WHEN IS A CHEAP LLM AUDIT DECISIVE? PROSPECTIVE FORECASTS AND MEASURED COSTS

Anonymous authors
Paper under double-blind review

ABSTRACT

The value of efficient evaluation lies in the decisions its evidence supports. We introduce a prospective audit protocol that separates false-assertion control, the probability of a correct definitive decision, and the cost of acquiring model outputs. Before observing new target outcomes, a fixed planner chooses among whole-question sampling, a paid target pilot, partial trajectories, and complete execution. On two new question pools shared by Qwen3 and Ministral3, the selected CommonsenseQA audits detect excessive risk with probabilities 97.51% and 98.80% at 47.5% and 40.1% of full-evaluation cost when source logs already exist. OpenBookQA audits cost 51.5% and 58.1% but confirm the acceptable pools with probabilities 63.97% and 26.83%. Constructing the source logs used by the planner raises its four evaluation-cost ratios to 1.25–1.76. All four prospective uniform-audit forecasts underestimate decision probability. Exact sample-size curves, same-observation inference controls, and an algorithmic CollabEval comparison distinguish the roles of sampling, inference, and information cost. These measurements connect evaluation efficiency to decision readiness under fixed finite-pool designs.

1 INTRODUCTION

Evaluation efficiency is a decision problem. To assess whether a model's accepted-answer error rate meets a threshold, an evaluator needs evidence precise enough to resolve that comparison. Statistical validity, the probability of reaching a conclusion, and the cost of obtaining that evidence play distinct roles. This distinction is especially consequential for long-reasoning models, whose trajectories have heterogeneous runtimes.

Existing methods make evaluation substantially more efficient by choosing informative examples, exploiting historical model outputs, or predicting an outcome before execution ends (Kossen et al., 2021; Fisch et al., 2026; Shi et al., 2026). These advances motivate a complementary question: *when does a reduction in evaluation computation provide enough evidence for the decision that an evaluator actually needs?* We study that question for a fixed benchmark roster and a fixed realized target policy. With public gold labels already available, target-model execution is the resource to allocate.

We distinguish three quantities. **Validity** controls the probability of an incorrect definitive assertion. **Decisiveness** is the probability of a correct definitive assertion; an unresolved result leaves the decision open. **Cost** includes the information that the evaluator must acquire. We measure their interaction prospectively: fix an acquisition forecast, execute the chosen audit, compare it with a hidden full census, and measure the cost of obtaining the evidence.

Two reasoning-model families share new OpenBookQA and CommonsenseQA pools, each with 32 pilot and 192 audit questions. Before target outputs, the planner chooses uniform or source-stratified sampling, a paid pilot, fixed prefix continuation, or census. All designs evaluate the same target policy, with partially observed trajectories treated as censored observations.

The selected CommonsenseQA audits exceed the .90 decisiveness target at less than half of census cost when source logs already exist. The selected OpenBookQA audits have similar cost fractions but often remain unresolved. Acquiring source information reverses the cost ordering of these evaluation designs. The hidden census separately evaluates the forecasts and audit conclusions.

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.

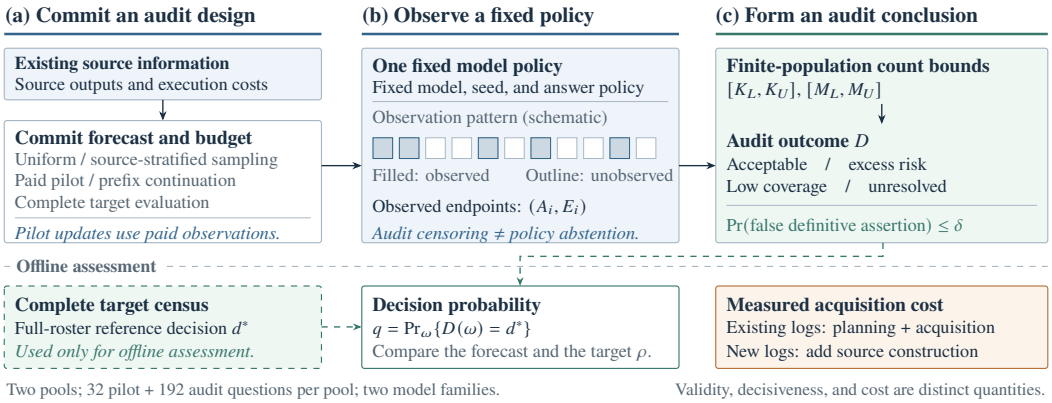

Figure 1: **Separate acquisition, inference, and offline evaluation.** An outcome-blind forecast selects a predefined acquisition design. The audit observes outputs of the same fixed target policy and applies finite-population tests. The target census provides the offline decision-probability reference; procurement uses only information available beforehand. Runtime measurements distinguish existing source information from information that must be acquired. Audit censoring leaves an outcome unobserved rather than changing the policy to abstention.

We contribute a prospective acquisition-and-audit protocol, exact finite-pool decision-power curves, and measured costs under existing versus newly acquired information. The curves distinguish first from sustained threshold crossings; matched-observation controls isolate the effect of the inference procedure from acquisition.

2 Related work

Benchmarking and efficient evaluation. MMLU measures broad multitask accuracy (Hendrycks et al., 2021), while HELM evaluates models across scenarios and multiple criteria, including calibration and efficiency (Liang et al., 2023). Reduced evaluations such as tinyBenchmarks estimate benchmark performance from selected examples (Maia Polo et al., 2024). Our target is a statistically qualified decision about a fixed policy's accepted-answer risk and coverage on one finite roster. Decision probability adds an operational target to aggregate accuracy estimation.

Historical predictions and acquisition. Prediction-powered inference corrects auxiliary prediction errors (Angelopoulos et al., 2023); cross-prediction reuses labeled observations through cross-fitting (Zrnic & Candès, 2023). CollabEval combines matrix completion and prediction-assisted inference over model-by-prompt scores (Fisch et al., 2026). FAQ combines historical factor models with active finite-bank accuracy estimation; its PAI intervals rely on a martingale central limit theorem under explicit regularity conditions (Wu et al., 2026). Active testing selects informative evaluation examples (Kossen et al., 2021), and Prediction-Powered Active Testing combines selective acquisition with residualized losses (Ashouritaklimi et al., 2026). EarlyEval predicts an agent's final outcome from intermediate behavior (Shi et al., 2026). Our comparisons use CollabEval's published algorithmic core, an anchor-mean control, and a fixed randomized prefix-continuation design. The empirical question is what these information purchases permit the evaluator to conclude, at their measured acquisition cost.

