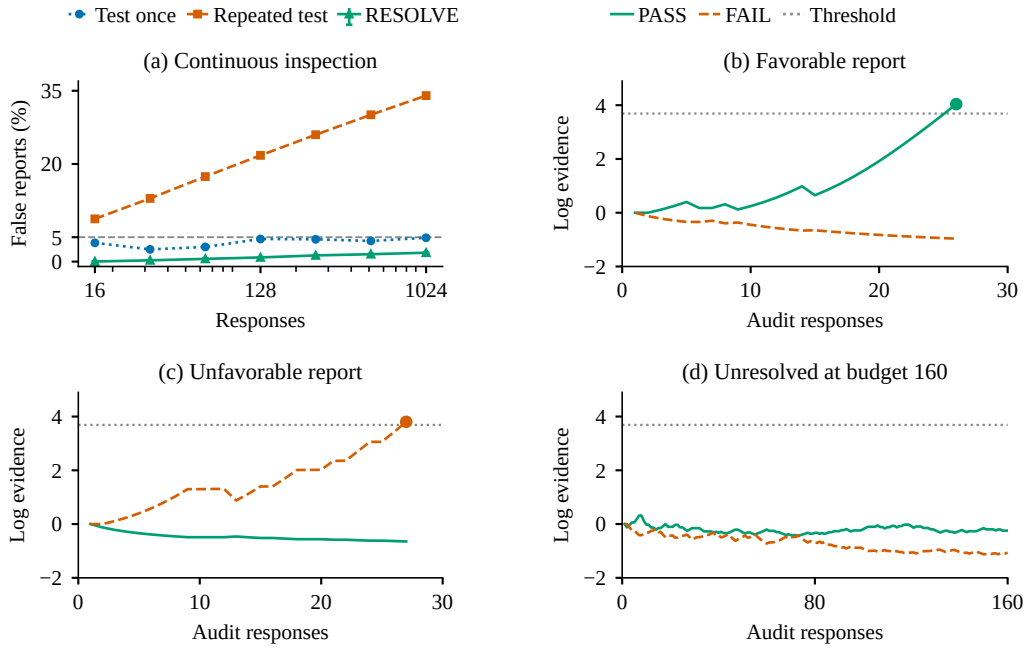


Figure 3: False reports under continuous inspection at a binary tie in (a), with one Monte Carlo SE for RESOLVE and a 5% reference line. The GPQA-Diamond paths in (b), (c) and (d) are favorable, unfavorable and unresolved, respectively. Dots mark confirmed reports using threshold $\log(40)$, and horizontal ranges differ.

Figures 3(b), 3(c) and 3(d) distinguish threshold crossings from an evidence lead on real queries. Two cross the fixed threshold. One remains unresolved despite a lead, which alone does not justify a report. Examples are nearest their outcome group's median stopping time or, if unresolved, median final maximum log evidence. Ties use query index, and selection uses no follow-up outcomes. Appendix B.2 shows all 140 paths.

We next examine reporting difficulty in controlled distributions and independent follow-up responses. At fixed p, increasing q lowers $\mathcal{I}_\Lambda(p, q)$ and raises the ideal cost bound in Theorem 4.3. Mean cost in Appendix Figure 8(b) rises with q and nearly linearly with $\log(1/\alpha)$. We fix $p = 0.350$, vary $q \in \{0.100, 0.200, 0.300\}$ and retain 20 smaller FAIL categories. Each distribution supplies 5000 ideal uncapped streams with five bets at seven levels from 0.100 to 0.001. All return by 65536. This finite comparison does not establish the theorem's fixed-P limiting constant.

On real queries, larger separation in independent follow-up blocks accompanies lower audit response use. Separation is the absolute difference between the largest PASS and FAIL frequencies, with zero for an absent label. After ranking within benchmarks, its pooled association with cost is -0.946. All eight nonreturns have low separation (Appendix B.2). This empirical gap is not Γ, and capped cost is not the ideal expected stopping time.

5.3 Label versus Identity Response Cost

To isolate the added cost of same-label competition, Fixed-label and Fixed-mode differ only in the opponent set. They share the candidate, counts from response one, five bets, weights and crossing rule. The nested opponent sets make Fixed-label cross no later on the same stream (Proposition C.5). Comparing $q = s$ with $q < s$ measures the size of the added identity cost in each regime.

We supply a FAIL winner of mass $p = 0.350$, strongest PASS mass $q = 0.200$ and another FAIL mass $s \in \{0.200, 0.280\}$. Twenty PASS categories share the residual. Each regime supplies 5000 paired streams with $B = 8192$. Fixed-label omits rivals with the candidate's label. The direction or target is prespecified and aligned with the alternative for cost measurement.

Table 1: Changing only the opponent set. Entries are RMST $\pm$ one Monte Carlo SE over 5000 paired streams. Every displayed run returns.

Regime	α	Fixed-label	Fixed-mode
$q = s$	0.050	188.9 ± 1.8	244.6 ± 2.0
	0.001	395.8 ± 2.7	471.1 ± 2.7
$q < s$	0.050	187.4 ± 1.8	946.8 ± 9.7
	0.001	394.0 ± 2.7	2032.8 ± 14.5

Table 1 shows higher identity cost with a closer same-label rival, but similar label cost. At $\alpha = 0.050$, Fixed-mode costs about 1.3 times Fixed-label for $q = s$, versus 5.1 for $q < s$. Equal masses need not give equal observed maxima, and differing total PASS masses limit the opponent-set comparison to within each regime. Full RESOLVE and CITE also differ in discovery, weighting, supplied information and certification conditions (Appendix Table 11). This comparison is therefore a mechanism check, not a general ranking.

Figure 4 repeats the matched check on 198 archived GPQA and 200 MMLU-Pro queries. Both controls receive the same candidate from a disjoint 64-response reference plurality and replay identical online blocks, orders, bets, weights and thresholds. Only the opponent set differs. Extra truncated mode RMST averages 28.2 responses on 46 queries with closer same-label reference rivals, versus 0.2 on 350 without them. All 46 candidates are FAIL on these single-answer tasks. These empirical groups are not population regimes or a general method ranking. Appendix B.3 gives report rates, intervals and sensitivities.

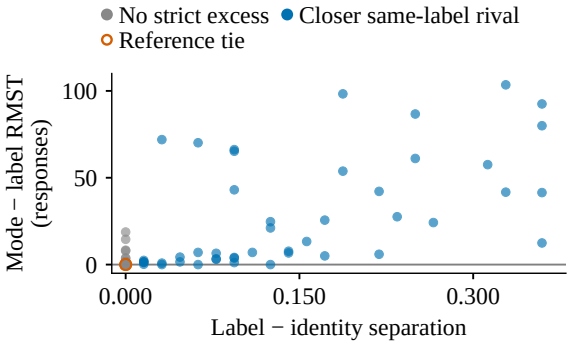


Figure 4: Fixed-mode minus Fixed-label RMST over 20 paired orders. Horizontal excess compares strongest other-category and opposite-label reference frequencies.

5.4 Alternative Label-Target Procedures

The target now stays fixed while supplied information differs. The CS methods in Table 2 know all categories and labels in advance, including unseen answers and extraction failure. RESOLVE discovers them online, so this is not a general ranking. All inspect every response on 198 archived GPQA-Diamond and 200 MMLU-Pro queries with Qwen3, using 20 paired orders and directional level 0.025. Outcome-CS (Hoeffding, 1963) uses Hoeffding intervals, Mixture-CS (Howard et al., 2021; Lindon and Malek, 2022) beta mixtures, and Multinomial-CS (Lindon and Malek, 2022) a joint Dirichlet-mixture region. Each reports when its region excludes the opposing label winning or tying.

Table 2: Label-target audits on option tasks with known complete category sets. Entries are means $\pm$ one ordering SE over 20 paired orders per query. Nonreturns pay the full response budget.

	GPQA-Diamond, $B = 128$		MMLU-Pro, $B = 64$	
Method	RMST	Report %	RMST	Report %
RESOLVE	34.54 ± 0.16	92.30 ± 0.13	23.05 ± 0.05	90.80 ± 0.13
Mixture-CS	36.47 ± 0.17	89.04 ± 0.11	21.29 ± 0.07	87.50 ± 0.14
Outcome-CS	61.02 ± 0.11	77.37 ± 0.10	36.43 ± 0.03	80.60 ± 0.05
Multinomial-CS	48.01 ± 0.16	84.62 ± 0.15	45.64 ± 0.02	75.68 ± 0.06

In Table 2, RESOLVE costs less and reports more often on GPQA-Diamond. On MMLU-Pro, Mixture-CS costs less but reports less often. Although nonreturns pay the full budget, sufficiently early returned audits can lower RMST despite fewer reports. Since methods need not resolve the same queries, the averages do not show which method reports sooner on a given query. Report rate measures returns, not correctness of the population claim. Appendix B.4 checks one shared 0.050 Multinomial-CS region and compares MMC (Cordero-Encinar and Duncan, 2025) under its different stopping target, not as a label audit. These comparisons concern a fixed policy, whereas an unfavorable audit can motivate reconsidering that policy.

From audit to re-audit. An unfavorable diagnosis concerns unchanged voting, so a proposed fix needs fresh evidence. In a separate prospective FEVEROUS study, we froze a prompt revision after development and audited both policies from response one on new queries. Of 29 baseline unfavorable reports, 2 became favorable, 24 stayed unfavorable and 3 became unresolved (Appendix Table 17). These are report transitions, not known population sign changes. The mixed result illustrates fresh assessment after revision, not reliable repair.

6 Discussion

RESOLVE diagnoses what a fixed voting policy reinforces rather than repairing that policy. A favorable report supports a PASS population label, not a sufficient finite voting budget. An unfavorable report gives anytime-valid evidence for a FAIL population plurality under the fixed policy. Additional unchanged sampling does not alter that response distribution. Unresolved means insufficient evidence, not a recommendation to keep sampling.

Validity requires i.i.d. responses and independently fixed policies, maps and deterministic evaluators, which excludes adaptive or stochastic judges. Error control is per query unless levels are allocated across a collection, and finite blocks alone do not identify population signs.

The capped monitor controls false reports, while its speed depends on separation, discovery, weights and numerical confirmation. The response bound applies to the uncapped process, not confirmation delay. Label certification avoids some contests between answers sharing a label.

Splitting or merging acceptable answers can change the plurality label without changing total PASS mass. Small cohorts and limited PASS diversity restrict generalization. Our GPQA-Diamond mirror lacks verified official rowwise equivalence, and archived maps lack demonstrated prespecification. When evaluation is available for selection, using it directly may be simpler. Changing the policy defines a new response distribution and requires fresh evidence.

AI USE STATEMENT

During this work, the authors used generative AI tools for literature search, experimental design, proof brainstorming and mathematical checks, method implementation and debugging, result interpretation, figure preparation, and language editing. The authors reviewed and verified all AI-assisted material and take full responsibility for the correctness and content of the work.

ETHICS STATEMENT

This work uses released benchmarks and model-generated responses and does not involve recruited human participants. PASS and FAIL refer only to the fixed task evaluator and should not be interpreted as general certificates of truth, safety, or fairness. The audit does not correct evaluator bias or incomplete benchmark evidence. Appendix A.2 documents dataset provenance and known limitations. Redistribution is subject to source licenses and access restrictions.

REPRODUCIBILITY STATEMENT

Appendix A documents the experimental protocol, response processing, data provenance, model settings, finite implementation, comparators, and uncertainty calculations. Appendix B provides extended results and sensitivity analyses. Appendix B.2.1 distinguishes retained historical CITE checks from unavailable simulation records, and Appendix B.5.3 describes the prompt revision and specifies its development selection rule. Appendix C contains the assumptions and complete proofs. The authors retain code, prompts, configurations, seeds, and derived records for the main analyses. Retained records do not recover missing prespecification for archived analyses, and redistribution remains subject to source licenses and access restrictions.

REFERENCES

Aggarwal, P., Madaan, A., Yang, Y., and Mausam (2023). Let's sample step by step: Adaptive-consistency for efficient reasoning and coding with LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12375–12396.

AI-MO (2024). Dataset card for AIMO validation AIME. Hugging Face dataset card, AI-MO/aimo-validation-aime.

Aly, R., Guo, Z., Schlichtkrull, M., Thorne, J., Vlachos, A., Christodoulopoulos, C., Cocarascu, O., and Mittal, A. (2021). FEVEROUS: Fact extraction and VERification over unstructured and structured information. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Amouyal, S., Wolfson, T., Rubin, O., Yoran, O., Herzig, J., and Berant, J. (2023). QAMPARI: A benchmark for open-domain questions with many answers. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 97–110.

Anand Jain, S., Shah, R., Gupta, S., Mehta, D., Nair, I. J., Vora, J., Khyalia, S., Das, S., Ribeiro, V. J., and Kalyanakrishnan, S. (2022). PAC mode estimation using PPR martingale confidence sequences. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 5815–5852.

Brown, B., Juravsky, J., Ehrlich, R., Clark, R., Le, Q. V., Ré, C., and Mirhoseini, A. (2024). Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*.

Chen, L., Davis, J., Hanin, B., Bailis, P., Stoica, I., Zaharia, M., and Zou, J. (2024). Are more LLM calls all you need? towards the scaling properties of compound AI systems. In *Advances in Neural Information Processing Systems*, volume 37, pages 45767–45790.

Chen, L., Gupta, A., Li, J., Qiao, M., and Wang, R. (2017). Nearly optimal sampling algorithms for combinatorial pure exploration. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 482–534.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. (2021). Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Cordero-Encinar, P. and Duncan, A. B. (2025). Certified self-consistency: Statistical guarantees and test-time training for reliable reasoning in LLMs. *arXiv preprint arXiv:2510.17472*.

Degenne, R. and Koolen, W. M. (2019). Pure exploration with multiple correct answers. In *Advances in Neural Information Processing Systems*, volume 32.

Dekoninck, J., Jovanović, N., Gehrunger, T., Rögnvaldsson, K., Petrov, I., Sun, C., and Vechev, M. (2026). Beyond benchmarks: MathArena as an evaluation platform for mathematics with LLMs. *arXiv preprint arXiv:2605.00674*.

Farquhar, S., Kossen, J., Kuhn, L., and Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Feng, A., Alonso, M., Odonnat, A., Feofanov, V., and Redko, I. (2026). Optimal self-consistency for efficient reasoning with large language models. In *Forty-Third International Conference on Machine Learning*.

Gao, T., Yen, H., Yu, J., and Chen, D. (2023). Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488.

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. (2021). Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

hendrydong (n.d.). gpqa_diamond_mc: Ungated dataset mirror, test split. Hugging Face dataset card, hendrydong/gpqa_diamond_mc.

Hoeffding, W. (1953). A lower bound for the average sample number of a sequential test. *The Annals of Mathematical Statistics*, 24(1):127–130.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30.

Howard, S. R., Ramdas, A., McAuliffe, J., and Sekhon, J. (2021). Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080.

Hugging Face H4 (2024). Dataset card for MATH-500. Hugging Face dataset card, HuggingFaceH4/MATH-500.

Kaufmann, E., Cappé, O., and Garivier, A. (2016). On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1):1–42.

Komiyama, J., Oba, D., and Oyamada, M. (2026). Best-of-infinity: Asymptotic performance of test-time LLM ensembling. In *International Conference on Learning Representations*.

Kuhn, L., Gal, Y., and Farquhar, S. (2023). Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*.

Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J. E., Zhang, H., and Stoica, I. (2023). Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

Kydlíček, H. (2026). Math-Verify: Math verification library. Software, version 0.9.0.

Li, Y., Yuan, P., Feng, S., Pan, B., Wang, X., Sun, B., Wang, H., and Li, K. (2024). Escape sky-high cost: Early-stopping self-consistency for multi-step reasoning. In *International Conference on Learning Representations*.

Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., and Cobbe, K. (2024). Let's verify step by step. In *International Conference on Learning Representations*, pages 39578–39601.

Lindon, M. and Malek, A. (2022). Anytime-valid inference for multinomial count data. In *Advances in Neural Information Processing Systems*, volume 35, pages 2817–2831.

MathArena (2026). AIME 2026 dataset card. Hugging Face dataset card, MathArena/aime_2026.

OpenAI (2025). gpt-oss-120b & gpt-oss-20b model card. OpenAI model card.

Ota, H., Iwase, N., Ichihara, Y., Komiyama, J., and Imaizumi, M. (2026). CITE: Anytime-valid statistical inference in LLM self-consistency. *arXiv preprint arXiv:2605.05873*.