Selective prediction and risk control. Selective classification trades answered-example coverage against conditional error, using a rejection rule around a trained predictor or learning prediction and rejection jointly (Geifman & El-Yaniv, 2017; 2019). Conformal prediction constructs prediction sets with marginal coverage guarantees (Angelopoulos & Bates, 2021); risk-controlling prediction sets and Learn then Test calibrate predictive procedures against explicit risk requirements (Bates et al., 2021; Angelopoulos et al., 2021). Our audit evaluates an answer policy fixed in advance. The labels acceptable, excess risk, and low coverage express conclusions about its finite-roster performance;

unresolved denotes unresolved audit evidence. This formulation places the cost and decisiveness of policy assessment alongside the policy's own coverage–risk tradeoff.

Sequential testing and finite-population auditing. Wald's sequential tests link error control to sampling cost (Wald, 1945); risk-limiting election audits can escalate sampling to a complete recount (Stark, 2008). Confidence sequences support time-uniform inference (Howard et al., 2021), including sampling without replacement (Waudby-Smith & Ramdas, 2020); financial audits also exploit weighted acquisition and side information (Shekhar et al., 2023). We use classical count-bound inversions under fixed within-arm sampling designs. This common inferential basis makes decision probabilities, forecasts, and acquisition costs directly comparable.

3 Auditing a fixed target policy

3.1 Outcomes, assertions, and audit randomness

For a roster of N questions, fix the model, prompt, decoding seed, maximum reasoning budget, and answer readout. Let $A_i \in \{0,1\}$ indicate that the deployment policy answers question i, and let $E_i = A_i \mathbf{1}\{\widehat{y}_i \neq y_i\}$ indicate an accepted error. Define

$$K = \sum_{i=1}^{N} A_i, \quad M = \sum_{i=1}^{N} E_i, \quad C = K/N, \qquad R = M/K \quad (K > 0). \tag{1}$$

A zero denominator leaves risk undefined. The audit samples observations and returns acceptable, excess risk, low coverage, or unresolved. At risk budget $\alpha = .10$ and coverage floor $c_0 = .20$, the first three statements assert $R \leq \alpha$ and $C \geq c_0$, $R > \alpha$, and $C < c_0$, respectively. The predeclared priority is acceptable, then excess risk, then low coverage; the supplement also records cases in which excess risk and low coverage both hold.

Let ω denote only the audit's random choices. Its decision $D(\omega)$ is valid when

$$\mathbb{P}_\omega\{D(\omega) \text{ is a false definitive assertion}\} \leq \delta. \tag{2}$$

Decisiveness additionally measures how often the evidence resolves the target. For the fixed census truth d^*, define the registered decisive probability

$$q = \mathbb{P}_\omega\{D(\omega) = d^*\}, \qquad D(\omega) = \text{unresolved contributes zero.} \tag{3}$$

We use $\delta = .05$ and, in the independent study, $\rho = .90$ as the desired value of q. The former controls erroneous audit assertions; the latter sets the desired probability of a correct audit conclusion.

3.2 Exact evidence and exact decision-power curves

For uniform sampling of n questions, let a be the observed number accepted and e the observed number of accepted errors. Classical finite-population tail inversion (Cochran, 1977) gives bounds $[K_L, K_U]$ and $[M_L, M_U]$. The four one-sided bounds each use error budget $\delta/4$, with inclusive integer-tail comparisons. The audit returns

$$\begin{aligned} \text{acceptable} &: K_L > 0, \quad M_U \leq \alpha K_L, \quad K_L \geq c_0 N, \\ \text{excess risk} &: K_L > 0, \quad M_L > \alpha K_U, \\ \text{low coverage} &: K_U < c_0 N, \end{aligned} \tag{4}$$

in that order, and otherwise returns unresolved. A union bound supplies Eq. (2). With H nonempty sampled strata, the same construction allocates $\delta/(4H)$ to each one-sided stratum bound and sums the bounds; fully observed strata are exact. Appendix C states the conditioning and proof.

Once a hidden census is available, the probability of each possible observed pair is

$$\mathbb{P}(a, e \mid N, K, M, n) = \frac{\binom{M}{e}\binom{K-M}{a-e}\binom{N-K}{n-a}}{\binom{N}{n}}. \tag{5}$$

Applying Eq. (4) to each feasible pair and summing Eq. (5) computes $q(n)$ without Monte Carlo error. This *offline diagnostic* uses census counts K, M reserved for the evaluator. We evaluate both the

frozen forecast worlds and the actual census with the same test. A conditional-risk hypergeometric construction provides a separately declared uniform-sampling sensitivity.

Discrete acceptance regions can make $q(n)$ locally nonmonotone. We report both the first threshold crossing $n_\rho = \min\{n : q(n) \geq \rho\}$ and the sustained crossing

$$n_\rho^+ = \min\{n : \min_{m \in \{n,\ldots,N\}} q(m) \geq \rho\}. \tag{6}$$

The unsmoothed curves characterize both thresholds for the specified audit.

3.3 Cost follows the information actually acquired

Let T_{run} be the measured acquisition window, T_{plan} the single-budget planning time, and T_{source} the cost of constructing the current source logs. We report

$$T_{\text{held}} = T_{\text{plan}} + T_{\text{run}} + T_{\text{paid}}, \qquad T_{\text{built}} = T_{\text{source}} + T_{\text{held}}, \tag{7}$$

where T_{paid} includes a pilot only for a method that purchases it. Direct census and standalone uniform sampling incur target-evaluation costs alone. The model is already resident in the primary windows; initialization-inclusive costs are a separate sensitivity.

A lower runtime qualifies as decision-ready evaluation only together with the requirement $q \geq \rho$. The acquisition budget constrains expected cost; selected trajectories are evaluated to completion.

4 Prospective experimental design

4.1 New questions, fixed models, and information separation

We use OpenBookQA (Mihaylov et al., 2018) and CommonsenseQA (Talmor et al., 2019). A deterministic ordering selects 224 previously unrun questions per pool, split into 32 pilot and 192 audit questions. OpenBookQA is disjoint from the development sample; CommonsenseQA is a new task for this study. The paired design contains 448 unique questions and 896 question–model evaluations. Model inputs consist of question text and the original ordered answer options.

Source observations use BF16 models and target observations use fixed NF4 representations of Qwen3-8B and Ministral-3-8B-Reasoning (Yang et al., 2025; Liu et al., 2026). Precision defines the source–target pair throughout the acquisition comparison. Models generate their own reasoning, stopping at the native close marker or a cap of 8,192 new tokens. A fixed option-likelihood readout scores the possible labels after completed reasoning. The primary policy answers every validly completed item; the secondary policy also requires maximum normalized option probability at least .99. Both acceptance policies are specified before target evaluation. Appendix B gives the sampler and readout.

All source logs are constructed first. Procurement forecasts for *both pools and both families* are committed together before any new NF4 output. Only the paid-pilot arm can update its plan using its own pilot. The complete target outcomes are reserved for offline evaluation. Each audit receives only its selected observations. Model seeds and audit randomizations are separate; live executions are compared against the same fixed target trajectories.

4.2 Procurement menu and the frozen forecaster

The menu contains uniform whole-question sampling (U), source-length stratification with relative long-stratum weights $.25, 1, 4$ (S.25/S1/S4), a target-prefix design (P), a paid pilot, and direct census. The source split is 1,024 generated tokens. P starts each item and, for trajectories still running at 1,024 tokens, uses a common preselected continuation probability in $\{.25, .5, .75, 1\}$. A discontinued suffix remains an unobserved outcome of the fixed target policy.

The forecaster resamples joint source/target records from separate development tasks, conditioning on model family, source completion, source correctness, and source length. It predicts target length through the paired length ratio and converts that length to time using frozen component measurements. The exact hierarchy and small-cell fallback are fixed. A paid-pilot forecast averages over simulated

Pool	Model	K/M	$R\,(\%)$	$\widehat{q}_U$	q_U	Chosen	q_{chosen}
CSQA	Qwen	192/40	20.83	48.75	97.51	U	97.51
CSQA	Ministral	190/46	24.21	33.28	99.81	S4	98.80*
OBQA	Qwen	192/8	4.17	61.47	63.97	U	63.97
OBQA	Ministral	190/11	5.79	25.37	26.83	U	26.83

Table 1: **Cheap audits can be decisive or inconclusive.** All four primary cells have $N = 192$ and $\beta = .50$. K/M are full-census accepted/error counts. Forecast and actual q_U are exact probabilities, shown as percentages. *S4's 98.80% uses 1,000 audit randomizations, with registered bounds [96.82, 99.71]%. CSQA is truly over the risk budget; OBQA satisfies the finite-pool conditions. A high audit q does not mean high model accuracy.

pilot outcomes; actual pilot acquisition must therefore be chosen *before* seeing whether the pilot helps. The experiment evaluates this fixed forecasting rule prospectively.

Budgets are $.25, .50, .75, 1.00$ times the forecast full-audit cost; $.50$ is primary. Plans maximize predicted decisive probability among forecast-affordable arms, with fixed cost and arm tie-breaks. For U at the primary budget, $n = 96$. The factorial evaluation covers 32 policy–budget cells and 224 acquisition designs, with 1,000 randomizations per design. The randomizations quantify sampling variation within each fixed pool.

4.3 Prospective hypotheses and cost measurement

The prospective analysis tests forecast calibration, low-cost decisiveness, and the cost of constructing source information. Appendix D gives the preregistered joint criteria and their directional assessment.

Acquisition cost measures the computation required to obtain the audit evidence. Each cell includes a timed evaluation of the selected design and of U, or S.25 when U is selected; a full target census provides the reference outcome and cost. Measurement covers input preparation, prefill, generation, and answer scoring, including output recording. Source acquisition, planning, pilot updates, and model initialization are accounted for separately. Each design–cell comparison uses one complete execution under the same configuration (Appendix B).

5 Results

5.1 The same budget buys different kinds of evidence

On CommonsenseQA, both full pools violate the 10% risk budget: Qwen makes 40 accepted errors out of 192, and Ministral 46 out of 190. Uniform half-pool sampling correctly identifies the violation with exact probabilities 97.51% and 99.81%. The selected audits are U for Qwen and S4 for Ministral; both exceed the .90 decision-service target. Uniform sampling has slightly higher decision probability than the selected S4 design; their respective execution costs are reported separately.

OpenBookQA yields the opposite decision problem. Qwen's 8/192 and Ministral's 11/190 accepted errors satisfy the full-pool risk requirement. Their selected U audits confirm it with probabilities 63.97% and 26.83%. Most remaining audits leave the decision unresolved. Both single live OpenBookQA selected windows return unresolved; both CommonsenseQA selected windows return excess risk. Table 1 separates those finite-pool facts from stochastic audit behavior.

The two pools exhibit distinct decision regimes under a common audit requirement. Their distances from the risk boundary, completion rates, and error patterns determine the evidence needed to resolve each case.

5.2 Prospective forecasts are conservative on the new pools

Every actual q_U exceeds its frozen forecast. On CommonsenseQA, the underestimates are 48.76 and 66.54 percentage points; on OpenBookQA, they are 2.50 and 1.46 points. The forecast errors therefore reverse the preregistered overoptimism direction. The two OpenBookQA audits illustrate

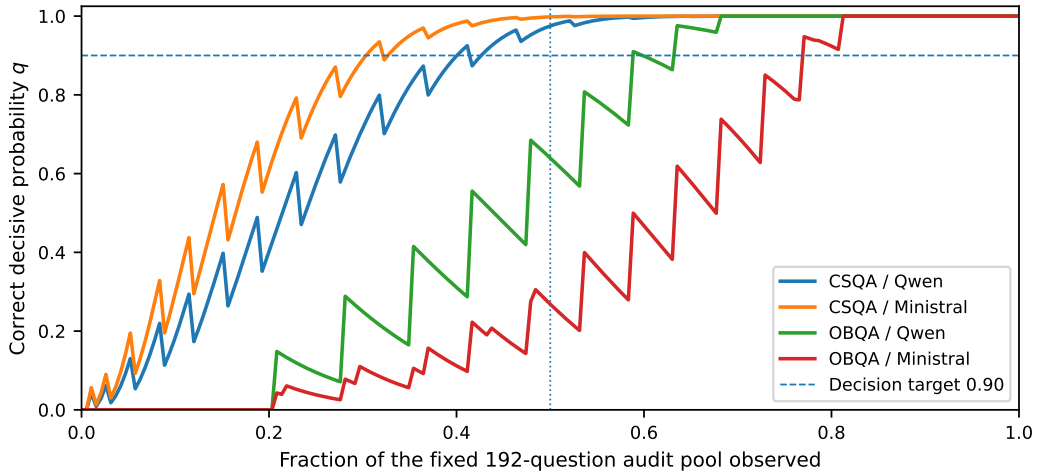

Figure 2: **Decision service is a curve, not a sample fraction.** Exact $q(n)$ for the original uniform audit on each realized primary pool. The dashed line is the prespecified .90 service target; the vertical line is half the pool. Visible local drops are retained consequences of discrete test boundaries. Each curve conditions on the complete census and the specified fixed-sample test.