Qwen (2024a). Qwen2.5-7B-Instruct model card. Hugging Face model card, Qwen/Qwen2.5-7B-Instruct.

Qwen (2024b). Qwen2.5-Math-1.5B-Instruct model card. Hugging Face model card, Qwen/Qwen2.5-Math-1.5B-Instruct.

Qwen (2025). Qwen3-30B-A3B-Instruct-2507 model card. Hugging Face model card, Qwen/Qwen3-30B-A3B-Instruct-2507.

Qwen, Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Tang, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., and Qiu, Z. (2024). Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Ramdas, A., Ruf, J., Larsson, M., and Koolen, W. (2020). Admissible anytime-valid sequential inference must rely on nonnegative martingales. *arXiv preprint arXiv:2009.03167*.

Ramdas, A. and Wang, R. (2025). Hypothesis testing with e-values. *Foundations and Trends® in Statistics*, 1(1-2):1–390.

Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J., and Bowman, S. R. (2024). GPQA: A graduate-level Google-proof Q&A benchmark. In *First Conference on Language Modeling*.

Scheffé, H. (1947). A useful convergence theorem for probability distributions. *The Annals of Mathematical Statistics*, 18(3):434–438.

Shafer, G. (2021). Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184(2):407–431.

Snell, C., Lee, J., Xu, K., and Kumar, A. (2025). Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations*, pages 10131–10165.

Ville, J. (1939). *Étude Critique de la Notion de Collectif*. Gauthier-Villars.

Wald, A. (1945). Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186.

Wang, P., Li, L., Chen, L., Cai, Z., Zhu, D., Lin, B., Cao, Y., Kong, L., Liu, Q., Liu, T., and Sui, Z. (2024a). Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450.

Wang, X., Wei, J., Schuurmans, D., Le, Q. V., Chi, E. H., Narang, S., Chowdhery, A., and Zhou, D. (2023). Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*.

Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S., Ren, W., Arulraj, A., He, X., Jiang, Z., Li, T., Ku, M., Wang, K., Zhuang, A., Fan, R., Yue, X., and Chen, W. (2024b). MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems*, volume 37, pages 95266–95290.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al. (2025). Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yang, A., Zhang, B., Hui, B., Gao, B., Yu, B., Li, C., Liu, D., Tu, J., Zhou, J., Lin, J., Lu, K., Xue, M., Lin, R., Liu, T., Ren, X., and Zhang, Z. (2024). Qwen2.5-Math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623.

Zhou, J., Lu, T., Mishra, S., Brahma, S., Basu, S., Luan, Y., Zhou, D., and Hou, L. (2023). Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.

APPENDIX

Overview of Appendices. Appendix A specifies the experimental protocol and implementation, and Appendix B extends the main experimental questions through sensitivity checks and policy re-auditing. Appendix C proves the population, validity and cost results and comparator guarantees, while Appendix D explains their relation to prior work.

A EXPERIMENTAL PROTOCOL AND IMPLEMENTATION

A.1 TASKS, RESPONSE CATEGORIES, AND FROZEN EVALUATORS

The category map determines which responses share a vote. Table 3 collects the five fixed evaluators. Reference answers and annotation membership are hidden from the model.

Table 3: Vote categories and evaluators. The last two tasks allow multiple PASS categories. Extraction failures remain in the response stream.

Task	Vote category	PASS rule	Important scope
AIME 2026	Normalized integer	Matches the answer key	Derivation is not evaluated.
GPQA-Diamond	Normalized option letter	Matches the science answer key	Missing or ambiguous extraction is a fixed FAIL category.
MMLU-Pro	Normalized option letter	Matches the subject answer key	Same extraction-failure rule as the other reasoning tasks.
QAMPARI-based QA	One normalized entity, with benchmark aliases merged	Validated gold entity with at least one retained annotated supporting passage matched in the frozen context	Different entities remain separate. A list is extraction failure. `NO_ANSWER` is failing abstention.
FEVEROUS-based verification	Verdict and unordered, deduplicated evidence-ID set	Reference verdict and at least one complete annotated evidence set	Valid supersets remain distinct. Wrong verdicts, unknown IDs and malformed outputs fail.

The reasoning parsers use the last nonempty balanced box or a guarded answer phrase, also allowing a final option-only line for option tasks. They ignore reasoning and retain missing or ambiguous extraction as FAIL. Normalization maps $\boxed{017}$ to `17` and `b` to `B`.

For the Cambridge museum query, Cambridge Science Centre and Cambridge University Museum of Zoology are distinct PASS entities. Museum of Cambridge is gold and mentioned in context but lacks a retained annotation match, so fails. The Boggeragh Mountains claim requires elevation 644 m and location in County Cork. With SUPPORTS, $\{\texttt{E004}, \texttt{E006}\}$ and $\{\texttt{E005}, \texttt{E006}\}$ pass, combining an elevation cell or sentence with a location sentence. The set $\{\texttt{E006}, \texttt{E044}\}$ fails because E044 concerns Fastnet Rock, not elevation. Both FAIL examples were first audit responses. PASS examples are post-collection rule checks, not claimed generations.

Under these fixed evaluators, FAIL need not mean factual error or absence of textual support. QAMPARI's non-gold, missing-annotation, extraction-failure and other FAIL reasons are diagnostics only. All tasks use reference information. No answer-free factual judge or adaptive or stochastic evaluator is tested.

A.2 DATA COHORTS, PROVENANCE, AND FREEZE CHRONOLOGY

Table 4 distinguishes the prospective cohorts from archived blocks. Prospective queries and rules are fixed before generation, and audit diagnoses are locked before follow-up generation under a frozen protocol. For the archived analyses reported here, records do not establish independent prespecification of maps, evaluators, unparseable-output rules and tie rules. Disjoint blocks prevent reuse but do not repair this chronology. Archived analyses have no claimed nominal anytime or familywise guarantee.

The policies are Qwen2.5-7B-Instruct, Qwen3-30B-A3B-Instruct-2507 and gpt-oss-20b. The inherited mathematics pool uses Qwen2.5-Math-1.5B-Instruct (Qwen et al., 2024; Qwen, 2024a; Yang et al., 2025; Qwen, 2025; OpenAI, 2025; Yang et al., 2024; Qwen, 2024b).

AIME 2026 uses the MathArena release. The inherited mathematics pool uses 20 AIMO validation AIME records and 13 MATH-500 items (MathArena, 2026; Dekoninck et al., 2026; AI-MO, 2024; Hendrycks et al., 2021; Lightman et al., 2024; Hugging Face H4, 2024). MMLU-Pro uses the TIGER-Lab release (Wang et al., 2024b). GPQA-Diamond uses the ungated multiple-choice mirror cited here, not a verified copy of the gated official release (Rein et al., 2024; hendrydong, nd). The prospective freezes pin the selected mirror contents before generation, but do not establish official rowwise equivalence or recover the archived acquisition revision.

Archived MMLU-Pro has 25 queries per domain across eight domains, whereas the prospective subset covers all 14. Archive splits use response identifiers where available, otherwise nonoverlapping array positions. Public pools retain their source parser and exact-match evaluator (Komiyama et al., 2026).

Table 4: Cohorts used in the main evidence. Decoder entries give temperature, top-p and top-k. Every audit uses $K_{\text{cap}} = 50$. Follow-up or reference blocks contain 64 responses per query. Model names are shortened here only.

Tasks	Queries	Policy	Status	Decoder	B
Three reasoning tasks	60	Qwen2.5-7B	Prospective	0.700/0.900/20	160
Two evidence tasks	80	Qwen2.5-7B	Prospective	0.700/0.900/20	160
Two evidence tasks	80	Qwen3-30B	Prospective	0.700/0.900/20	160
Two evidence tasks	80	gpt-oss-20b	Prospective	0.400/0.900/20	160
Archived GPQA-Diamond	198	Qwen3-30B	Retrospective	0.700/0.800/20	128
Archived MMLU-Pro	200	Qwen3-30B	Retrospective	0.700/0.800/20	64

Prospective Audits and Independent Follow-Up. Selection uses no generated outcomes. AIME and GPQA-Diamond take the first 20 source indices in a fixed hash ranking. MMLU-Pro cycles through alphabetically ordered domains, taking ranked questions until 20 are selected. A local freeze pins questions, checkpoint, prompts, decoder, maps, evaluator, tie rule, monitor, seeds and summaries. Each query uses $\alpha_{\text{query}} = 0.050$ with $\alpha_0 = \alpha_1 = 0.025$, not 5% simultaneous protection for the collection.

Plurality uses first-appearance ties and the fixed categories, including entity aliases, but never PASS/FAIL labels. Prompts withhold reference answers, verdicts and annotation membership. Truncated and unparseable responses remain, and ambiguous requests are not retried. Within a shard, each active query receives one response and monitor update before another request. Stopped audits are never continued with historical or follow-up responses.

Evidence-Based Answers and Verification. Evidence cohorts are selected from metadata before generation and use the same Qwen2.5 audit configuration. Exact arithmetic replay checks every audit prefix.

All 1000 ALCE QAMPARI query IDs, questions and aliases match the original development split (Amouyal et al., 2023; Gao et al., 2023). We retain the top 10 GTR passages without oracle reranking. Selection requires at least three validated answers and two with all retained annotations present, excluding ambiguous aliases and numeric year/season questions. The 40 selected queries have two to six PASS-eligible entities each. The evaluator needs only one retained support passage per entity, matched by normalized text and source title. Selection is thus stricter than evaluation. Annotation matches need not establish every relation in a composed question or exhaust factual support.

FEVEROUS retrieval uses only the claim and matching Wikipedia snapshot, not gold evidence (Aly et al., 2021). It searches 20 candidate titles and keeps five pages. It ranks their elements by BM25 with $k_1 = 1.200$ and $b = 0.750$, then considers the top 60 candidates for packing with their structural context within 5800 tokens. Annotations then identify 383 eligible claims. Selection cycles through text, structured and mixed evidence profiles in claim-ID order within each verdict, taking 20 SUPPORTS and 20 REFUTES claims and excluding not-enough-information claims. The

selected profiles contain 12, 15 and 13 claims. Thirty-eight claims have two complete eligible sets and two have three.

Interfaces and Decoder Settings. Token caps are 6144 for Qwen2.5 reasoning, 128 for entity QA and 256 for verification. Native chat templates preserve task messages and frozen contexts, with boxed answers requested for reasoning. The Qwen2.5 audit uses FP16, two-way tensor parallelism, vLLM 0.8.5 post-release 1 and no prefix caching (Kwon et al., 2023). Table 4 gives decoder settings.

Clean gpt-oss-20b uses native Harmony, medium reasoning effort and a 12288-token cap covering analysis and final output. Only the final channel enters extraction. Native quantized weights use BF16 activations and two-way tensor parallelism. Interface checks precede generation. Archived gpt-oss has the same recorded decoder values. Archived Qwen3 uses caps 8192 on MMLU-Pro and 12288 on AIME and GPQA-Diamond. Inherited Qwen2.5-Math decoding settings are undocumented, not inferred from later runs.

Follow-up protocols fix inclusion, seeds, generation and summaries before collection, which begins only after all audit diagnoses are locked. The 60 reasoning queries retain all 3840 responses, including 46 truncations. The 80 evidence queries add 5120, giving 8960 follow-up responses with separate costs. Response and engine seeds are disjoint from earlier campaigns. These local records are not externally timestamped, and disjoint seeds do not prove independent, stationary sampling.

A.3 Finite RESOLVE Implementation

Algorithm 1 makes the post-discovery counting and exact confirmation in Section 3 explicit. Account (c, j) stores the opponent snapshot $S_{c,j}(b) = N_{T_{c,j}}(b)$, with absent count and snapshot entries zero. This gives the largest opponent count strictly after discovery,

$$M_t(R_{c,j}) = \max\big(\{0\} \cup \{N_t(b) - S_{c,j}(b) : b \in \mathcal{A}_{1-c}\}\big).$$

Algorithm 1 Finite RESOLVE with post-discovery counts and confirmed reporting

Require: Fixed policy, response-wise category map and deterministic evaluator h
Require: Fixed rational levels α_0, α_1, grid Λ and weights $v_\lambda, \bar{w}_j$ specified below
Require: Response budget B and account cap K_{cap} per direction
Initialize an empty label cache, count map N and accounts, with $K_0 = K_1 = 0$.
for $t = 1, \ldots, B$ **do**
  Generate and map $a = X_t$. Record whether it is new and cache $h(a)$ if so.
  Set $c_a = h(a)$ from the cache and increment $N(a)$.
  Update every existing account's C, M by (1), using its snapshot S.
  if a is new and $K_{c_a} < K_{\text{cap}}$ **then**
    Open $j = K_{c_a} + 1$ with $R_{c_a,j} = a$, $T_{c_a,j} = t$ and $C(a) = M(a) = 0$. Set $K_{c_a} = j$.
    Store $S_{c_a,j}(b) = N(b)$ for every observed opposite-label category b.
  end if
  for $c = 1, 0$, in this order to give PASS priority **do**
    Compute the log evidence screen from (2), including unused mass.
    if the screen is at least $\log(1/\alpha_c)$ **then**
      Recompute all component accounts and $\bar{E}_t^c$ exactly using rational arithmetic.
      If $\alpha_c \bar{E}_t^c \geq 1$, return favorable for $c = 1$ or unfavorable for $c = 0$, with count t.
    end if
  end for
end for
Return unresolved after B responses.

We use $K_{\text{cap}} = 50$, $\Lambda = \{2^{-1}, \ldots, 2^{-5}\}$ and $v_\lambda = 1/5$. Stored binary64 candidate weights are exact rational approximations to $6/(\pi^2 j^2)$, not ideal weights, and are nonnegative with $\sum_{j=1}^{K_{\text{cap}}} \bar{w}_j < 1$. Positive levels are exact decimal fractions satisfying $\alpha_0 + \alpha_1 \leq \alpha_{\text{query}} < 1$. Any finite rational grid in $(0, 1)$ with nonnegative rational weights summing to one is allowed. Log-sum-exp screening can delay exact confirmation, so $\rho_c \geq \bar{\tau}_c(\alpha_c)$, with no delay bound.

Let D_B count distinct categories observed by budget B. With five bets, incremental updates and screening cost $O(K_{\text{cap}})$ per response. Opening each of at most $2K_{\text{cap}}$ accounts copies observed opponent counts, costing $O(K_{\text{cap}} D_B)$ in total. Since $D_B \leq B$, total bookkeeping is $O(BK_{\text{cap}})$ and memory is $O(D_B + K_{\text{cap}} D_B)$, treating dictionary and arithmetic operations as unit cost. Evaluator work and exact confirmation, including rational bit complexity, are additional.

Smaller bets can extend the grid's range of positive growth without changing validity, but a finite grid can still miss narrow label margins.

A.4 Comparator Definitions

Table 5 distinguishes targets and information. Fixed-label and Fixed-mode share counts from response one, five bets and level α, differing only in opponent set. Their simulation targets are supplied. Archived replays instead use 64 reference responses and directional level 0.025 (Appendix B.3). Fixed-mode is not full CITE.

Table 5: Comparator information and allocations. Reference responses are additional to the monitored stream. Levels describe complete-block replays unless marked as controlled.

Method	Target	Supplied candidate	Known categories / labels	Extra block	Level allocation	Key distinction
RESOLVE	Label	No	No / online	None	0.025 per direction	Discovers candidates
Fixed-label	Label	Yes	Yes / yes	None	α, controlled	Opposite-label rivals only
Fixed-mode	Identity	Yes	Yes / unused	None	α, controlled	All other categories
Outcome-CS	Label	No	Yes / yes	None	0.025 per direction	Hoeffding intervals
Mixture-CS	Label	No	Yes / yes	None	0.025 per direction	Beta mixtures
Multinomial-CS	Label	No	Yes / yes	None	0.025 per direction	Joint region
CITE	Identity	Yes	No / no	64 or none	$0.050/3$ per component	Reference or oracle target
MMC	Different stopping target	No	Yes / unused	None	Nominal $\delta = 0.050$	Changing contenders

Let d be the number of known categories, including extraction failure, with $d = 5$ for GPQA-Diamond and $d = 11$ for MMLU-Pro. Outcome-CS uses clipped Hoeffding intervals with coordinate/time level $0.025/[dt(t+1)]$ (Hoeffding, 1963). Mixture-CS inverts Jeffreys beta mixtures at $0.025/d$ per coordinate (Howard et al., 2021; Lindon and Malek, 2022). Both report when the largest candidate-label lower endpoint exceeds all opposing upper endpoints. Multinomial-CS inverts a uniform Dirichlet-mixture region over the label null at 0.025, using Lemma C.6 (Lindon and Malek, 2022). One shared region can instead protect both directions (Appendix B.4). Mixture-CS and Multinomial-CS proposals are exactly confirmed. All methods inspect from response one. An empty candidate class cannot report. With no opposing class, interval methods need a positive lower endpoint and Multinomial-CS may report immediately. Appendix C.5 gives the proofs. Future-support replays in Appendix B.4 lack the guarantee for a category set known in advance.