Pool	Model	Forecast $n_{.90}$	First $n_{.90}$	Sustained $n^{+}_{.90}$	$n^{+}_{.90}/192$ (%)
CSQA	Qwen	149	77	82	42.7
CSQA	Ministral	172	59	63	32.8
OBQA	Qwen	131	113	122	63.5
OBQA	Ministral	176	148	148	77.1

Table 2: **A first discrete crossing need not persist.** $n^{+}_{.90}$ is the smallest size after which every larger size meets .90, for this fixed U test. Forecast sizes use the committed mixture of worlds. All counts are out of 192. Multiplying these fractions by census component cost gives an expected component-cost diagnostic, not a new live timing.

inexpensive but frequently unresolved certification, whereas the CommonsenseQA audits are both inexpensive and decisive.

Forecasted outcome counts reveal where the discrepancy arises. On CommonsenseQA, predicted accepted-error counts average 28.81 and 25.59, versus actual counts 40 and 46; both actual counts exceed all 200 forecast worlds, whose maxima are 37 and 36. On OpenBookQA, the corresponding means are 8.19 and 11.80, versus 8 and 11 (Appendix D.1). At fixed n, the census counts (N, K, M) and the test fully determine $q(n)$. A post-hoc substitution of true component prices changes only two of the 32 policy/budget choices, both on OpenBookQA/Qwen at $\beta = .75$. The corresponding outcome-count discrepancy and limited change in selected designs distinguish probability forecasting from runtime prediction.

5.3 Exact curves expose the service requirement

Table 2 reports the first and sustained .90 crossings. For CommonsenseQA, the first crossings are 77 and 59 questions, below half the pool; the sustained crossings are 82 and 63. OpenBookQA needs 113 and 148 questions for a first crossing, with sustained crossings at 122 and 148. For OBQA/Qwen, the first crossing at 113 and sustained crossing at 122 therefore describe different sample-size requirements.

Inference design changes decision probability even when the observations are identical. The separately specified conditional-risk construction gives the same q at $n = 96$ in all four primary cells. This alternative therefore preserves the primary comparison. On the same 1,000 uniform observation masks, global versus stratified bounds give decisive rates of 65.9% versus 8.2% for OBQA/Qwen

Pool	Model	Design	n	q (%)	Time (min)	Held	Built
CSQA	Qwen	U	96	97.51	73.0	0.472	0.472
CSQA	Qwen	Selected U	96	97.51	73.0	0.475	1.249
CSQA	Qwen	Census	192	100.00	154.5	1.000	1.000
CSQA	Ministral	U	96	99.81	112.0	0.488	0.488
CSQA	Ministral	Selected S4	71	98.80*	91.5	0.401	1.493
CSQA	Ministral	Census	192	100.00	229.4	1.000	1.000
OBQA	Qwen	U	96	63.97	53.7	0.511	0.511
OBQA	Qwen	Selected U	96	63.97	53.7	0.515	1.758
OBQA	Qwen	Census	192	100.00	105.2	1.000	1.000
OBQA	Ministral	U	96	26.83	94.3	0.578	0.578
OBQA	Ministral	Selected U	96	26.83	94.3	0.581	1.689
OBQA	Ministral	Census	192	100.00	163.2	1.000	1.000

Table 3: **Decision probability and acquisition cost on the same rosters.** Time is measured target acquisition; Held and Built are ratios to the resident-model census. Selected rows add planner cost in Held and its source construction in Built. Source-free U and census purchase neither. *S4 uses 1,000 audit randomizations; other probabilities are exact. The service target is $q \geq .90$.

and 27.2% versus 0.5% for OBQA/Ministral (Appendix E). Because the observed outcomes are unchanged, this contrast isolates the effect of the specified inference procedure.

The global–stratified difference persists across the tested error-budget allocations. This post hoc sensitivity uses the same observations. In addition to equal tail budgets, we examine stratum-size weighting and an allocation that gives each error-count tail twice the budget of each acceptance-count tail (Appendix E.1). The three stratified decision rates are 8.2/7.1/9.1% for OBQA/Qwen and 0.5/0.5/0.8% for OBQA/Ministral. The difference from global inference persists under these alternatives. For global U, error-weighting leaves all four $q(96)$ values unchanged but moves the sustained .90 crossings from 82/63/122/148 to 77/59/120/146. The sampling design and measured execution costs are unchanged.

Acceptance criteria also shape decision readiness. The prespecified .99 confidence filter changes the audited policy. It increases OBQA/Qwen's $q(96)$ from 63.97% to 77.78%, while both CommonsenseQA pools remain above .90 and both OpenBookQA pools below it. Appendix B.1 gives all eight policy–pool–model results.

5.4 Source ownership reverses the measured cost comparison

Source ownership determines the cost of the selected audit. With existing source logs, CommonsenseQA evaluation costs 47.5% and 40.1% of direct execution while achieving the declared decision-service target. Selected OpenBookQA evaluation costs 51.5% and 58.1%, but falls short of that target. The joint cost–decisiveness comparison separates these operating regimes.

Adding source acquisition raises all evaluation-cost ratios to 1.25–1.76, meeting the separately prespecified source-cost criterion. Three selected arms are U: a standalone uniform audit can omit both source construction and planning. Each cost comparison includes precisely the information purchased by its acquisition design. For CSQA/Ministral, source-free U observes 96 questions in 112.0 minutes (48.81% of the census), while the selected S4 acquisition observes 71 in 91.5 minutes (39.89%) before planning. Their decision probabilities are 99.81% (exact U) and 98.80% (S4 randomizations). These operating points compare acquisition designs; the allocation sensitivity holds their observations fixed.

Including the same measured target initialization in both selected and direct evaluation changes the held-log ratios to .477, .402, .517, and .582. All selected acquisition windows fall within their expected budgets, without outcome-dependent truncation. Figure 3 compares the two source-information assumptions.

5.5 PREDICTION-ASSISTED INFERENCE UNDER MATCHED OBSERVATION BUDGETS

We reproduce CollabEval's published Algorithms 1–3 and covariance formula independently, using two BF16 anchor rows and two NF4 target rows on the same 192 columns. A mean of anchors supplies a simpler CrossPPI comparator; classic IID estimation supplies another. The 4-row matrix defines the attainable ranks and the scale of this comparison.