Our independent full-CITE implementation follows Algorithm 1 and Definitions 3.1, 3.3 and 3.5 of Ota et al. (2026). Its fixed target is either a disjoint 64-response reference plurality with first-appearance ties, charged to cost, or the true mode supplied free in simulations. The monitored block never selects the target. CITE splits its level into $0.050/3$ per component in replays or $\alpha/3$ in controlled sweeps. It requires the pairwise check and a target-mass lower bound above the unseen-category upper bound. Empty competition passes only the pairwise check. Uniform grids $\{2^{-1}, \ldots, 2^{-13}\}$ and $\{2^{-1}, \ldots, 2^{-8}\}$ implement the published pairwise and target-mass families, not claimed original defaults. Appendix C.6 gives exact confirmation and validity. Appendix B.2.1 describes retained historical checks and their replication limits.

For Table 14, we use the authors' released ratio-update MMC implementation with $M = 2$ and our nominal $\delta = 0.050$. It compares the predictable leader with the runner-up and the aggregate outside their top-two set. Each channel needs at least one update and log evidence at least $\log(20)$. Before an update, the point alternative is the success count plus $1/2$, divided by the total update count plus one, clipped to $[0.50001, 0.99999]$. Multiply evidence by twice this alternative on success or twice its complement on failure. Counts start at zero and wealth at one, without resets when contenders change. Ties use fixed alphabetic order. The code returns the leader selected before the latest response, whereas the printed pseudocode returns the updated leader (Cordero-Encinar and Duncan, 2025, Algorithm 1). The repository's LLM evaluation code uses a different cumulative score update. Labels are applied only after return.

A.5 Metrics, Uncertainty, and Replay Protocol

Report rate and RMST follow Section 5, including nonreturns. CITE adds 64 reference responses per query when reference-selected. Follow-up costs are separate.

To compute PASS availability for a block of N_i responses containing K_i passes, we use

$$A_i(n) = 1 - \frac{\binom{N_i - K_i}{n}}{\binom{N_i}{n}}, \qquad 1 \le n \le N_i.$$

A binomial coefficient is zero when its lower argument exceeds its upper one. Fresh-block plurality uses first-appearance ties and 2000 sampled orders per query. Availability and endpoints whose tied leaders share a label are exact conditional on the blocks, not on new responses. Archived behavior retains its recorded 200–300 orders. Audit groups are descriptive, not randomized.

Monte Carlo SE measures uncertainty across independent simulated streams from a known distribution. Ordering SE combines within-query permutation variances with equal query weights. Unless stated otherwise, $\pm$ gives one SE of the stated type. Reorderings are not new responses, and original-order counts have no ordering SE. Across-query SE describes heterogeneity of query means, not new responses or population sampling. Query-bootstrap intervals condition on selected strata. Zero observed errors require an upper bound, not a claim of zero risk.

B Additional Experimental Results

B.1 Availability, Evaluation, and Plurality

Table 6 checks whether the distinction between availability and selection in Section 5.1 persists across all five benchmarks, using equal-query endpoints. Unfavorable groups can contain PASS responses without selecting them. Every favorable query has a passing full-block plurality, while unresolved groups have mixed endpoints. Finite blocks identify neither population signs nor convergence.

Table 6: Fresh-block endpoints for unfavorable and unresolved Qwen2.5 audit groups, exact conditional on the retained blocks. "None" means no queries, not a zero probability.

	Unfavorable group			Unresolved group
Benchmark	Queries	$A(64)$	$V(64)$	Queries / $V(64)$
AIME 2026	14	0.143	0.000	5/0.000
GPQA-Diamond	15	1.000	0.000	1/1.000
MMLU-Pro	11	0.818	0.000	2/0.500
QAMPARI-based QA	23	0.087	0.000	none
FEVEROUS-based verification	30	0.033	0.000	none

Evidence Requirements and Generated Diversity. Both Qwen2.5 evidence cohorts resolve by $B = 160$, but only two QAMPARI queries and four FEVEROUS claims generate multiple PASS categories. Three FEVEROUS claims have different matched-annotation collections, yet their passing outputs share a complete set within each claim. Extra IDs can change identity without a new

route to PASS. Correct verdicts appear in $30/40$ blocks, but only $11/40$ contain a response meeting the full evidence requirement, explaining Figure 2(c). Diagnostic FAIL reasons never change stopping.

Fixed PASS Mass and Fragmentation. To separate total PASS mass from its concentration, the controlled distributions fix $\pi_1 = 0.600$ and a FAIL maximum 0.300. Their strongest PASS masses are 0.400, 0.310, 0.200, 0.100 and 0.050. Each distribution supplies 10000 independent streams to budget 1024, with query level 0.050 split equally. The same evaluator labels and voting rule apply throughout a stream.

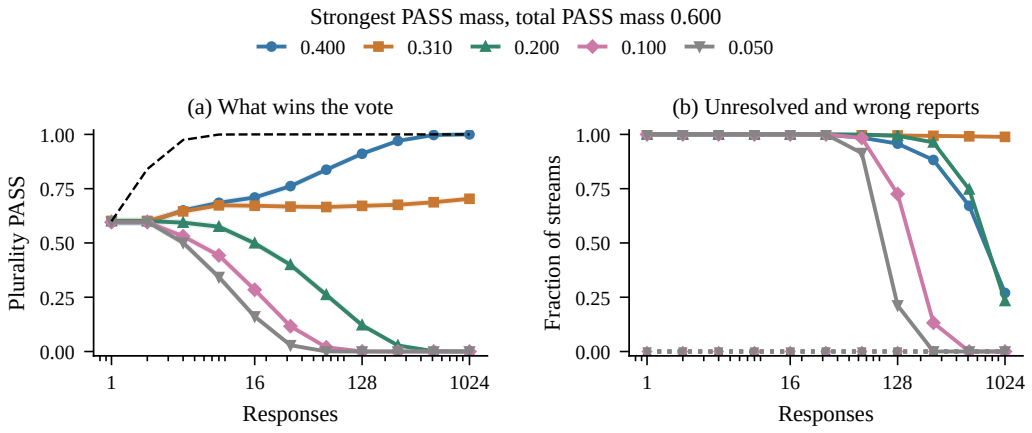


Figure 5: Fixed total PASS mass with different strongest PASS masses, identified by markers. (a) Plurality selection with pointwise 95% binomial Monte Carlo intervals and analytic availability in dashed black. (b) Unresolved fractions in solid lines and wrong reports in dotted lines.

In Figure 5, unchanged availability accompanies a switch to FAIL plurality when its leader exceeds every PASS category, illustrating Proposition 2.1. At gap 0.010, most audits remain unresolved even at the largest budget, not an information-theoretic lower bound. Across QAMPARI and FEVEROUS with all three fixed policies, no query block has empirical PASS frequency at least 0.500 and a negative empirical gap. Less restrictive cutoffs yield isolated cases, not population prevalence.

Pooling labels answers a different question. A Jeffreys beta-mixture confidence sequence (Howard et al., 2021; Lindon and Malek, 2022) reports whether total PASS mass exceeds $1/2$. Its shared region at threshold 40 has anytime coverage error at most 0.025. It reports PASS on 99.65% to 99.75% of streams by 1024 responses, with Monte Carlo SE at most 0.06 percentage points. All five settings have total PASS mass 0.600, yet plurality favors FAIL at strongest PASS masses 0.200, 0.100 and 0.050. Thus a valid audit of total label mass does not answer the plurality question.

B.2 Continuous Monitoring and Audit Difficulty

Figure 6 extends Section 5.2 to all 140 primary paths, including eight nonreturns. Paths stop at a confirmed crossing of 40 or budget 160. Median use is 21 responses. Conditional on returning, mean and median are 31.8 and 20, using 80.1% less than the full allocation. Follow-up costs are excluded, and paths do not reveal population signs.

One unfavorable AIME 2026 report is driven by extraction failure alone, unlike every unfavorable GPQA-Diamond and MMLU-Pro report. It concerns the full extraction-and-voting policy, not necessarily a recurring substantive wrong answer.

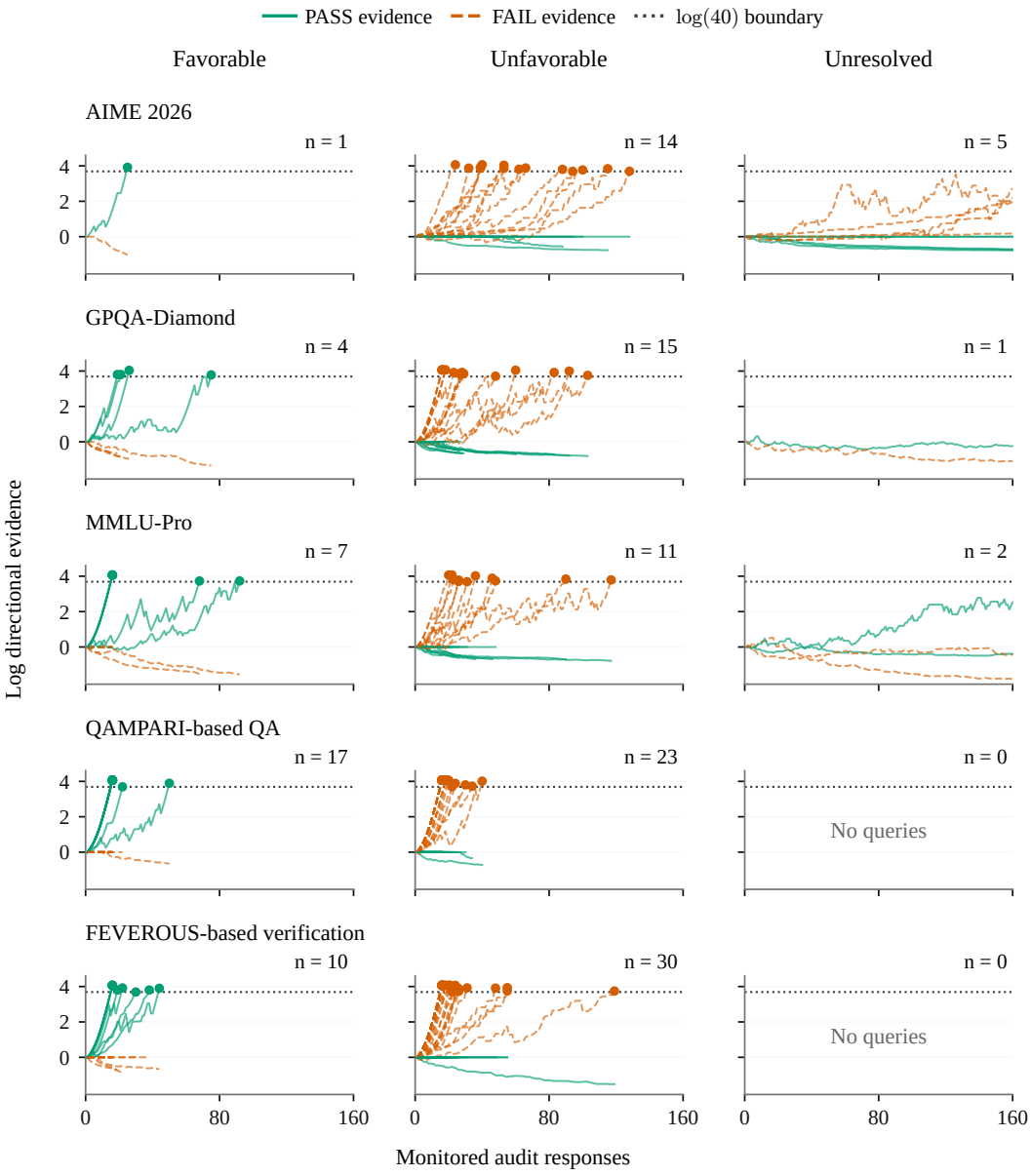


Figure 6: Monitoring paths for every primary query, grouped by benchmark and recorded outcome. Each query contributes a solid PASS path and a dashed FAIL path. Overlapping paths are retained, and n gives the query count. Dots mark confirmed reports. The dotted line is the reporting boundary $\log(40)$.

B.2.1 False Reports under Repeated Inspection

Figure 7 checks whether false reports accumulate under repeated inspection, first at a binary tie and then with more categories. At the binary tie, integer recursion counts noncrossing prefixes by removing newly rejecting states. Enumeration checks it through 16, and integer tails give boundaries through 1024. One prespecified direction uses level 0.050 on 30000 RESOLVE streams. At 1024, once-only, repeated and RESOLVE errors are 4.881%, 34.013% and (1.823±0.077)%, respectively, the last comprising 547 false reports.

Six multicategory nulls retain archived supports and labels, projecting masses to rational label ties. Raising a mode of the lower-max label and clipping opposite categories above a common level preserves total mass. One setting is tied already, and two need substantial transfer (Table 7). Each supplies 50000 streams through 2000, splitting query level 0.050 equally. The invalid comparator repeatedly selects leaders and direction for an unadjusted one-sided binomial tail.

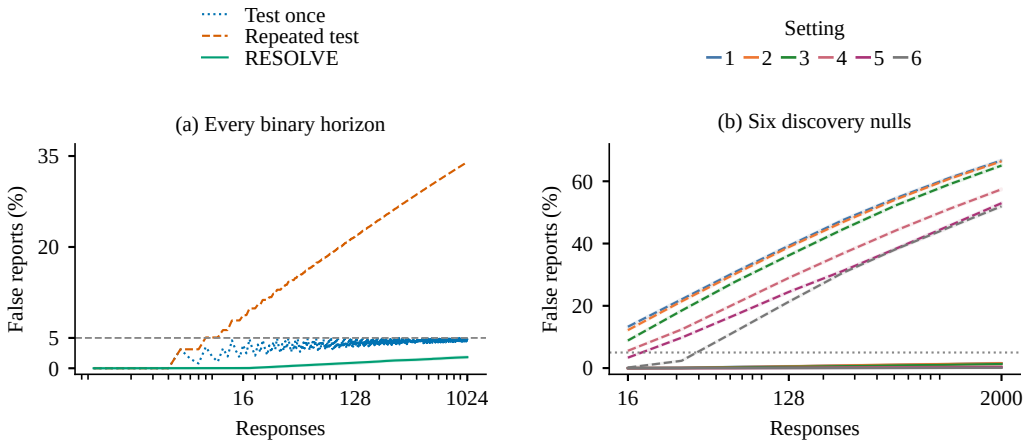


Figure 7: Continuous-inspection checks. (a) Binary tie through every horizon to 1024. (b) Six projected multicategory label ties through 2000 responses. Solid curves show RESOLVE and dashed curves repeated tests. Bands give one Monte Carlo SE. Horizontal lines mark 5%.

Table 7: Six projected label ties and their false-report percentages at 2000 responses. Entries give one independent-stream Monte Carlo SE.

Setting	Categories	Common max.	Moved mass	RESOLVE error (%)	Repeated test (%)
1	13	0.40625	0.26563	1.534 ± 0.055	66.700 ± 0.211
2	16	0.38281	0.32031	1.558 ± 0.055	66.414 ± 0.211
3	16	0.32031	0.11719	1.298 ± 0.051	65.050 ± 0.213
4	20	0.18229	0.16667	0.468 ± 0.031	57.426 ± 0.221
5	22	0.14063	0.00000	0.240 ± 0.022	52.964 ± 0.223
6	49	0.04297	0.02734	0.090 ± 0.013	51.962 ± 0.223

In Table 7, all six RESOLVE errors lie below the budget, an upper bound rather than a target rate. These finite null checks do not prove infinite-horizon validity. The comparator also reselects its target, so excess error cannot be attributed solely to inspection frequency.

Boundary and Numerical Checks. The supplied FAIL candidate and strongest PASS category each have mass 0.300, another FAIL has 0.200, and 20 PASS categories share the residual. Each method tests one prespecified target or direction at 0.050 on 20000 streams. By 2000, RESOLVE, Fixed-label and Fixed-mode have 261, 622 and 510 false reports, or $(1.305 \pm 0.080)\%$, $(3.110 \pm 0.123)\%$ and $(2.550 \pm 0.111)\%$. The three CS methods have zero, with one-sided 95% upper limit 0.015%. RESOLVE reports are rationally confirmed and fixed controls use exact integer boundaries. Null RMST mainly measures censoring.

The retained full-CITE validation record includes deterministic count checks and aggregate results from two historical null simulations. Replication-level simulation records are unavailable, so the aggregates alone cannot reproduce those runs.

Paired replays of these binary streams split query level 0.050 equally between directions. False-report percentages are 1.767 ± 0.076, 1.463 ± 0.069, 1.070 ± 0.059 and 0.740 ± 0.049 by 1024, inspecting every response or every 4, 16 or 64 responses, with one Monte Carlo SE. Every response updates evidence. Scheduled checks use its actual value, not rounded dense crossing times. These are not new replications. Sparser inspection reduces observed finite-horizon error but is unnecessary for Theorem 4.1.

B.2.2 Reporting near a Label Tie

To separate validity from the chance of obtaining a report, we fix a passing maximum $p = 0.350$, failing maximum $q \in \{0.050, 0.150, 0.250, 0.300, 0.330\}$ and 20 failing categories sharing the residual. Each distribution supplies 10000 streams with $B = 1024$ and query level 0.050 split equally. The cap, bets, discovery and exact confirmation are unchanged. Complemented labels reuse streams, not add replications.

In Figure 8(a), at gap 0.020 and budget 160, there are 56 correct reports, 11 wrong reports and 9933 nonreturns. The unconditional error is 0.110%, but $11/67$ returns are wrong. This conditional error of 16.418% has pointwise 95% Clopper–Pearson Monte Carlo interval $[8.491, 27.480]\%$, conditional on the number of returns. Theorem 4.1 does not bound this conditional quantity. With the larger gap 0.300, 9995 correct reports and no errors occur by 160. Zero errors give a one-sided 95% upper bound of 0.030%, not zero risk.

Forced plurality on the same streams at 160, with first-appearance ties, makes 3709 errors at gap 0.020 and none at gap 0.300. Abstention matters most near the boundary. This replay varies P across curves, unlike Theorem 4.3's fixed-P limit.

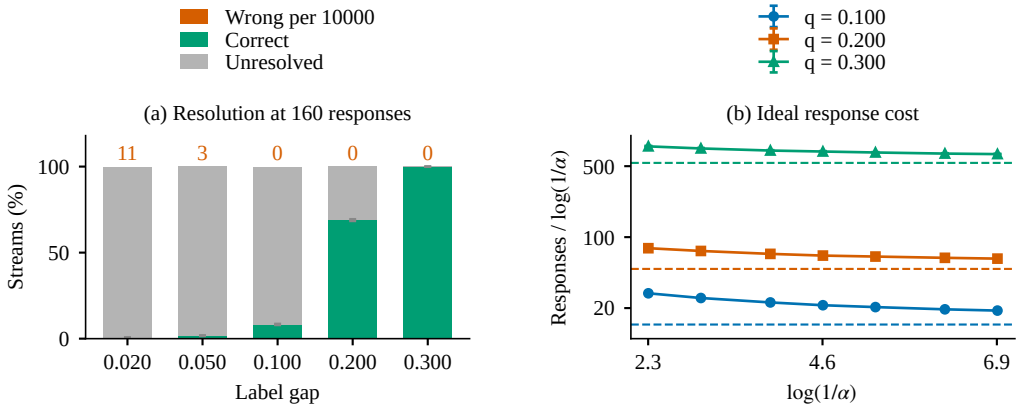


Figure 8: Finite-budget reports and ideal response cost. (a) Correct, wrong and unresolved streams at 160 responses as the label gap varies. Orange labels count wrong reports. (b) Mean stopping time divided by $\log(1/\alpha)$, with dashed $1/\mathcal{I}_\Lambda(p, q)$ references and one pointwise Monte Carlo SE.

B.2.3 Response Cost after Validity

For Theorem 4.3's error-level dependence, category 0 alone passes with mass 0.350. One FAIL category has mass $q \in \{0.100, 0.200, 0.300\}$ and 20 others each have mass $(0.650 - q)/20$. Each distribution supplies 5000 independent streams shared across seven levels from 0.100 to 0.001, shown in Figure 8(b) with endpoints in Table 8.

With $\Lambda = \{2^{-1}, \ldots, 2^{-5}\}$, equal grid weights and $w_j = 6/(\pi^2 j^2)$, evidence is one before PASS discovery and $E_t^1 = 1 + w_1\{\frac{1}{5}\sum_{\lambda\in\Lambda} G_t(0;\lambda) - 1\}$ afterward. Unopened accounts remain at one,

counts exclude discovery, and no account cap is imposed. This specializes the theorem but does not test competition among multiple passing candidates.

Configuration, seeds and code were recorded before sampling. All streams return before 65536, without censoring. Interval arithmetic enclosing the mathematical constant π checks neighboring integer boundaries, and a separate replay checks every saved prefix.

Table 8: Endpoint ideal mean stopping times behind Figure 8(b), in responses $\pm$ one pointwise Monte Carlo SE over 5000 paired streams per distribution.

α	$q = 0.100$	$q = 0.200$	$q = 0.300$
0.100	64.4 ± 0.4	179.1 ± 1.7	1803.7 ± 19.8
0.001	130.6 ± 0.6	424.3 ± 2.8	4540.7 ± 30.5

Stronger competition raises cost, while dividing by $\log(1/\alpha)$ in Figure 8(b) reveals approximate logarithmic scaling. Reciprocal rates are asymptotic references, not finite-level upper bounds. These ideal-weight costs differ from Table 11's stored-weight replay and do not describe the capped monitor.

Online Discovery at Fixed Maximizing Masses. The leading rate need not capture discovery cost at a finite budget. We fix $p = 0.350$ and $q = 0.200$, with $m \in \{0, 3, 30\}$ PASS distractors sharing 0.300 mass when present. Twenty smaller FAIL categories share 0.450 for $m = 0$ and 0.150 otherwise. Comparing 3 and 30 fixes total PASS mass too. Each setting supplies 1000 independent streams with $B = 1024$, $K_{\text{cap}} = 50$, five uniform dyadic bets and 0.025 per direction. Rational checks cover every report and preceding prefix, including every prefix of 30 streams.

Table 9: Discovery overhead at fixed p, q. RMST and mean discovery ranks have one Monte Carlo SE. The decisive rank is the discovery index of the largest weighted account at a report, conditional on returning. Unresolved entries are counts out of 1000.

Distractors	RMST	Unresolved	Best-category rank	Decisive rank
0	256.0 ± 4.8	0	1.000 ± 0.000	1.000 ± 0.000
3	295.5 ± 5.4	1	1.637 ± 0.026	1.636 ± 0.026
30	307.8 ± 5.6	1	1.816 ± 0.037	1.812 ± 0.037

In Table 9, higher costs reflect discovery and weighting together, not rank alone, despite an unchanged leading rate. Every stream retains the maximizing candidate. Each nonzero nonreturn rate has Monte Carlo SE 0.1 percentage points. No wrong report is observed, with one-sided 95% upper bound 0.299% per setting.

Held-Out Separation and Audit Difficulty. Figure 9 checks difficulty on all 140 primary queries. Table 10 adds alternative scores and 160 secondary policy/query pairs from the same 80 evidence queries under Qwen3 and gpt-oss. Audits were locked before the 64 follow-up responses. Maps, evaluators, failure and truncation handling remain fixed, with no selection by follow-up outcomes.

For query i, let $\widehat{m}_{i,c}(n)$ be the largest category frequency with label c among the first n follow-up responses in generation order. It is zero if no category with that label appears. Define

$$\widehat{g}_i(n) = \left|\widehat{m}_{i,1}(n) - \widehat{m}_{i,0}(n)\right|,$$
$$\widehat{p}_i(n) = \max\{\widehat{m}_{i,1}(n), \widehat{m}_{i,0}(n)\}, \qquad \widehat{q}_i(n) = \min\{\widehat{m}_{i,1}(n), \widehat{m}_{i,0}(n)\},$$
$$\widehat{I}_{i,\Lambda}(n) = \max_{\lambda \in \Lambda}\{\widehat{p}_i(n)\log(1+\lambda) + \widehat{q}_i(n)\log(1-\lambda)\}, \qquad \Lambda = \{2^{-1}, \ldots, 2^{-5}\}.$$

Scores use only follow-up prefixes $n \in \{32, 64\}$, retaining two nonpositive information scores at 32. All at 64 are positive. For confirmed direction times $\rho_{i,c}$, infinite for no report, cost is $C_i = \min\{\rho_{i,0}, \rho_{i,1}, 160\}$. The eight primary and seven secondary nonreturns pay 160. Follow-up responses never enter the audit.

Within-stratum Spearman correlations average tied ranks. The resolved-only check excludes nonreturns and uses $\widehat{g}_i(64)$. Pooling transforms rank r in a stratum of size N to $(r-(N+1)/2)/(N+1)$,

then computes Pearson correlation across concatenated strata. Intervals take 2.5th and 97.5th percentiles of 10000 bootstrap resamples, recomputing ranks. Primary resampling draws queries within benchmarks, and secondary resampling keeps both policies together per query ID. Intervals concern these cohorts, not a benchmark population.

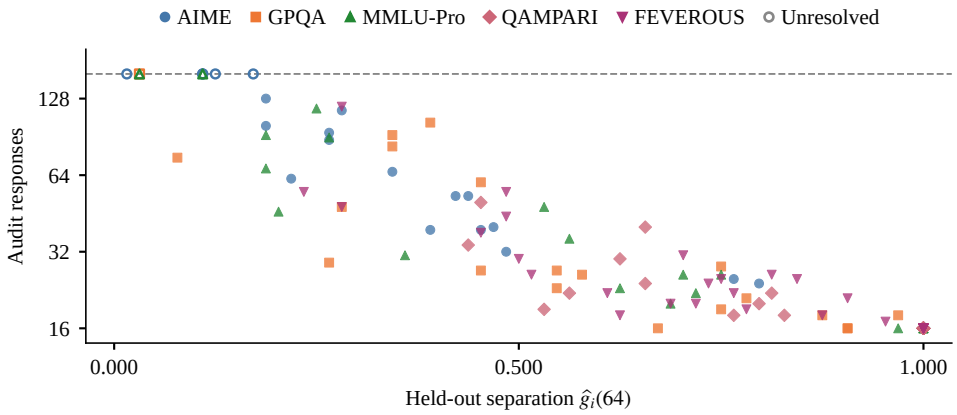


Figure 9: Audit cost and independent follow-up separation for all 140 primary queries. Hollow points remain unresolved at the response budget $B = 160$, shown by the dashed line. Larger empirical separation accompanies lower cost. Table 10 gives the rank associations and sensitivity checks.

Table 10's negative association holds in every stratum, with primary pooled value -0.946, across proxies, prefixes, resolved-only analysis and secondary policies. All 74 primary queries with $\widehat{g}_i(64) \geq 0.750$ report within 28 responses, while the eight nonreturns have gaps 0.016–0.172. These are descriptive associations, not causal effects or fitted predictions. Neither proxy identifies signed Γ, and truncated costs do not verify the expected stopping bound or its leading constant.

Table 10: Correlations with audit cost capped at $B = 160$. Pooled intervals are pointwise 95% query-bootstrap intervals. Resampling is stratified by benchmark and keeps both secondary policies together.

Policy and benchmark	Gap 64	Information 64	Gap 32	Resolved
Qwen2.5, AIME 2026	−0.951	−0.959	−0.930	−0.901
Qwen2.5, GPQA-Diamond	−0.867	−0.866	−0.816	−0.844
Qwen2.5, MMLU-Pro	−0.946	−0.944	−0.903	−0.927
Qwen2.5, QAMPARI-based QA	−0.986	−0.989	−0.990	−0.986
Qwen2.5, FEVEROUS-based verification	−0.956	−0.953	−0.936	−0.956
Qwen3, QAMPARI-based QA	−0.901	−0.895	−0.901	−0.870
Qwen3, FEVEROUS-based verification	−0.879	−0.872	−0.932	−0.866
gpt-oss, QAMPARI-based QA	−0.851	−0.848	−0.844	−0.824
gpt-oss, FEVEROUS-based verification	−0.944	−0.946	−0.923	−0.937
Pooled analysis	**Records**	**Correlation**	**Bootstrap interval**	
Primary, gap 64	140	−0.946	[−0.962, −0.894]	
Primary, information 64	140	−0.946	[−0.963, −0.894]	
Primary, gap 32	140	−0.922	[−0.949, −0.854]	
Primary, resolved only	132	−0.933	[−0.954, −0.872]	
Secondary, gap 64	160	−0.894	[−0.936, −0.829]	

B.3 Label versus Identity

Matched Controls behind Table 1. Under the matched conditions in Section 5.3, the nested opponent sets make Fixed-label cross no later (Proposition C.5). All RESOLVE and fixed-control streams return correctly. At 0.050, the Fixed-mode/Fixed-label ratio is 1.295 ± 0.008 for $q = s$ and 5.053 ± 0.061 for $q < s$, with paired-stream delta-method SEs. These ratios measure the added identity cost, not asymptotic constants or RESOLVE/full-CITE speedups.

Tightening the Error Level at Fixed Distribution. Table 11 extends Table 1 to six levels on 5000 paired streams per setting. Distribution, target, grid, weights and $B = 8192$ stay fixed. The prespecified target or direction is aligned with the alternative for cost measurement. RESOLVE and the controls use α, while full CITE splits it among three components.

Table 11: Complete error-level sweep behind Table 1. Entries are RMST $\pm$ one Monte Carlo SE over 5000 paired streams per distribution, including nonreturns at response budget $B = 8192$.

		(a) Matched controls		(b) Full procedures	
Regime	α	Fixed-label	Fixed-mode	RESOLVE	Oracle-target CITE
$q = s$	0.100	151.2 ± 1.6	199.0 ± 1.8	205.1 ± 2.0	325.8 ± 2.2
	0.050	188.9 ± 1.8	244.6 ± 2.0	243.9 ± 2.1	365.9 ± 2.4
	0.020	237.5 ± 2.1	300.3 ± 2.2	294.5 ± 2.4	418.3 ± 2.6
	0.010	273.6 ± 2.2	341.1 ± 2.3	331.9 ± 2.5	458.7 ± 2.7
	0.005	312.3 ± 2.4	381.1 ± 2.4	367.5 ± 2.6	495.2 ± 2.8
	0.001	395.8 ± 2.7	471.1 ± 2.7	447.2 ± 2.9	583.4 ± 3.0
$q < s$	0.100	148.8 ± 1.6	749.6 ± 8.5	192.6 ± 1.9	1318.7 ± 11.5
	0.050	187.4 ± 1.8	946.8 ± 9.7	236.4 ± 2.1	1519.0 ± 12.4
	0.020	237.5 ± 2.0	1209.1 ± 11.0	289.4 ± 2.3	1783.1 ± 13.6
	0.010	273.5 ± 2.2	1402.5 ± 11.9	326.5 ± 2.5	1967.1 ± 14.3
	0.005	310.3 ± 2.4	1598.3 ± 12.7	365.2 ± 2.6	2153.5 ± 15.0
	0.001	394.0 ± 2.7	2032.8 ± 14.5	452.4 ± 2.9	2580.4 ± 16.6

Tighter levels raise every cost in Table 11, but not necessarily CITE's ratio to RESOLVE. Returns are exactly confirmed. CITE's one nonreturn among 5000 streams at $q < s$ and $\alpha = 0.001$ pays the full budget. These finite-grid, censored costs identify neither an untruncated expectation nor Theorem 4.3's ideal constant.

Natural Same-Label Competition. The real-response comparison retains all 198 archived Qwen3 GPQA-Diamond and 200 MMLU-Pro queries. The first 128 or 64 responses are monitored. The next 64 supply a fixed reference plurality, with first-appearance ties, but never enter monitored counts. Fixed-label and Fixed-mode share this candidate, the online block, 20 paired permutations, counts from response one, bets $\{2^{-1}, \ldots, 2^{-5}\}$ with weights $1/5$, directional level 0.025, exact crossing threshold 40 and budget. Only opponents differ. An empty opponent set gives maximum count zero, not an automatic report. Neither control is full RESOLVE or CITE.