At half-pool observation, the CollabEval core's decisive rates under its *asymptotic* analysis are 93.7%, 99.5%, 93.8%, and 44.1%; the anchor comparator gives 90.3%, 99.2%, 92.5%, and 56.0%. These estimates use a different validity analysis from the exact audit and are reported separately. At quarter-pool observation, undefined delta-method risk intervals remain in the coverage denominators. Appendix G reports risk MSE, coverage, decisive returns, and undefined intervals together.

Auxiliary information contributes its own acquisition cost. CollabEval purchases observed companion-target outputs, whereas anchor-only estimation uses the source observations alone. Table 14 pairs the existing decision probabilities with own-target, companion-target, and source-anchor component costs. Shared execution is charged once across model rows, keeping statistical precision and computational expenditure separately measurable.

6 IMPLICATIONS AND SCOPE

Decision-oriented evaluation. An audit is characterized jointly by its target, error guarantee, decisive probability, and acquisition cost. The measured regimes show how this description distinguishes useful partial evaluation from an equally inexpensive but unresolved audit. Source-free uniform sampling remains a distinct alternative to source-assisted procurement. Matched-observation comparisons identify inference construction as a further determinant of decision readiness.

Statistical and experimental scope. The guarantees apply to the specified finite rosters, fixed trajectories, and fixed within-arm sampling designs. Total Evaluation Error addresses variability across evaluation configurations (Messing, 2026). The present guarantee is fixed-design rather than anytime-valid, and the sample-size curves characterize the stated tests rather than optimal sample complexity. The allocation and forecast-count sensitivities are post hoc evaluations of existing designs. Each allocation retains its own fixed-rule guarantee. The forecast-count analysis locates discrepancies in the saved predictive distribution; it does not isolate their donor-model cause.

Prospective interpretation and transfer. All new forecasts are conservative, opposite to the preregistered overoptimism hypothesis. The observed task contrast reflects the two pools' risk margins and outcome patterns; it does not establish an intrinsic disadvantage for acceptability certification. Timing measures one selected and one comparator execution per cell on a single workstation. CollabEval is evaluated on a four-row matrix; Prediction-Powered Active Testing and EarlyEval provide related context. New populations, decoding policies, hardware, and adaptive acquisition procedures require their own measurements.

7 CONCLUSION

Decision-ready evaluation connects valid inference with the probability and cost of reaching a conclusion. Our prospective study measures decisive low-cost risk detection, frequently unresolved acceptability audits, and cost reversals when source information must be acquired. Exact decision-power curves and matched-observation comparisons identify how sampling, inference, and information ownership shape these outcomes. Together, they provide an operational basis for deciding what reduced evaluation computation delivers.

ETHICS STATEMENT

The study uses public questions and released models without new human annotation, participant data, or personal records. Finite-roster declarations concern benchmark outcomes, not safety-critical deployment. Model and dataset redistribution retains the original licenses and access conditions.

REPRODUCIBILITY STATEMENT

The appendices specify data partitions, prompts, decoding settings, acquisition rules, exact count inference, and runtime definitions. The numerical materials contain per-question endpoint summaries, original observation masks, frozen forecast counts, timing summaries, and reconstruction scripts for the reported statistics and decision curves. The included CPU commands reconstruct statistics from these recorded observations. The release reconstructs the saved numerical layer; fresh model execution and hardware timing require a separate implementation of the specified inference and acquisition settings.

STATEMENT ON THE USE OF LARGE LANGUAGE MODELS

Generative AI tools assisted with research design, code implementation, mathematical derivations, statistical analysis, literature discovery, and manuscript and figure preparation. Numerical results are grounded in recorded experiments and verification code. The authors take responsibility for the scientific claims, references, and submitted materials.

REFERENCES

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021. URL https://arxiv.org/abs/2107.07511.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*, 2021. URL https://arxiv.org/abs/2110.01052.

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023. doi: 10.1126/science.adi6000.

Kianoosh Ashouritaklimi, Valentin Kilian, Daolang Huang, Tom Rainforth, and François Caron. Prediction-powered active testing. *arXiv preprint arXiv:2607.08347*, 2026. URL https://arxiv.org/abs/2607.08347.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):43:1–43:34, 2021. doi: 10.1145/3478535.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018. URL https://arxiv.org/abs/1803.05457.

William G. Cochran. *Sampling Techniques*. John Wiley & Sons, New York, third edition, 1977. ISBN 9780471162407. URL https://www.wiley-vch.de/en/areas-interest/mathematics-statistics/sampling-techniques-978-0-471-16240-7.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2305.14314.

Adam Fisch, Daniel Deutsch, Joshua Maynez, Alekh Agarwal, Jonathan Berant, William Cohen, Amir Globerson, and Jacob Eisenstein. CollabEval: Statistically efficient collaborative model evaluation via matrix completion. *arXiv preprint arXiv:2607.05046*, 2026. URL https://arxiv.org/abs/2607.05046.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159, 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2009.03300.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021. doi: 10.1214/20-AOS1991.

Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Tom Rainforth. Active testing: Sample-efficient model evaluation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5753–5763, 2021. URL https://proceedings.mlr.press/v139/kossen21a.html.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=iO4LZibEqW.

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, et al. Ministral 3. *arXiv preprint arXiv:2601.08584*, 2026. URL https://arxiv.org/abs/2601.08584.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinyBenchmarks: Evaluating LLMs with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 34303–34326, 2024. URL https://proceedings.mlr.press/v235/maia-polo24a.html.

Solomon Messing. Hidden measurement error in LLM pipelines distorts annotation, evaluation, and benchmarking, 2026. URL https://arxiv.org/abs/2604.11581v6. Version 6.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2381–2391, 2018. doi: 10.18653/v1/D18-1260.

Shubhanshu Shekhar, Ziyu Xu, Zachary C. Lipton, Pierre J. Liang, and Aaditya Ramdas. Risk-limiting financial audits via weighted sampling without replacement. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pp. 1932–1941, 2023. URL https://proceedings.mlr.press/v216/shekhar23a.html.

Yuling Shi, Zhensu Sun, Junsen Dong, Chengcheng Wan, David Lo, and Xiaodong Gu. EarlyEval: Cheaper agent evaluation via early outcome prediction. *arXiv preprint arXiv:2609.02783*, 2026. URL https://arxiv.org/abs/2609.02783.

Philip B. Stark. Conservative statistical post-election audits. *The Annals of Applied Statistics*, 2(2): 550–581, 2008. doi: 10.1214/08-AOAS161.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pp. 4149–4158, 2019. doi: 10.18653/v1/N19-1421.