Figure 10(a) describes the reference blocks. Label separation subtracts the largest opposite-label frequency from the candidate frequency. Identity separation instead subtracts the largest other-category frequency. Group A has a unique reference plurality and a positive difference between these separations, indicating a closer same-label rival. Group B has a unique plurality and zero difference. Group C retains reference plurality ties before tie breaking.

Table 12 reports online RMST, charging nonreturns the full budget and excluding the common 64 reference responses. Each query averages its 20 orders equally. Report rates count threshold crossings, not verified correct population certificates. Pointwise 95% percentile intervals use 10000 paired query-bootstrap samples, carrying both controls and all orders together. Pooled resampling stays within benchmark and gives queries equal weight. These intervals describe variation across the selected archived queries, not coverage for a benchmark population.

Table 12: Matched real-response costs and report rates. A marks closer same-label reference rivals, B no strict excess and C reference ties. Brackets are paired query-bootstrap intervals for extra mode RMST. Single-query ties have no informative bootstrap interval.

			RMST			Report %	
Task	Group	N	Label	Mode	Extra mode [95%]	Label	Mode
GPQA	A	30	37.1	72.9	35.835 [23.688, 48.587]	93.3	63.0
	B	167	29.0	29.2	0.213 [0.082, 0.378]	93.1	93.1
	C	1	128.0	128.0	0.000	0.0	0.0
MMLU-Pro	A	16	23.4	37.4	13.994 [7.156, 21.678]	93.8	62.5
	B	183	19.6	19.8	0.201 [0.010, 0.492]	91.6	91.1
	C	1	32.2	32.2	0.000	100.0	100.0
Pooled	A	46	32.3	60.6	28.238 [20.043, 36.998]	93.5	62.8
	B	350	24.1	24.3	0.207 [0.077, 0.377]	92.3	92.1

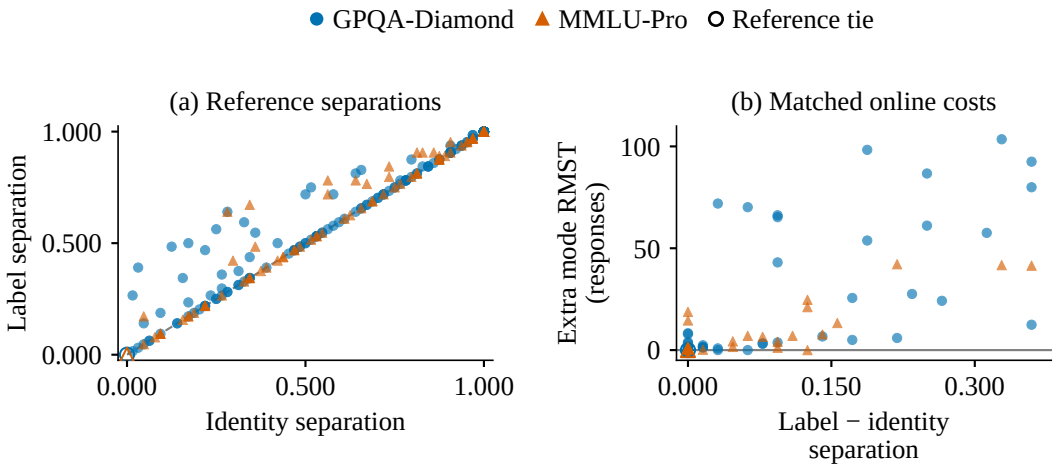


Figure 10: Reference competition and matched cost on all 398 archived queries. (a) Label versus identity separation. (b) Fixed-mode minus Fixed-label truncated RMST against the difference between these separations. Each point averages 20 paired orders. Shapes distinguish benchmarks and hollow markers retain the two reference ties. Coincident points overlap without jitter. Both controls exclude the common reference cost.

The extra cost in Table 12 is concentrated in Group A, where Fixed-mode also reports less often. Figure 10(b) shows that this association is not deterministic. Spearman correlations are 0.788 on GPQA and 0.811 on MMLU-Pro. Correlating ranks centered and scaled within benchmark gives 0.796 with paired query-bootstrap interval $[0.709, 0.869]$, retaining all 398 queries. All 46 A candidates are FAIL because these tasks have only one PASS category. Group B includes 286 PASS candidates with identical opponent sets. Restricting B to its 64 FAIL candidates still gives extra mode cost 1.132 with interval $[0.449, 2.024]$, versus 28.238 in A. Original-order differences are 30.435 and 0.129 for A and B. At directional level 0.050, they are 28.505 and 0.208. Both reference ties have zero difference, too few for a general conclusion about ties.

Fixed-mode's opponent set contains Fixed-label's, preventing earlier crossing on the same path. The finding is the size of its extra truncated cost and association with reference competition. A reference winner need not be the population mode, and equal reference gaps need not imply equal online opponent maxima. Archived maps lack demonstrated independent prespecification, budgets differ and nonreturns are censored. This comparison establishes neither Corollary 4.4 nor a causal effect or universal method ranking.

B.4 Alternative Procedures and Their Information

These comparisons separate the cost of the target from the value of supplied categories, candidates and error allocations.

Paired Complete-Block Costs. For the comparisons in Section 5.4, the 198 GPQA and 200 MMLU-Pro archives have 128 and 64 online responses, then 64 disjoint reference responses for CITE. Methods use original orders and 20 paired seeded permutations without query selection or retuning. Table 2's label methods need no reference block. CITE charges it and the full budget for nonreturns, including possibly uncertifiable targets. These retrospective costs cannot identify population labels or give a general ranking across the different targets and information in Appendix A.4.

The independent full-CITE implementation has online RMST 43.97 ± 0.15 on GPQA and 27.61 ± 0.05 on MMLU-Pro, with report rates $(86.16 \pm 0.14)\%$ and $(87.50 \pm 0.10)\%$. Uncertainty is one ordering SE over 20 permutations, not Table 12's query-bootstrap uncertainty. Reference blocks bring total RMST to 107.97 ± 0.15 and 91.61 ± 0.05. Closer same-label reference rivals accompany higher CITE costs and fewer returns. Unlike Figure 10's matched controls, full CITE also differs from RESOLVE in discovery and certification checks.

A Shared Query-Level Multinomial Region. Table 13 replaces the 0.025 directional split with one 0.050 Multinomial-CS region protecting both labels. The same 20 orders are replayed with rational confirmation. The default replay reproduces all 8358 records, including original orders.

Table 13: Multinomial-CS allocation sensitivity on the complete archived blocks. Entries give one ordering SE. The shared region retains query-level 0.050 protection under the assumptions for a known category set, not a restored prospective guarantee for these archives.

	GPQA-Diamond		MMLU-Pro	
Allocation	RMST	Report %	RMST	Report %
0.025 per direction	48.01 ± 0.16	84.62 ± 0.15	45.64 ± 0.02	75.68 ± 0.06
One 0.050 region	45.51 ± 0.17	85.73 ± 0.15	44.79 ± 0.03	75.90 ± 0.09

The shared region improves both metrics in Table 13, but still costs more and returns less often than RESOLVE here. Under the known-category assumptions, either false direction excludes the true vector from the shared region, avoiding a second union bound.

MMC Replay. MMC (Cordero-Encinar and Duncan, 2025), specified in Appendix A.4, uses the same blocks and orders, budgets 128 and 64 and our nominal $\delta = 0.050$. Categories are supplied, labels withheld until return. Nonreturns pay the budget, with original orders kept separate.

Table 14: MMC stopping behavior on archived blocks. Costs and percentages have one ordering SE. Exclusive returns and joint nonreturns are percentages over the same paired orders as RESOLVE. The two methods have different targets and guarantee interpretations.

Task	MMC RMST	MMC return %	Only MMC %	Only RESOLVE %	Neither %
GPQA	35.95 ± 0.18	85.66 ± 0.15	0.30 ± 0.08	6.94 ± 0.14	7.40 ± 0.14
MMLU-Pro	19.09 ± 0.07	88.58 ± 0.11	0.45 ± 0.09	2.67 ± 0.09	8.75 ± 0.12

MMC uses similar responses on GPQA and fewer on MMLU-Pro, but returns less often on both (Table 14). A direct replay from the pinned released source reproduced all stored MMC results. MMC's channel theorem constrains predictably selected contenders at every round (Cordero-Encinar and Duncan, 2025, Theorem 3.1). Given the different target and return convention, we report stopping behavior, not verified correct certificates or a matched 5% identity or RESOLVE label guarantee. CITE remains the fixed-target identity comparator.

Support-Oracle Diagnostic. CS methods receive every realized category and label from 80 Qwen3 follow-up blocks generated after diagnosis lock. RESOLVE discovers them online. Each

block has 64 responses, with $B = 64$, 20 paired orders and 0.025 per direction. Original orders remain separate. On each task, $34/40$ blocks have one label. Known absence of opponents lets Mixture-CS and Multinomial-CS return at response one. Outcome-CS needs a positive lower endpoint. Table 15 isolates the six blocks per task with both labels.

Table 15: Retrospective support-oracle diagnostic on Qwen3 evidence blocks. Costs and report percentages are means ± one ordering SE over 20 paired orders per query.

		All 40 blocks		Both labels, 6 blocks	
Task	Method	RMST	Report %	RMST	Report %
Entity QA	RESOLVE	21.44 ± 0.10	93.00 ± 0.36	48.23 ± 0.65	53.33 ± 2.42
	Mixture-CS	8.19 ± 0.09	90.25 ± 0.17	48.90 ± 0.61	35.00 ± 1.15
	Outcome-CS	12.66 ± 0.08	90.00 ± 0.00	58.25 ± 0.48	33.33 ± 0.00
	Multinomial-CS	7.95 ± 0.12	90.63 ± 0.26	47.32 ± 0.79	37.50 ± 1.74
Verification	RESOLVE	20.81 ± 0.05	92.63 ± 0.13	41.63 ± 0.21	50.83 ± 0.83
	Mixture-CS	6.73 ± 0.05	92.50 ± 0.00	39.18 ± 0.33	50.00 ± 0.00
	Outcome-CS	11.79 ± 0.07	92.50 ± 0.00	49.59 ± 0.37	50.00 ± 0.00
	Multinomial-CS	6.76 ± 0.06	92.63 ± 0.13	39.42 ± 0.43	50.83 ± 0.83

The overall CS cost advantage mostly disappears with both labels present, and report rates fall (Table 15). This small subset does not show faster learning of unknown support. Recomputed maps, labels and FAIL reasons match records. Independent prefix checks cover RESOLVE and Multinomial-CS, rational checks Mixture-CS, and higher-precision inequalities Outcome-CS. None restores future-support prespecification or prospective validity. Older archives with different maps are not pooled.

B.5 Sensitivity and Scope

B.5.1 Finite Implementation Sensitivity

Table 16 varies the cap or uses seven equal-weight bets $\{2^{-1}, \ldots, 2^{-7}\}$, reducing each original bet's weight from $1/5$ to $1/7$. Levels and maps stay fixed. The 132 resolved prefixes stop at their frozen reports, censoring variants that need more responses. Only the eight original nonreturns reach 160.

Table 16: Finite implementation sensitivity on 140 stored audit prefixes. Time and direction changes count returned reports only. Prefix censoring counts previously resolved queries without a return by the end of their stored prefix. All eight originally unresolved queries remain unresolved through 160 in these replays.

K_{cap} / equal-weight bets	Returns	Time changes	Direction changes	Prefix censored
$10/5$	131	0	0	1
$25/5$	132	0	0	0
$50/5$	132	0	0	0
$100/5$	132	0	0	0
Uncapped / 5	132	0	0	0
$50/7$	76	0	0	56

Raising the cap changes no observed report. The $K_{\text{cap}} = 10$ variant loses one return within its stored prefix, while the seven-bet variant loses many through both grid expansion and weight redistribution. No variant returns earlier, so every prefix-truncated mean is 39.1. This is not RMST through 160 or evidence of nonresolution beyond the stored prefixes. No continuation is imputed.

B.5.2 Response Representation Defines the Target

On MMLU-Pro, replaying the first 64 responses at 0.025 per direction changes $75/200$ diagnoses, including 21 direct favorable/unfavorable reversals. The alternative parser rereads a boxed option, then an explicit answer phrase, then a bold or standalone letter, instead of lightly normalizing stored

extracted strings. Keys and evaluators stay fixed. Merging formats and recovering missed extractions explains how the target changes, not why this post hoc parser should be preferred.

As a negative control, replacing wrapper/whitespace normalization with complete-pool grouping by the first symbolically equivalent representative (Kydlíček, 2026) changes no diagnoses among 93 mathematics queries, comprising 30 AIME 2026 queries each for Qwen3 and gpt-oss and 33 AIME/MATH-500 queries for Qwen2.5-Math. Dropping failed extractions changes $9/30$ gpt-oss diagnoses, with failures accounting for 21.5% of responses. None replaces the primary map. A new representation defines a new audit target and requires its own fixed protocol.

B.5.3 Audit, Policy Revision, and Re-Audit

After an unfavorable audit, a revised policy defines a new P requiring fresh evidence. The frozen FEVEROUS metadata order supplied 24 development and 40 confirmatory queries, disjoint from each other and the original 40-query cohort. Model, retrieval, category map and evaluator stay fixed. The selected suffix asks for "the smallest set of evidence IDs that is by itself enough to decide the claim" in the original output format.

Development used 64 responses per query under each policy. The candidates requested a minimal sufficient set, added a check for unnecessary IDs, or further preferred the earliest IDs among equally small sufficient sets. The frozen rule maximized, in order, the count of negative-to-positive gap changes from baseline, median paired gap change, and the fraction of positive gaps. Gap is largest PASS minus largest FAIL category frequency in each development block. Ties favored fewer suffix characters, then the candidate order just given. The selected revision was frozen before confirmatory generation. Development showed no clear systematic improvement. Median gap change was 0.0000, with 10 improving and 8 worsening queries.

Both confirmatory policies use the Qwen2.5 verification settings in Appendix A.2, including the 256-token cap, and Algorithm 1 with $B = 160$ and 0.025 per direction. They start separately at response one. Only the suffix changes. Retained validation records confirm cohort separation, freeze chronology, evaluator consistency and prefix replay. These are local records, not external preregistration.

Table 17: Paired audit reports on 40 new FEVEROUS queries. Rows give the baseline report and columns the report under the frozen prompt modification. All queries are retained. Counts describe reports rather than known population labels.

	Modified policy $P^{(1)}$			
Baseline $P^{(0)}$	Favorable	Unfavorable	Unresolved	Total
Favorable	10	0	1	11
Unfavorable	2	24	3	29

In Table 17, most unfavorable reports persist, two become favorable and three become unresolved. The reversals illustrate a different diagnosis after changing the policy, but the mixed result does not establish reliable repair. No favorable-to-unfavorable report is observed, not zero risk. The two unfavorable-to-favorable cases use 76 and 93 responses across both audits, establishing no compute, token or time savings. Each query/policy receives level 0.050, not joint protection over the 40 queries or both policies. Report fractions do not give error-controlled estimates of population sign changes.

After audit lock, 64 fresh responses per policy/query show mean largest FAIL frequency falling from 0.624 to 0.507. Mean total PASS frequency rises by 0.066, but its largest category rises by only 0.005. Queries with at least two observed PASS categories rise from 0 to 9. This descriptive change concerns mainly the FAIL leader, not the intended PASS concentration.

Two amendments affected no confirmatory audit. A development launch failed before storing responses and restarted from zero. A follow-up seed above $2^{32} - 1$ was corrected after diagnosis lock. The follow-up explains the observed frequency changes without establishing a population mechanism.

C PROOFS

We prove the population target, anytime validity and response-cost results in order, then justify the comparison procedures in Tables 1 and 2.

We retain the notation of Sections 2 to 4. Unless stated otherwise, probabilities and expectations are under $\mathbb{P} = P^{\otimes\mathbb{N}}$. The history is $\mathcal{F}_t = \sigma(X_1, \ldots, X_t)$, with trivial $\mathcal{F}_0$ and $\mathcal{F}_\infty = \sigma(X_1, X_2, \ldots)$. Counts $N_t(a)$ start at response one, while $C_t(z), M_t(z)$ start strictly after discovery. Bars mark capped evidence and stored weights. Appendix C.3 defines local time after discovery.

C.1 POPULATION TARGET

A countable category space requires two facts. Positive maxima are attained, so winning categories and null opponents exist. Uniform frequency convergence then controls the empirical winner even as its identity changes.

Lemma C.1. *For any $S \subseteq \mathcal{A}$ with $\sup_{a\in S} p_a > 0$, some category attains that supremum. With $\widehat{p}_n(a) = n^{-1}\sum_{t=1}^n \mathbb{1}\{X_t = a\}$,*

$$\sum_{a\in\mathcal{A}} |\widehat{p}_n(a) - p_a| \to 0, \qquad \sup_{a\in\mathcal{A}} |\widehat{p}_n(a) - p_a| \to 0 \quad \textit{almost surely as } n \to \infty.$$

Proof. We show attainment first, then uniform convergence. Let $M = \sup_{a\in S} p_a > 0$ and suppose it is not attained. There is at least one mass in S above $M/2$. If there were only finitely many, their maximum would be strictly below M. All remaining masses are at most $M/2$, so that finite maximum would be the supremum, a contradiction. Thus there are arbitrarily many distinct $a_1, \ldots, a_K \in S$ with $p_{a_k} > M/2$, so

$$1 \geq \sum_{k=1}^K p_{a_k} > \frac{KM}{2}.$$

This is impossible for $K > 2/M$.

For convergence, write $\widehat{P}_n(B) = \sum_{a\in B} \widehat{p}_n(a)$ for the empirical probability of $B \subseteq \mathcal{A}$. Apply the strong law to each category indicator and intersect the countably many probability-one events. Fix $\varepsilon > 0$ and choose finite $F \subseteq \mathcal{A}$ with $\mathbb{P}(X_1 \notin F) < \varepsilon$. Eventually $\sum_{a\in F} |\widehat{p}_n(a) - p_a| < \varepsilon$. Conservation of mass gives $\widehat{P}_n(\mathcal{A} \setminus F) < 2\varepsilon$, whence

$$\sum_{a\in\mathcal{A}} |\widehat{p}_n(a) - p_a| \leq \sum_{a\in F} |\widehat{p}_n(a) - p_a| + \widehat{P}_n(\mathcal{A} \setminus F) + \mathbb{P}(X_1 \notin F) < 4\varepsilon.$$

Let $\varepsilon \downarrow 0$. The supremum is bounded by this sum. This is the countable discrete form of Scheffé's theorem (Scheffé, 1947). □

Uniform convergence now transfers the strict population gap to the empirical maxima.

Proof of Proposition 2.1. It suffices to show that the empirical label gap eventually has the population gap's sign. For $c \in \{0, 1\}$ define $\widehat{m}_{c,n} := \sup_{a\in\mathcal{A}_c} \widehat{p}_n(a)$, with value zero for an empty label class. The deterministic supremum inequality and Lemma C.1 give

$$|\widehat{m}_{c,n} - m_c| \leq \sup_{a\in\mathcal{A}} |\widehat{p}_n(a) - p_a| \to 0 \quad \text{almost surely as } n \to \infty.$$

If $\Gamma > 0$, then almost surely $\widehat{m}_{1,n} > \widehat{m}_{0,n}$ for every sufficiently large n. Every empirical plurality therefore has label one, regardless of ties among label-one categories. Exchanging the labels handles $\Gamma < 0$. Empty label classes are included through the zero convention. The argument does not cover $\Gamma = 0$, and no convergence claim is made there. □

The same argument explains the availability distinction. If $\pi_1 > 0$, the probability of seeing no pass in n draws is $(1 - \pi_1)^n \to 0$ as $n \to \infty$. When $\Gamma < 0$, Proposition 2.1 and bounded convergence also give $\mathbb{P}(h(\widehat{a}_n) = 1) \to 0$. A union bound therefore shows that finding a pass while plurality fails has probability tending to one. Figure 2 examines this distinction at finite budgets.

C.2 Anytime Validity

The validity argument must handle both the discovered candidate and the changing observed opponent. We bound each account by a fixed-opponent supermartingale under the null, then mix these bounds.

Fix c and $z \in \mathcal{A}_c$. For an opponent $b \in \mathcal{A}_{1-c}$, write $Z_k^{(z,b)} = \mathbb{1}\{X_k = z\} - \mathbb{1}\{X_k = b\}$. The pairwise evidence behind (1), started after discovery, is

$$E_t^{(z,b)}(\lambda) := \prod_{k=T(z)+1}^{t} \left(1 + \lambda Z_k^{(z,b)}\right)$$
$$= (1+\lambda)^{N_t(z) - N_{T(z)}(z)} (1-\lambda)^{N_t(b) - N_{T(z)}(b)}, \qquad t \geq T(z).$$

This process stays at one through discovery. For a nonempty opposite class, $G_t(z;\lambda) = \inf_{b \in \mathcal{A}_{1-c}} E_t^{(z,b)}(\lambda)$ because $1 - \lambda < 1$. For an empty class the convention remains $G_t(z;\lambda) = (1+\lambda)^{C_t(z)}$.

The strongest observed opponent can change with the data. We therefore bound the account using one opponent fixed under each null, rather than claim that the account itself is a supermartingale.

Candidate Discovery. Let $D_c(t)$ count distinct categories discovered in $\mathcal{A}_c$ and $K_c(t) = \min\{D_c(t), K_{\text{cap}}\}$ count admitted candidates. Recall $T_{c,j} = \inf\{t \geq 1 : D_c(t) \geq j\}$, with infinity for no discovery. Write $\mathcal{F}_T$ for the history at a stopping time T. For $H \in \mathcal{F}_T$, we use $H \in \mathcal{F}_\infty$ and $H \cap \{T = k\} \in \mathcal{F}_k$ at every finite k.

Set $R_{c,j} = X_{T_{c,j}}$ on finite discovery and a cemetery symbol otherwise. Then

$$\{T_{c,j} \leq t\} = \{D_c(t) \geq j\} \in \mathcal{F}_t,$$
$$\{R_{c,j} = a\} \cap \{T_{c,j} \leq t\} = \bigcup_{k=1}^{t} \{T_{c,j} = k, X_k = a\} \in \mathcal{F}_t.$$

The cemetery event is $\{T_{c,j} = \infty\} \in \mathcal{F}_\infty$, with empty intersection in the second condition. Thus discovery is a stopping time and the complete identity is $\mathcal{F}_{T_{c,j}}$-measurable. In particular, on $\{T_{c,j} < k\}$ it is known before X_k. Post-discovery counts are adapted, and their countable maximum is measurable. Every account is bounded by 2^t, so the normalized evidence mixture is finite, adapted, and has a stopping-time first crossing. Undiscovered accounts remain one.

The next lemma supplies a supermartingale that bounds each discovered account under the null.

Lemma C.2. *Fix a distribution P satisfying $H_{0,c}$, a discovery position j, and $\lambda \in (0,1)$. There is a P-dependent nonnegative supermartingale $W_{j,t}^{c,P}(\lambda)$ with $W_{j,0}^{c,P}(\lambda) = 1$ such that, $\mathbb{P}$-almost surely,*

$$G_t(R_{c,j};\lambda) \leq W_{j,t}^{c,P}(\lambda) \qquad \textit{for every } t.$$

Proof. A valid account against a fixed opponent. If $m_c = 0$, countability implies that no label-c category is ever observed almost surely. The candidate is never discovered, and $G_t(R_{c,j};\lambda) = W_{j,t}^{c,P}(\lambda) = 1$ provides the dominator for every t.

Now suppose $m_c > 0$. Under $H_{0,c}$, $m_{1-c} \geq m_c > 0$. Lemma C.1 supplies a fixed $b \in \mathcal{A}_{1-c}$ with $p_b = m_{1-c}$. Choose a fixed $a_0 \in \mathcal{A}_c$ as a placeholder before discovery and define the predictable candidate

$$z_{j,k-1} := \begin{cases} R_{c,j}, & T_{c,j} < k, \\ a_0, & T_{c,j} \geq k. \end{cases}$$

The event $\{T_{c,j} < k\}$ and the discovered identity on that event are determined by $\mathcal{F}_{k-1}$. Hence $z_{j,k-1}$ is $\mathcal{F}_{k-1}$-measurable and satisfies $p_{z_{j,k-1}} \leq m_c \leq p_b$. Since X_k is independent of $\mathcal{F}_{k-1}$ and the candidate takes values in the countable set $\mathcal{A}$,

$$\mathbb{P}(X_k = z_{j,k-1} \mid \mathcal{F}_{k-1}) = p_{z_{j,k-1}}.$$

To avoid using the discovery response as evidence for its own category, start the wealth only on subsequent responses,

$$W_{j,t}^{c,P}(\lambda) := \prod_{k=1}^{t} \Big[1 + \lambda \mathbb{1}\{T_{c,j} < k\}\big(\mathbb{1}\{X_k = z_{j,k-1}\} - \mathbb{1}\{X_k = b\}\big)\Big].$$

The two categories in the difference have opposite labels and are therefore distinct. Each multiplier consequently equals $1 + \lambda$, $1 - \lambda$, or 1, and is nonnegative. Its conditional mean given $\mathcal{F}_{k-1}$ is

$$1 + \lambda \mathbb{1}\{T_{c,j} < k\}\big(p_{z_{j,k-1}} - p_b\big) \le 1.$$

Each product is bounded by 2^t, so it is integrable. Multiplying the conditional inequality by the nonnegative previous wealth proves the supermartingale property.

Domination of the discovered account. Both processes remain one through discovery because $\mathbb{1}\{T_{c,j} < k\} = 0$ at $k = T_{c,j}$. After finite discovery, the new wealth equals $E_t^{(R_{c,j},b)}(\lambda)$. The observed maximum opposing count is at least this fixed opponent's count, giving

$$\begin{aligned} G_t(R_{c,j};\lambda) &= (1+\lambda)^{C_t(R_{c,j})}(1-\lambda)^{M_t(R_{c,j})} \\ &\le (1+\lambda)^{C_t(R_{c,j})}(1-\lambda)^{N_t(b)-N_{T_{c,j}}(b)} = E_t^{(R_{c,j},b)}(\lambda) = W_{j,t}^{c,P}(\lambda). \end{aligned}$$

The inequality uses $M_t(R_{c,j}) \ge N_t(b) - N_{T_{c,j}}(b)$ and $0 < 1 - \lambda < 1$. A larger opponent count makes the evidence smaller, proving domination for every t. □

Using the same weights for the accounts and their bounds transfers domination to the capped monitor.

Proof of Theorem 4.1. Fix c and a distribution P satisfying $H_{0,c}$. By countability, the componentwise inequalities in Lemma C.2 hold simultaneously. Keep the unadmitted discovery mass at one and mix the dominators using the capped monitor's weights,

$$\bar{W}_t^{c,P} = 1 - \sum_{j=1}^{K_{\mathrm{cap}}} \bar{w}_j + \sum_{j=1}^{K_{\mathrm{cap}}} \bar{w}_j \sum_{\lambda\in\Lambda} v_\lambda W_{j,t}^{c,P}(\lambda).$$

The constant term is nonnegative and $\bar{W}_0^{c,P} = 1$. Fixed nonnegative mixing preserves the supermartingale property. An undiscovered index contributes one here and to the unused mass in (2). Lemma C.2 bounds each discovered account, so $\bar{E}_t^c \le \bar{W}_t^{c,P}$ for all t almost surely. Ville's inequality (Ville, 1939) applies to this nonnegative supermartingale starting at one, giving

$$\mathbb{P}\big(\bar{\tau}_c(\alpha_c) < \infty\big) \le \mathbb{P}\Big(\sup_{t\ge 0} \bar{W}_t^{c,P} \ge \frac{1}{\alpha_c}\Big) \le \alpha_c.$$

The monitor and level are the same for every distribution satisfying $H_{0,c}$ and the setup, proving uniformity without comparing capped and ideal evidence pathwise.

The ideal process uses the corresponding mixture

$$W_t^{c,P} := \sum_{j\ge 1} w_j \sum_{\lambda\in\Lambda} v_\lambda W_{j,t}^{c,P}(\lambda).$$

Each component is nonnegative and at most 2^t. The normalized mixture is therefore finite and integrable at every fixed time. Conditional Tonelli's theorem permits termwise conditioning even for countable mixtures, giving $\mathbb{E}[W_t^{c,P} \mid \mathcal{F}_{t-1}] \le W_{t-1}^{c,P}$. Normalization gives $W_0^{c,P} = 1$, and the same weights preserve domination, so $E_t^c \le W_t^{c,P}$. Ville's inequality also yields

$$\mathbb{P}\Big(\sup_{t\ge 0} E_t^c \ge \frac{1}{\alpha}\Big) \le \mathbb{P}\Big(\sup_{t\ge 0} W_t^{c,P} \ge \frac{1}{\alpha}\Big) \le \alpha.$$

□

Proof of Corollary 4.2. Fix any distribution satisfying the setup. Exact confirmation gives $\{\rho_c < \infty\} \subseteq \{\bar{\tau}_c(\alpha_c) < \infty\}$. Since Γ is nonrandom, the error event in (3) satisfies

$$\begin{aligned}\mathbb{P}(\mathcal{E}_{\mathrm{err}}) &\le \mathbb{1}\{\Gamma \le 0\}\mathbb{P}(\rho_1 < \infty) + \mathbb{1}\{\Gamma \ge 0\}\mathbb{P}(\rho_0 < \infty) \\ &\le \mathbb{1}\{\Gamma \le 0\}\mathbb{P}\big(\bar{\tau}_1(\alpha_1) < \infty\big) + \mathbb{1}\{\Gamma \ge 0\}\mathbb{P}\big(\bar{\tau}_0(\alpha_0) < \infty\big) \\ &\le \alpha_1 + \alpha_0 \le \alpha_{\mathrm{query}}.\end{aligned}$$

Every nonzero term corresponds to a true directional null, so Theorem 4.1 supplies its bound. No independence between directions is needed. At $\Gamma = 0$ both directions are included. Otherwise only the losing direction is included. The fixed simultaneous-crossing rule and exact numerical confirmation cannot add a report outside this event. Nothing conditions on returning or retaining a maximizer. □

C.3 Evidence Growth and Response Cost

Almost-sure evidence growth does not by itself bound expected delay. After finding the growth rate, we control rare long waits after discovery and average over discovery time and rank.

For this subsection, fix a dominant direction c and a maximizing category $a^\star \in \mathcal{A}_c$ supplied by Lemma C.1, with $p = p_{a^\star} = m_c > q = m_{1-c}$. Its discovery time and rank, and the subsequent counts, are

$$\begin{aligned}T_\star &:= \inf\{t \ge 1 : X_t = a^\star\}, & J_\star &:= D_c(T_\star), \\ C_n^\star &:= C_{T_\star+n}(a^\star), & M_n^\star &:= M_{T_\star+n}(a^\star).\end{aligned}$$

Here t is the global response count, while $n = t - T_\star$ counts responses strictly after discovery. The discovery time is geometric with success probability p, so $J_\star \le T_\star < \infty$ almost surely and $\mathbb{E}[T_\star] = 1/p$.

We must justify using concentration after the random discovery time. For $H \in \mathcal{F}_{T_\star}$ and bounded measurable f of d responses, independence at each fixed time gives

$$\begin{aligned}\mathbb{E}\big[\mathbb{1}_H f(X_{T_\star+1}, \dots, X_{T_\star+d})\big] &= \sum_{k\ge1} \mathbb{E}\big[\mathbb{1}_{H\cap\{T_\star=k\}} f(X_{k+1}, \dots, X_{k+d})\big] \\ &= \mathbb{P}(H)\,\mathbb{E}[f(X_1, \dots, X_d)].\end{aligned}$$

Here $H \cap \{T_\star = k\} \in \mathcal{F}_k$, and boundedness justifies the sum. Thus future responses remain i.i.d. given $\mathcal{F}_{T_\star}$, with each block having distribution $P^{\otimes d}$.

For $0 < u < p$, the rate $\ell_\lambda(p, u)$ from Section 4.2 is strictly concave in λ. Its derivative and maximizing bet are

$$\frac{\partial \ell_\lambda(p,u)}{\partial \lambda} = \frac{p}{1+\lambda} - \frac{u}{1-\lambda}, \qquad \lambda^\star = \frac{p-u}{p+u}.$$

Substitution gives the stated $\mathcal{I}(p, u)$. At $u = 0$, the rate increases to $p \log 2$ as $\lambda \uparrow 1$, agreeing with the continuous extension. This calculation identifies the growth of the account. It does not claim that the rate is an information lower bound for every label audit.