Abraham Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16 (2):117–186, 1945. doi: 10.1214/aoms/1177731118.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://arxiv.org/abs/2406.01574.

Ian Waudby-Smith and Aaditya Ramdas. Confidence sequences for sampling without replacement. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/e96c7de8f6390b1e6c71556e4e0a4959-Abstract.html.

Skyler Wu, Yash Nair, and Emmanuel J. Candès. Efficient evaluation of LLM performance with statistical guarantees, 2026. URL https://arxiv.org/abs/2601.20251v3. Version 3.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Tijana Zrnic and Emmanuel J. Candès. Cross-prediction-powered inference. *arXiv preprint arXiv:2309.16598*, 2023. URL https://arxiv.org/abs/2309.16598.

A Datasets and sampling design

Question selection. OpenBookQA uses the test split at revision 388097ea7776314e93a529163e0fea805b8a6454. Its 500 question IDs are sorted by a deterministic input-hash key. The first 224, used in development, are excluded; the next 32 form the pilot and the following 192 the audit roster. The remaining 52 are outside the study. CommonsenseQA uses the validation split at revision 94630fe30dad47192a8546eb75f094926d47e155. A separate deterministic hash order selects 224 of its 1,221 questions, with 32 assigned to the pilot and 192 to the audit. Source and target models receive identical selected question IDs within each dataset. Pilot, audit, and previously evaluated development inputs are disjoint.

Public answer keys are used for scoring. Source correctness is available to the forecaster after source execution; target gold labels do not enter target prompts, acceptance rules, or acquisition decisions. The unit of analysis is a question, giving 448 unique new questions shared by the two model families. Forecast simulations and repeated audit samples characterize variation conditional on this question set. Training-set contamination remains outside the experimental partitioning guarantee (Section 6).

Information available to each method. Both datasets' source outputs are available before the acquisition forecasts are fixed, and the forecasts precede all new target execution. A paid-pilot update uses only the corresponding dataset and model family before audit outcomes are observed. The complete target census is available only to the offline evaluator. Source-only acquisition decisions depend on neither unobserved target outcomes nor final target lengths. Pilot costs are assigned only to methods that purchase pilot observations.

Repeated execution and the estimand. Each question–model pair defines one seed-fixed target trajectory. Timed executions reuse the same seeds, and their generated tokens and answer scores agree with the corresponding reference trajectories. Their agreement fixes the outcomes while acquisition cost is measured. Question-sampling and decoding variability are consequently separate from these timing repetitions.

B Model, prompt, and endpoint specification

Source and target instantiations use Qwen/Qwen3-8B at revision b968826d9c46dd6066d109eabc6255188de91218 and mistralai/Ministral-3-8B-Reasoning-2512 at revision 81eaece1948f3875421d9a45bc55487d10e2d894. The source uses BF16 weights; the target uses the corresponding fixed bitsandbytes NF4 representation with BF16 computation and double quantization

disabled. The NF4 representation follows Dettmers et al. (2023); no adapter training is performed. Precision fixes the source and target instantiations for the acquisition study.

Qwen uses its thinking-enabled native template, temperature .6, top-$p = .95$, top-$k = 20$, and min-$p = 0$. Ministral uses its pinned reasoning system prompt represented with native structured thinking content, temperature .7, top-$p = .95$, top-$k = 0$, and min-$p = 0$. Both sample one return sequence, use repetition penalty 1, no repeated-ngram constraint, one beam, and an 8,192-new-token cap. Seeds are deterministic hashes of the fixed study namespace, pool, model, and question ID. Source, target, and timed evaluation receive the same integer seed for an item. Each model representation nevertheless defines its own trajectory.

The user message is:

> Solve the multiple-choice question. Think through the problem before selecting an answer.
> Question:
> `{question}`
> Choices:
> `{label}. {choice text}`
> Choose exactly one of the listed option labels.

Original ordered choices are substituted without outcome-aware reordering. Qwen stops at the first generated `</think>` marker; Ministral at `[/THINK]`. Reaching the cap or EOS before that marker counts as policy noncompletion. After a valid close, the exact prompt-plus-generated-token prefix receives the fixed suffix `\nFinal answer:`. Each continuation consisting of a space plus an option label is scored by its summed conditional token log likelihood. Scoring preserves the generated prefix token sequence. Aggregation and the option softmax use float64; a tie within 10^{-8} follows original option order. The endpoint softmax uses unscaled logits.

The primary acceptance event is valid native completion with finite endpoint scores. The secondary event additionally requires the maximum normalized option probability to be at least .99. Correctness uses the winning label and public gold. An unobserved suffix caused by audit cancellation supplies neither A nor E; it is not assigned $(0, 0)$. A completed but wrong answer supplies $(1, 1)$; deployment noncompletion supplies $(0, 0)$.

Inference uses batch size one, eager attention, deterministic execution settings, and no CPU offload, with Transformers 5.16.1 and bitsandbytes 0.50.2. Wall-clock measurements use an NVIDIA RTX 5090 workstation with 32 GiB VRAM and the same serial inference configuration for each method and its census comparator. The common hardware and serial configuration define the timing comparison.

B.1 Acceptance-policy sensitivity

Table 4 evaluates the two prespecified policies with the original global uniform-sampling test. The confidence filter changes (K, M) and the policy risk and coverage on the same roster, by selecting the answers admitted by the policy. Both CommonsenseQA policies violate the risk requirement; both OpenBookQA policies satisfy it. The .90 decisiveness target retains the same across-pool pattern at $n = 96$, although the OBQA/Qwen probability changes from 63.97% to 77.78%.

Pool	Model	Policy	K	M	C (%)	R (%)	$q(96)$ (%)	$n_{.90}/n^{+}_{.90}$
CSQA	Qwen	Complete	192	40	100.00	20.83	97.51	77/82
CSQA	Qwen	Conf.99	190	39	98.96	20.53	96.40	79/84
CSQA	Ministral	Complete	190	46	98.96	24.21	99.81	59/63
CSQA	Ministral	Conf.99	187	44	97.40	23.53	99.71	61/65
OBQA	Qwen	Complete	192	8	100.00	4.17	63.97	113/122
OBQA	Qwen	Conf.99	191	7	99.48	3.66	77.78	103/113
OBQA	Ministral	Complete	190	11	98.96	5.79	26.83	148/148
OBQA	Ministral	Conf.99	190	11	98.96	5.79	26.83	148/148

Table 4: **Two prespecified acceptance policies under the same audit.** All rows have $N = 192$. Probabilities are exact global-U values, not estimates for S4. The final column gives first and sustained .90 crossings.