For fixed p and $0 < u < p$,

$$\frac{\partial}{\partial u}\mathcal{I}(p,u) = \log\frac{2u}{p+u} < 0.$$

Together with continuity at zero, this derivative makes $\mathcal{I}(p, u)$ strictly decreasing on $[0, p)$. A stronger competitor leaves less distinguishing information per response.

One maximizing account with positive weight is enough to transfer this rate to the ideal mixture.

Proposition C.3. *For the fixed P and label c above, with $p = m_c > q = m_{1-c}$, suppose the uncapped process has $w_j > 0$ for every j and $v_\lambda > 0$ for every $\lambda \in \Lambda$. Then*

$$\liminf_{t\to\infty} \frac{1}{t} \log E_t^c \ge \mathcal{I}_\Lambda(p, q) \qquad \textit{almost surely.}$$

For a dense grid, the right-hand side equals $\mathcal{I}(p, q)$.

Proof. We transfer one maximizing account's growth to the ideal mixture in (4). Conditional on discovery, future responses are i.i.d. as established in Appendix C.3. Lemma C.1 bounds the error of their largest opposing frequency by the uniform frequency error. Hence $C_n^\star/n \to p$ and $M_n^\star/n \to q$ almost surely as $n \to \infty$, including $q = 0$. For every fixed $\lambda \in \Lambda$,

$$\frac{\log G_{T_\star+n}(a^\star;\lambda)}{n} \to \ell_\lambda(p,q) \quad \text{almost surely as } n \to \infty.$$

The bound $E^c_{T_\star+n} \geq w_{J_\star} v_\lambda G_{T_\star+n}(a^\star;\lambda)$ transfers this rate to the mixture. Dividing by $t = T_\star + n$ removes the fixed discovery-time and log-weight terms. Countability of Λ makes the probability-one statements simultaneous, so taking their supremum proves the first claim. Density and continuity of $\ell_\lambda(p,q)$ give the full-grid rate. No uniqueness of $a^\star$ is required. □

To control rare, long waits, we separate finitely many opponents from the remaining categories, whose total mass is small. A union bound over all categories need not be finite.

Lemma C.4. *For $\varepsilon > 0$, choose deterministic finite $F_\varepsilon \subseteq \mathcal{A}_{1-c}$ with remaining mass below $\varepsilon/2$. Conditional on $\mathcal{F}_{T_\star}$, for every $n \geq 1$,*

$$\mathbb{P}\big(C_n^\star/n < p - \varepsilon \mid \mathcal{F}_{T_\star}\big) \leq e^{-2n\varepsilon^2},$$
$$\mathbb{P}\big(M_n^\star/n > q + \varepsilon \mid \mathcal{F}_{T_\star}\big) \leq |F_\varepsilon| e^{-2n\varepsilon^2} + e^{-n\varepsilon^2/2}.$$

If $q = 0$, then $M_n^\star = 0$ almost surely. The finite set depends only on fixed P and ε, not the discovery history.

Proof. Conditional on discovery, $C_n^\star$ sums n independent Bernoulli indicators with mean p. Hoeffding's inequality (Hoeffding, 1963) gives the lower-tail bound, also when the event is empty.

Countable summability supplies F_ε with remaining set $\mathcal{T}_\varepsilon := \mathcal{A}_{1-c} \setminus F_\varepsilon$ satisfying $\mathbb{P}(X_1 \in \mathcal{T}_\varepsilon) < \varepsilon/2$. For a set $B \subseteq \mathcal{A}$, let $N_n^\star(B) := \sum_{k=1}^n \mathbb{1}\{X_{T_\star+k} \in B\}$ count its occurrences after discovery. Then

$$\left\{\frac{M_n^\star}{n} > q + \varepsilon\right\} \subseteq \bigcup_{b \in F_\varepsilon} \left\{\frac{N_n^\star(\{b\})}{n} > q + \varepsilon\right\} \cup \left\{\frac{N_n^\star(\mathcal{T}_\varepsilon)}{n} > q + \varepsilon\right\}.$$

Outside F_ε, the tail count bounds each opposing count. Since $p_b \leq q$, conditional Hoeffding bounds each finite-set event by $e^{-2n\varepsilon^2}$. The binomial tail count has success probability below $\varepsilon/2$. Exceeding $q + \varepsilon$ requires a deviation greater than $\varepsilon/2$, giving the bound $e^{-n\varepsilon^2/2}$. The union bound proves the claim. These constants are not uniform over P. When $q = 0$, countability gives zero total opposing mass, so no opposing category appears almost surely. □

The summable deviation bounds now control expected delay through the tail-sum identity. Throughout, $\tau_c(\alpha)$ is a directional crossing on the infinite stream, not the reporting time of a two-direction procedure.

Proof of Theorem 4.3. We control the delay after discovery of one maximizing account. Sufficient candidate support and small opposing counts force a crossing. Exponential bounds on failures give a conditional expected delay. Averaging over discovery time and rank adds only a finite term at fixed P, which vanishes on the $\log(1/\alpha)$ scale.

Since $\mathcal{I}_\Lambda(p,q) > 0$, fix $\lambda \in \Lambda$ with $\ell := \ell_\lambda(p,q) > 0$. With positive grid weight v_λ, write

$$L := \log\frac{1+\lambda}{1-\lambda} > 0, \qquad b_\alpha := \log\frac{1}{\alpha w_{J_\star} v_\lambda}.$$

The account's log threshold b_α is finite and $\mathcal{F}_{T_\star}$-measurable. Fix $\xi > 0$ and set $\varepsilon_\xi := \xi\ell/\{(1+\xi)L\}$. Define the bad event that candidate support is too small or some opposing count too large,

$$\mathcal{B}_n := \{C_n^\star < n(p - \varepsilon_\xi)\} \cup \{M_n^\star > n(q + \varepsilon_\xi)\}.$$

Since $\log(1-\lambda) < 0$, outside $\mathcal{B}_n$ we have

$$\log E^c_{T_\star+n} \geq \log\{w_{J_\star} v_\lambda\} + n\ell - n\varepsilon_\xi L.$$

The last two terms equal $n\ell/(1+\xi)$. Thus outside $\mathcal{B}_n$ a crossing occurs whenever $n \geq (1+\xi)b_\alpha/\ell$.

Lemma C.4 and a union bound give, for every $n \geq 1$,

$$\mathbb{P}(\mathcal{B}_n \mid \mathcal{F}_{T_\star}) \leq (|F_{\varepsilon_\xi}| + 2)e^{-n\varepsilon_\xi^2/2} = C_0 e^{-c_0 n}.$$

Here $C_0 = |F_{\varepsilon_\xi}| + 2 < \infty$ and $c_0 = \varepsilon_\xi^2/2 > 0$ depend on fixed P, λ, and ξ, not on α or the discovery history. The same bound holds when $q = 0$.

The conditional response cutoff is $H_\alpha = \lceil (1+\xi)b_\alpha/\ell \rceil \geq 1$. For each integer n, the event $\{n \geq H_\alpha\}$ is $\mathcal{F}_{T_\star}$-measurable. On it, failure to cross by $T_\star + n$ requires $\mathcal{B}_n$. Hence

$$\mathbb{P}(\tau_c(\alpha) > T_\star + n \mid \mathcal{F}_{T_\star}) \leq C_0 e^{-c_0 n} \qquad \text{on } \{n \geq H_\alpha\}.$$

The conditional tail-sum identity for $(\tau_c(\alpha) - T_\star)_+$ has H_α terms bounded by one and a geometric remainder, giving

$$\begin{aligned}\mathbb{E}\big[(\tau_c(\alpha) - T_\star)_+ \mid \mathcal{F}_{T_\star}\big] &\leq H_\alpha + \sum_{n \geq H_\alpha} C_0 e^{-c_0 n} \\ &\leq \frac{(1+\xi)b_\alpha}{\ell} + C_\xi,\end{aligned}$$

where $C_\xi = 1 + C_0/(1 - e^{-c_0})$ is finite and independent of α and the discovery history. The positive part also covers paths on which another account crossed before $T_\star$.

Only the discovery weight in b_α is random. Since $w_j = 6/(\pi^2 j^2)$ and $J_\star \leq T_\star$ with $\mathbb{E}[T_\star] = 1/p$, concavity of the logarithm gives

$$\begin{aligned}\mathbb{E}\Big[\log \frac{1}{w_{J_\star}}\Big] &= \log \frac{\pi^2}{6} + 2\mathbb{E}[\log J_\star] \\ &\leq \log \frac{\pi^2}{6} + 2\log \mathbb{E}[T_\star] = \log \frac{\pi^2}{6} + 2\log \frac{1}{p} < \infty.\end{aligned}$$

Thus $\mathbb{E}[b_\alpha] = \log(1/\alpha) + O(1)$ as $\alpha \downarrow 0$. Using $\tau_c(\alpha) \leq T_\star + (\tau_c(\alpha) - T_\star)_+$ gives the finite bound $\mathbb{E}[\tau_c(\alpha)] \leq 1/p + (1+\xi)\mathbb{E}[b_\alpha]/\ell + C_\xi$. Dividing by $\log(1/\alpha)$ gives

$$\limsup_{\alpha \downarrow 0} \frac{\mathbb{E}[\tau_c(\alpha)]}{\log(1/\alpha)} \leq \frac{1+\xi}{\ell_\lambda(p,q)}.$$

Finally, send $\xi \downarrow 0$. The resulting bound holds for every grid point with positive rate, so take a sequence of grid points whose rates approach $\mathcal{I}_\Lambda(p,q) > 0$. This proves the stated upper bound. If Λ is dense, continuity in λ gives $\mathcal{I}_\Lambda(p,q) = \mathcal{I}(p,q)$. When $q = 0$, the full-grid supremum is approached as $\lambda \uparrow 1$, and density supplies such a sequence even though the endpoint itself is excluded. □

Theorem 4.3 concerns the uncapped process. A finite cap can omit every maximizing account, and a fixed finite grid can have no positive rate even when $p > q$. These limitations do not affect Theorem 4.1. We do not claim an unconditional response cost bound for the capped process or a bound on numerical confirmation delay. A directional crossing bound also does not bound the time of a correct final report.

C.4 Label versus Identity Information

We combine the ideal label upper bound with CITE's existing information lower bound for a fixed target. Here $q = m_{1-c}$ excludes same-label rivals, whereas s includes them. Corollary 4.4 assumes $p > s > q \geq 0$, so its positive runner-up mass meets the condition of the published bound.

Ota et al. (2026, Corollary C.11 and Lemma C.9) give $\mathbb{E}[\tau_{\mathrm{CITE}}(\alpha)] \geq \log(1/\alpha)/\mathcal{I}(p,s)$ for a fixed target with a positive runner-up. The finite bound assumes almost sure return and finite expectation under P. For our comparison, infinite expectation already satisfies the lower bound, and finite expectation implies almost sure return. We use the pointwise inequality, not its separate small-gap expansion or the paper's minimax theorem.

The supplied target and certification rule must stay fixed across distributions. If auxiliary input U is used, its distribution must also stay fixed and it must be independent of the responses. Its likelihood factor then cancels from the stopped-record likelihood ratio, leaving the same information bound. A target selected from a reference block whose distribution changes with P does not automatically satisfy this premise. Our empirical reference-selected CITE comparisons therefore do not inherit the oracle-target lower-bound interpretation.

Proof of Corollary 4.4. Keep $P, h, a^\star$, the dense RESOLVE grid and weights, and every non-α CITE configuration fixed. CITE (Ota et al., 2026, Theorem 4.1) with an oracle target supplied before sampling is a level-α certifier. Theorem 4.3 and the cited pointwise lower bound, with $q < s$ and strict decrease of $\mathcal{I}(p, u)$, give

$$\limsup_{\alpha \downarrow 0} \frac{\mathbb{E}[\tau_c(\alpha)]}{\log(1/\alpha)} \le \frac{1}{\mathcal{I}(p,q)} < \frac{1}{\mathcal{I}(p,s)} \le \liminf_{\alpha \downarrow 0} \frac{\mathbb{E}[\tau_{\mathrm{CITE}}(\alpha)]}{\log(1/\alpha)}.$$

This is a strict separation of leading response-count bounds at fixed P. It is not an exact asymptotic formula for CITE and does not claim finite-budget dominance. □

C.5 Validity of Label-Target Comparison Procedures

We first compare fixed label and identity targets with matched evidence. We then prove validity for the three label confidence-sequence methods. These arguments require the prespecified inputs in Appendix A.4.

Matched Fixed-Candidate Targets. The controls share a supplied candidate, counts, weights, and boundary. Only their opponent sets differ. The following proposition establishes their pathwise ordering. Table 1 measures the size of the added identity cost.

Fix a direction c and a candidate $z \in \mathcal{A}_c$ before sampling. These controls use the global counts $N_t(a)$ from Section 3, including response one. On a deterministic finite or countable grid $\Lambda \subset (0, 1)$ with shared nonnegative weights summing to one, define

$$G_t^{\mathrm{label}}(z) := \sum_{\lambda \in \Lambda} v_\lambda (1+\lambda)^{N_t(z)} (1-\lambda)^{\sup_{b \in \mathcal{A}_{1-c}} N_t(b)},$$
$$G_t^{\mathrm{mode}}(z) := \sum_{\lambda \in \Lambda} v_\lambda (1+\lambda)^{N_t(z)} (1-\lambda)^{\sup_{b \in \mathcal{A} \setminus \{z\}} N_t(b)}.$$

Empty maxima and unseen counts are zero. With no observed competitor, either statistic is $\sum_{\lambda \in \Lambda} v_\lambda (1+\lambda)^{N_t(z)} \le 2^t$, not an automatic certificate. Fixed-label and Fixed-mode use these exact statistics at the same boundary. Neither is discovery-based RESOLVE or full CITE.

Proposition C.5. *Under the frozen i.i.d. setup, for the prespecified z, grid, and weights above and $\alpha \in (0, 1)$,*

$$\mathbb{P}\Big(\sup_{t \ge 0} G_t^{\mathrm{label}}(z) \ge 1/\alpha\Big) \le \alpha \quad \text{if } m_c \le m_{1-c},$$
$$\mathbb{P}\Big(\sup_{t \ge 0} G_t^{\mathrm{mode}}(z) \ge 1/\alpha\Big) \le \alpha \quad \text{if } p_z \le \sup_{b \in \mathcal{A} \setminus \{z\}} p_b.$$

Moreover, $G_t^{\mathrm{label}}(z) \ge G_t^{\mathrm{mode}}(z)$ for every response sequence and time. Their first exact crossing times at the same boundary satisfy $\tau^{\mathrm{fixed}}_{\mathrm{label}} \le \tau^{\mathrm{fixed}}_{\mathrm{mode}}$, with infinity allowed.

Proof. If $p_z = 0$, both statistics are at most one almost surely. Otherwise each null supplies a fixed opponent b with $p_b \ge p_z$ by Lemma C.1. The process $(1+\lambda)^{N_t(z)}(1-\lambda)^{N_t(b)}$ has conditional mean multiplier $1 + \lambda(p_z - p_b) \le 1$. It dominates the maximum-opponent account because that maximum is at least $N_t(b)$. Mixing these nonnegative supermartingales and applying Ville proves validity. If $\mathcal{A} = \{z\}$, the mode null is empty. No computed maximum-opponent statistic is asserted to be a supermartingale.

The label opponent set is contained in the mode opponent set. Since $1 - \lambda < 1$, the evidence ordering holds termwise, so every exact mode crossing is also a label crossing, including infinite stopping times. □

The ordering checks the matched experiment, not optimality. RESOLVE versus Fixed-label also changes discovery, count origins, and rank mixing, with no claimed pathwise ordering. Current matched replay verifies exact integer crossing boundaries and their neighboring counts. Historical log-domain crossings remain numerical results.

Full CITE handles empty observed competition differently. Its pairwise condition then passes, but its target-mass lower bound must still exceed an unseen-category upper bound. Dropping the latter requirement is invalid. At a binary tie, reporting immediately after the first target response would already have error $1/2$. Our fixed controls never use that automatic pairwise branch.

Validity of the Comparison Procedures. Appendix A.4 specifies the comparator inputs and allocations. Here we check the label-null adaptations and numerical guards. These arguments assume a fixed finite category set of size d and evaluator specified independently of the i.i.d. audit stream. We first treat nonempty label classes. If the candidate class is empty, it cannot report. If only the opposite class is empty, the directional null is empty and its error guarantee is vacuous. The interval procedures compare a candidate lower endpoint with zero, and the joint region excludes the empty null at the first inspection. These conventions do not establish validity for a support chosen from future responses, as in the evidence-task replay.

Outcome-CS. Each coordinate count sums independent indicators in $[0, 1]$. At directional level α, Hoeffding's inequality (Hoeffding, 1963) gives coordinate error at most $\alpha/[dt(t + 1)]$ with radius $\sqrt{\log(2dt(t + 1)/\alpha)/(2t)}$. Summing over d coordinates and $t \geq 1$ bounds any coverage failure by α. On joint coverage, a label-c lower endpoint above all opposite upper endpoints implies $m_c > m_{1-c}$. Clipping to $[0, 1]$ does not weaken coverage. The replay's strict comparison uses floating-point radii. Its retained returns have also been checked by exact inequalities bounding the logarithm. We do not assert equality to every mathematical first crossing.

Mixture-CS. For $n = N_t(a)$ and $\theta \in [0, 1]$, define the fixed Jeffreys mixture using the beta function B. Calligraphic $\mathcal{M}$ denotes mixture evidence, distinct from the opposing count $M_t(z)$. We use

$$\mathcal{M}_{a,t}(\theta) = \frac{\mathrm{B}(n + 1/2, t - n + 1/2)}{\mathrm{B}(1/2, 1/2)\theta^n(1 - \theta)^{t-n}}.$$

This is the binary Dirichlet-mixture construction with parameters $(1/2, 1/2)$ (Lindon and Malek, 2022, Theorems 2.2 and 2.4). At a boundary parameter it remains a nonnegative supermartingale on null-supported paths, since an alternative's next expected likelihood multiplier is its mass on that support, at most one. Use $0^0 = 1$ and infinite evidence for a positive count assigned zero probability. With coordinate allocation α/d, simultaneous inversion gives $\mathcal{M}_{a,t}(p_a) < d/\alpha$ for every a, t with probability at least $1 - \alpha$. Separating label-wise interval maxima then implies the strict label claim.

To confirm a report, let n and m be the largest candidate-label and opposing counts, and write $\mathcal{M}_{k,t}(x)$ for the mixture above with count k. The implementation proposes $m/t < x < n/t$ and exactly checks $\mathcal{M}_{n,t}(x), \mathcal{M}_{m,t}(x) \geq d/\alpha$. For any opposing count $k < m$, the beta recurrence gives

$$\frac{\mathcal{M}_{k+1,t}(x)}{\mathcal{M}_{k,t}(x)} = \frac{k + 1/2}{t - k - 1/2}\frac{1 - x}{x} < 1,$$

because $k + 1/2 < m < tx$. Hence every opposing mixture at x is at least $\mathcal{M}_{m,t}(x)$. Each mixture decreases below its empirical frequency and increases above it. On the joint no-crossing event, the candidate mass must exceed x and every opposing mass must be below x. Thus even the nonstrict exact confirmation implies the strict label claim on that event. The numerical root bracket only proposes a report. Rational beta identities confirm it and may delay, not create, a crossing.

Multinomial-CS. For d known categories, identify a categorical distribution with its mass vector θ. Write $\mathbf{p} = (p_a)_{a\in\mathcal{A}}$ for the true mass vector, $\mathbf{N}_t = (N_t(a))_{a\in\mathcal{A}}$ for the counts, and $\mathbf{1}$ for the all-ones vector. With multivariate beta function B, the uniform Dirichlet mixture for the ordered sequence is

$$\mathcal{M}_t(\theta) = \frac{\mathrm{B}(\mathbf{1}+\mathbf{N}_t)}{\mathrm{B}(\mathbf{1})}\prod_{a\in\mathcal{A}}\theta_a^{-N_t(a)}.$$

No multinomial coefficient is needed for an ordered sequence. The confidence sequence follows the Dirichlet-mixture construction (Lindon and Malek, 2022, Theorem 2.4), with the same boundary extension just checked above. Our inversion reports when $\inf_{\theta\in H_{0,c}}\mathcal{M}_t(\theta) \geq 1/\alpha$. Under a true null this implies a crossing by $\mathcal{M}_t(\mathbf{p})$, so Ville bounds the report probability by α.

If both directions use the same region at level α_{query}, either false report implies $\sup_{t\geq 0}\mathcal{M}_t(\mathbf{p}) \geq 1/\alpha_{\text{query}}$. The same Ville bound therefore protects their union without dividing the level. This justifies Appendix B.4.

The remaining question is how to invert the label null without a numerical optimizer changing the target. The following lemma establishes the pooling calculation used in the implementation.

Lemma C.6. *Let $\mathcal{A}$ be finite, with both $\mathcal{A}_c$ and $\mathcal{A}_{1-c}$ nonempty, and let nonnegative counts n_a sum to $t > 0$. For each opposing $b \in \mathcal{A}_{1-c}$, start $S_b = \{b\}$ and visit candidate-label counts in decreasing order. Add a candidate while its count exceeds the current mean in S_b, stopping at the first that does not. Define*

$$\widehat{\theta}_a^{(b)} = \begin{cases} \sum_{j\in S_b} n_j/(t|S_b|), & a \in S_b, \\ n_a/t, & a \notin S_b. \end{cases}$$

Then $\widehat{\theta}^{(b)}$ maximizes $L(\theta) := \sum_{a\in\mathcal{A}} n_a \log\theta_a$ over probability vectors satisfying $\theta_a \leq \theta_b$ for every $a \in \mathcal{A}_c$. Zero-count terms are zero. A largest-count opponent $d_0 \in \mathcal{A}_{1-c}$ attains the maximum over the full label null $H_{0,c}$, including ties and zero counts.

Proof. For each $b \in \mathcal{A}_{1-c}$, let Θ_b be the set of probability vectors satisfying $\theta_a \leq \theta_b$ for every $a \in \mathcal{A}_c$. Because the category set is finite and both labels are present,

$$H_{0,c} = \bigcup_{b\in\mathcal{A}_{1-c}}\Theta_b.$$

Under the null, some opposite-label category is at least as probable as every candidate-label category. Minimizing $\mathcal{M}_t(\theta)$ amounts to maximizing $L(\theta)$ over this union. We solve each constraint set by pooling, then show that a largest-count opponent suffices.

A fixed opponent. Each addition raises the mean to below the newly added count. Descending order leaves all included candidate counts above the final mean and all excluded ones at or below it, including ties. Pooling preserves the count sum, so $\widehat{\theta}^{(b)}$ is normalized and feasible.

If the pooled mean is zero, n_b and all candidate-label counts are zero. The empirical vector is already feasible and maximizes the unconstrained likelihood. Otherwise set $u = \widehat{\theta}_b^{(b)} > 0$, $\mu_a = n_a/u - t \geq 0$ for $a \in S_b \setminus \{b\}$, and $Z = \{a \notin S_b : n_a = 0\}$. The pooled sum gives $n_b/u = t - \sum_{a\in S_b\setminus\{b\}}\mu_a$. For any feasible θ assigning positive probability to every positive count, $\log x \leq x - 1$ gives

$$\begin{aligned} L(\theta) - L(\widehat{\theta}^{(b)}) &\leq \sum_{a:n_a>0}\frac{n_a}{\widehat{\theta}_a^{(b)}}(\theta_a - \widehat{\theta}_a^{(b)}) \\ &= -t\sum_{a\in Z}\theta_a + \sum_{a\in S_b\setminus\{b\}}\mu_a(\theta_a - \theta_b) \leq 0. \end{aligned}$$

The equality uses normalization and the common pooled probability. The last inequality uses $\mu_a \geq 0$, $\theta_a \leq \theta_b$ for pooled candidates, and nonnegative probabilities. A positive count assigned zero probability gives log likelihood $-\infty$. Thus optimality does not require differentiation at zero.

A largest-count opponent. A vector feasible for b is already feasible for a largest-count opponent d_0 when $\theta_{d_0} \geq \theta_b$. Otherwise swapping those probabilities makes it feasible for d_0 and changes log

likelihood by $(n_{d_0} - n_b)\log(\theta_b/\theta_{d_0}) \geq 0$. For zero coordinates, first mix the vector with weight $\eta > 0$ of the uniform vector, which preserves feasibility. Apply the interior swap argument and let $\eta \downarrow 0$ in the continuous likelihood product, using $0^0 = 1$. Thus the largest opposing count attains the union maximum, with ties allowed. □

The lemma justifies the guard's rational checks at every opposing constrained maximizer. At time zero $\mathcal{M}_0 = 1$, so no report occurs for $\alpha < 1$, and we use no division by t. An empty candidate class cannot win. An empty opposite class gives an empty null and a report at the first inspection, without pooling. Both classes exist in the full option category sets but need not in realized evidence-task supports. These branches preserve the lemma's assumptions and do not give the support-oracle replay a prospective guarantee.

C.6 Reproduced CITE Procedure and Validity

Here α is the total level of the fixed-target certifier. Use the global counts $N_t(a)$ from Section 3. For CITE's supplied target z, write $n = N_t(z)$ and $m = \max_{a \neq z} N_t(a)$, with zero for an empty observed competitor set. Its pairwise mixture must reach $3/\alpha$. Empty observed competition passes this component only. A target-mass lower bound L_t must also exceed an unseen-category bound U_t. We fix grids from the published families, not assumed experimental defaults. The pairwise grid is $\{2^{-1}, \ldots, 2^{-13}\}$ with uniform weights, corresponding to geometric-grid parameter 0.001. The target-mass grid is $\{2^{-1}, \ldots, 2^{-8}\}$ with uniform weights. For $t \geq 1$ and $0 \leq u \leq n/t$, define

$$\mathcal{M}_t^{\text{mass}}(u) = \frac{1}{8}\sum_{k=1}^{8}[1 + 2^{-k}(1-u)]^n[1 - 2^{-k}u]^{t-n},$$
$$L_t = \sup\{u \in [0, n/t] : \mathcal{M}_t^{\text{mass}}(u) \geq 3/\alpha\},$$
$$(1 - U_t)^t/U_t = \alpha/3.$$

An empty lower-bound set gives $L_t = 0$. These positive mass bets are admissible on the full unit interval. The unseen equation has a unique solution in $(0, 1)$. A report requires both the pairwise condition and $L_t > U_t$.

The implementation verifies $(1-u)^t \leq \alpha u/3$ for a dyadic u, and requires $n/t > u$ and $\mathcal{M}_t^{\text{mass}}(u) > 3/\alpha$ by rational arithmetic. The unseen-tail ratio decreases in u, so $u \geq U_t$. The strict mass check and $u < n/t$ let continuity supply a larger feasible point, giving $L_t > u$. The nonempty pairwise condition is also exactly confirmed. Appendix B.2.1 describes the retained validation record and its missing simulation-level detail.

Condition on a genuinely independent reference block and all frozen analysis rules. Its selected target is then supplied before the monitored i.i.d. stream. CITE's three-component validity theorem (Ota et al., 2026, Algorithm 1 and Theorem 4.1) applies with the three declared component budgets. The geometric pairwise and positive target-mass grids satisfy their component conditions. The rational upper bracket and strict mass check above imply $L_t > U_t$. Its pairwise confirmation checks the actual nonempty-competitor statistic. Thus our reports are a subset of full mathematical CITE reports. Averaging over an independent reference preserves the bound. A reference target need not be the true mode, so this argument gives validity, not guaranteed return. Missing archived map pre-specification remains a separate limitation. These reference-selected experiments do not satisfy the invariant-side-information premise of the oracle-target lower-bound comparison automatically.

D Extended Related Work

Extending the Introduction, we distinguish RESOLVE's target from related inference goals and its contribution from the sequential tools and information bounds it inherits.

D.1 Voting, Sampling, and Early Stopping

Stopping rules ask when further samples are unnecessary for answer selection. Self-consistency (Wang et al., 2023) selects by answer frequency without a verifier. Adaptive-Consistency (Aggarwal et al., 2023) uses a beta-based stopping criterion, while ESC (Li et al., 2024) stops on agreement

within a window. These methods seek useful selection at lower cost, a different objective from certifying the population winner's evaluator label.

Cost reduction also depends on what further voting favors. A mixture of queries helped and harmed by voting can make aggregate performance first improve and then worsen (Chen et al., 2024). Best-of-Infinity (Komiyama et al., 2026) studies limiting voting performance and adaptive sampling. Blend-ASC (Feng et al., 2026) allocates samples across questions. RESOLVE instead audits one fixed query's population label before interpreting more voting as progress.

Repeated sampling can make a useful answer available without ensuring its selection (Brown et al., 2024). RESOLVE audits this distinction under Proposition 2.1's unweighted plurality of fixed i.i.d. outputs, not arbitrary inference pipelines. Auditing a larger frozen pipeline requires independent repetitions, each costing a full execution.

D.2 Population Modes and Anytime Inference

Mode certification asks whether one category dominates every rival. CITE (Ota et al., 2026) answers this question for a supplied target on an unknown category space. RESOLVE instead discovers candidate categories online and compares the strongest category in each evaluator class. This also differs from CITE's top-k extension, which separates the weakest supplied member from the strongest outsider. We do not evaluate adaptive target selection for CITE.

PPR-1v1 (Anand Jain et al., 2022) compares pairs on a finite category set using prior-posterior-ratio confidence sequences. PPR-Adaptive (Anand Jain et al., 2022, Section 5.2) assigns summable error budgets as answers appear. Identity certification therefore need not assume a known category space. Certified Self-Consistency (Cordero-Encinar and Duncan, 2025) combines MMC with distribution sharpening. Our replay of the released ratio-update MMC implementation compares a different stopping target, with the guarantee qualification in Appendix B.4. It is not a substitute for CITE or a label audit.

With supplied categories and their labels, confidence sequences can also test the label question. Lindon and Malek (2022) develop a multinomial mixture that covers the probability vector uniformly in time. Appendix C.5 derives our label-null inversion and coordinatewise alternatives. They receive information that RESOLVE discovers online. Fixed-label and Fixed-mode instead isolate the opponent set and are neither PPR nor full CITE. The mixed MMLU-Pro results in Table 2 rule out a blanket ranking.

D.3 Evaluators, Selection, and Representation

An evaluator may select an answer or assess an unchanged selection rule. Verifier selection and adaptive compute serve the first role (Cobbe et al., 2021; Snell et al., 2025), whereas RESOLVE serves the second. Instruction checks and LLM judges offer possible criteria, but model-based judgments have documented biases (Zhou et al., 2023; Zheng et al., 2023; Wang et al., 2024a). Our theorem requires a fixed label for each category. Adaptive or stochastic judges need a different validity argument even under frozen generation.

The category map determines which responses share a vote. Semantic uncertainty groups responses by meaning rather than surface form (Kuhn et al., 2023; Farquhar et al., 2024). Such grouping addresses representation, not by itself whether the resulting plurality passes an evaluator. Learning groups from monitored responses falls outside our frozen-map guarantee. Appendix B.5.2 checks parser sensitivity, and Appendix A.1 explains why a correct entity or verdict need not meet the full evidence criterion. FAIL alone establishes neither factual error nor absence of document support.

D.4 Sequential Evidence and Information Bounds

The sequential question is how to preserve error control while evidence is repeatedly inspected. Betting mixtures and Ville's inequality provide established tools (Ville, 1939; Shafer, 2021; Howard et al., 2021; Ramdas et al., 2020). Theorem 4.1 uses them through a nonnegative supermartingale supplied by each null distribution and dominating the discovered evidence. The latter need not itself be a supermartingale. RESOLVE's contribution is their construction for the evaluator target with online discovery, not a new generic pairwise testing principle.

Response cost is separate. Sequential information bounds link error levels to observations needed to distinguish distributions (Wald, 1945; Hoeffding, 1953; Kaufmann et al., 2016). Appendix C.4 uses CITE's fixed-target bound, with a positive runner-up and invariant independent side information. Our uncapped upper bound separates leading response scales at fixed P as $\alpha \downarrow 0$, not exact finite-budget costs. In region identification (Chen et al., 2017) and bandits with several correct recommendations (Degenne and Koolen, 2019), the learner chooses what to sample. RESOLVE instead performs inference on one fixed response stream.