C EXACT COUNT INFERENCE AND ITS GUARANTEE

For a binary attribute on N fixed items with unknown total T, a simple random sample without replacement of size n produces $X \sim \mathrm{Hypergeom}(N, T, n)$. At one-sided error allocation η, define

$$L(x) = \min\{t : \mathbb{P}_t(X \geq x) \geq \eta\}, \tag{8}$$

$$U(x) = \max\{t : \mathbb{P}_t(X \leq x) \geq \eta\}, \tag{9}$$

with t restricted to feasible integer totals. Integer numerator/denominator comparisons implement the inclusive tail convention exactly. A full census returns $L = U = x$; an empty sample returns $[0, N]$.

Proposition 1 (Classical fixed-design validity)**.** *For fixed stratification and independent simple random samples within sampled strata, the count procedure in Section 3 yields simultaneous coverage of K and M at least $1 - \delta$. Every definitive statement made on that event is true.*

Proof. For each sampled stratum, hypergeometric tail inversion gives a one-sided noncoverage probability at most $\eta = \delta/(4H)$ for each of the two attributes and two directions. A union bound across at most $4H$ events gives simultaneous stratum-count coverage at least $1 - \delta$. Summing valid stratum bounds covers the population totals; fully enumerated strata have no sampling error. Intersecting with $0 \leq M \leq K \leq N$ preserves coverage. On this event, the acceptable inequalities imply $K > 0$, $M/K \leq \alpha$, and $K/N \geq c_0$; the excess risk inequalities imply $M/K > \alpha$; and the low coverage inequality implies $K/N < c_0$. A false definitive assertion therefore requires count noncoverage. □

The guarantee is conditional on the stated fixed within-arm sampling designs. Conditioning on source information and an independently purchased pilot is valid for the registered fixed within-arm design. For P, condition on the complete potential trajectories: early completers are observed in full, and all survivors have the same continuation probability. Conditional on their realized sample count, the continued survivors form a simple random sample. Audit randomizations are independent of unseen suffix outcomes. This conditioning establishes validity for the specified prefix-continuation design.

Equalities. For $N = 66, n = 53, x = 53$, the upper tail at $T = 63$ is exactly

$$\frac{\binom{63}{53}}{\binom{66}{53}} = \frac{1}{160}.$$

At $\eta = 1/160$, the inclusive rule retains $T = 63$, giving $[63, 66]$. The boundary illustrates why the specified inclusive comparison matters for the discrete confidence interval.

Conditional decision-power calculation. Equation (5) partitions all possible samples into counts of accepted errors, accepted correct answers, and abstentions. The total mass is one. Summing the mass of outcomes whose frozen test returns the census truth gives the finite-pool $q(n)$. We compute all eight curves, covering both answer policies, at every $n = 0, \ldots, 192$. The calculation conditions on the question roster and the realized decoding seed.

Conditional-risk sensitivity. Given a accepted observations in a uniform sample, feasible population acceptance counts are $k \in [a, N - n + a]$. To test $R > .10$, maximize the lower-tail probability of seeing at most e errors under $\mathrm{Hypergeom}(k, \lfloor k/10 \rfloor + 1, a)$ over feasible k. To test $R \leq .10$, maximize the upper tail under $\mathrm{Hypergeom}(k, \lfloor k/10 \rfloor, a)$. Separate coverage tests use the boundary $k_{\min} = \lceil N/5 \rceil$. Each test uses threshold $1/80$ and rejects only for a strictly smaller p-value. This maximizes over the nuisance count rather than plugging in a point estimate. It is a fixed-SRS sensitivity, not an extension to the stratified or adaptive procedures. The primary $n = 96$ results agree with the original exact procedure in all four cells.

D FORECASTING AND ACQUISITION RULES

The development bank contains paired source and target outputs from ARC-Challenge (Clark et al., 2018) and MMLU-Pro (Wang et al., 2024), using one fixed decoding seed and equal total task weights. Within each model family, donors are grouped by source native completion, source correctness, and

whether source length exceeds 1,024. When fewer than eight donors are available, the fixed backoff drops correctness, then length, then uses the model-wide bank. Target completion state, endpoint outcome, target length, and source length are resampled jointly. The predicted target length is the new item's source length times the donor target/source length ratio, rounded and clipped to $[1, 8192]$.

The time predictor uses the fixed family-level median generation-seconds-per-token and measured prefill and endpoint terms from the earlier component study. The coefficients are fixed before exploratory OpenBookQA target evaluation and both independent pools. Each procurement forecast uses 200 worlds and 100 audit randomizations per world. Exact U service curves are also computed for each forecast world. This source-conditioned model supplies procurement forecasts; the observed-data test supplies the audit guarantee.

For U, $n = \lfloor N \min(1, B/\sum_i \widehat{c}_i) \rfloor$. For source-stratified designs, a monotone scalar search sets layer probabilities with ratios $.25$, 1, or 4, capped at one, under the predicted budget; final layer sample sizes are floored and sampled without replacement. The P arm runs all prefixes, then uses a common continuation probability selected from the fixed menu. The menu excludes P when its predicted prefix cost exceeds the budget. The paid arm purchases 32 separate target pilot results, charges their acquisition and update, and chooses a remaining uniform/stratified design using a frozen equal-weight donor/pilot mixture where the fixed fallback permits it.

At each policy and budget, the planner maximizes forecast decisive probability among forecast-affordable designs. A tie within .01 favors lower target acquisition cost and then the fixed order: uniform sampling, the three stratified designs, prefix continuation, paid pilot, and direct census. The pilot purchase decision integrates over possible outcomes before the actual pilot is observed. Both pools' plans are committed before any new target output. Target values and final lengths are hidden from source-only plans.

Prospective directional assessment. The preregistered replication requires the following conditions jointly. Condition (a) required at least .10 forecast overoptimism in exact U decisive probability. Condition (b) required a selected live target cost at most .90 of direct cost and decisive probability at most .80. Condition (c) required decisive probability below the .90 service target under both the primary test and the uniform conditional sensitivity. At least two eligible cells had to span both tasks and model families. All four forecast errors have the conservative sign, opposite to condition (a). The two OBQA cells exhibit the inexpensive-but-inconclusive component; the two CSQA cells exhibit inexpensive decisive detection. The separately prespecified source-construction criterion required cost at least 1.10 times direct execution across tasks; the comparison exceeds that threshold in all four primary cells.

D.1 Forecast count distributions

Table 5 summarizes the original 200 forecast worlds per cell. The two actual CommonsenseQA error counts exceed their respective saved-world maxima. The OpenBookQA counts lie within their forecast ranges and close to their means. These comparisons locate the discrepancy in the forecasted outcome-count distribution. The reported forecast averages the world-specific $q(n)$ values. Because the test is nonlinear, replacing this average by $q(n)$ evaluated at mean counts would change the forecast.

Pool	Model	Predicted K	Actual K	Predicted M	M range	Actual M
CSQA	Qwen	190.04	192	28.81	20–37	40
CSQA	Ministral	186.62	190	25.59	17–36	46
OBQA	Qwen	190.77	192	8.19	4–14	8
OBQA	Ministral	187.04	190	11.80	6–18	11

Table 5: **Frozen outcome-count forecasts and complete-pool observations.** Predicted counts are means over 200 saved worlds; the range is the minimum and maximum predicted accepted-error count. Table 1 reports the corresponding decision-probability forecasts.

E ALL FIXED PROCUREMENT COMPARISONS

Tables 6–7 report all prespecified acquisition designs. Entries are Monte Carlo percentages of correct definitive decisions over the 1,000 audit randomizations; main Table 1 uses exact U probabilities. A sampled estimate can equal 100% while exact power is slightly below one. "Complete" denotes the primary completion policy and "Conf.99" its fixed confidence-filtered sensitivity. Direct is retained as a reference even when it is not forecast-affordable. Cost and guarantee information accompany the rates when comparing acquisition designs.

Model	Policy	β	Chosen	U	S.25	S1	S4	Prefix	Pilot	Census
Qwen	Complete	0.25	U	59.7	7.4	30.3	39.3	80.9†	27.6	100.0
Qwen	Complete	0.50	U	97.4	57.3	93.5	99.9	81.3†	83.7	100.0
Qwen	Complete	0.75	U	100.0	100.0	100.0	100.0	81.7†	100.0	100.0
Qwen	Complete	1.00	U	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Qwen	Conf.99	0.25	U	55.5	6.7	26.4	30.5	79.0†	26.2	100.0
Qwen	Conf.99	0.50	U	96.6	50.3	91.4	99.9	81.3†	79.1	100.0
Qwen	Conf.99	0.75	U	100.0	100.0	100.0	100.0	79.2†	99.8	100.0
Qwen	Conf.99	1.00	U	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Ministral	Complete	0.25	U	76.1	29.5	47.1	42.5	98.6†	44.6	100.0
Ministral	Complete	0.50	S4	100.0	99.1	99.0	98.8	98.3†	77.1	100.0
Ministral	Complete	0.75	S4	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Ministral	Complete	1.00	U	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Ministral	Conf.99	0.25	U	69.1	25.4	37.1	34.7	97.9†	39.0	100.0
Ministral	Conf.99	0.50	S4	99.9	98.5	97.7	96.7	97.9†	70.5	100.0
Ministral	Conf.99	0.75	S4	100.0	100.0	100.0	100.0	100.0	99.9	100.0
Ministral	Conf.99	1.00	U	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Table 6: **CommonsenseQA: all policies, budgets, and designs.** Four budgets and two policies for each model are retained. Values are Monte Carlo percentages; chosen arms were frozen before target results. Component costs for every design are in the machine-readable supplement.

Model	Policy	β	Chosen	U	S.25	S1	S4	Prefix	Pilot	Census
Qwen	Complete	0.25	U	10.8	0.0	0.0	0.0	35.5†	0.0	100.0
Qwen	Complete	0.50	U	65.9	0.0	2.9	0.0	33.2†	21.2	100.0
Qwen	Complete	0.75	S4	100.0	28.3	100.0	100.0	36.5†	73.9	100.0
Qwen	Complete	1.00	U	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Qwen	Conf.99	0.25	U	14.5	0.0	0.0	0.0	37.9†	0.0	100.0
Qwen	Conf.99	0.50	U	79.8	0.0	4.5	0.0	38.1†	25.4	100.0
Qwen	Conf.99	0.75	S4	100.0	36.2	100.0	100.0	40.5†	86.2	100.0
Qwen	Conf.99	1.00	U	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Ministral	Complete	0.25	U	3.1	0.0	0.0	0.0	6.9†	0.0	100.0
Ministral	Complete	0.50	U	27.2	0.0	0.1	0.0	5.8†	11.7	100.0
Ministral	Complete	0.75	U	83.4	50.5	16.3	24.0	5.8	36.1	100.0
Ministral	Complete	1.00	U	100.0	100.0	100.0	100.0	100.0	81.3	100.0
Ministral	Conf.99	0.25	U	3.1	0.0	0.0	0.0	5.2†	0.0	100.0
Ministral	Conf.99	0.50	U	27.2	0.0	0.1	0.0	4.4†	10.4	100.0
Ministral	Conf.99	0.75	U	83.4	53.8	16.7	22.3	5.7	35.6	100.0
Ministral	Conf.99	1.00	U	100.0	100.0	100.0	100.0	100.0	76.3	100.0

Table 7: **OpenBookQA: all policies, budgets, and designs.** A value of 0.0 denotes zero correct definitive returns among 1,000 evaluated randomizations. A dagger marks a design that was ineligible under its forecast cost; its measured return rate is still reported. Census is an unconstrained reference.

Same observation, different inference. Uniform masks receive both global and source-stratified count bounds. Table 8 changes no model output or observed item. The OBQA reduction therefore measures the consequence of stratum-wise error allocation and rectangular aggregation on identical evidence.

Pool	Model	Global interval	Stratified, same mask
CSQA	Qwen	97.4	93.6
CSQA	Ministral	100.0	98.7
OBQA	Qwen	65.9	8.2
OBQA	Ministral	27.2	0.5

Table 8: **Inference alone changes decisive return.** Correct decisive percentages at $\beta = .50$ on the same 1,000 U masks. These Monte Carlo values are distinct from the exact probabilities in Table 1.

E.1 Fixed error-budget allocation sensitivity

This post hoc analysis changes the confidence construction while retaining the original outcomes and all 1,000 observation masks for each design. Let H denote the number of nonempty, non-census strata and let N_h be a stratum's fixed size. Equal allocation gives each of its four one-sided bounds $\delta/(4H)$. Size-weighted allocation gives each bound $\delta N_h/(4N)$, where N includes fully observed strata. Error-weighted allocation assigns $\delta/(6H)$ to each K bound and $\delta/(3H)$ to each M bound. Fully observed strata are exact; an unobserved stratum receives its full feasible count range. Unused error shares are retained.

Each rule assigns total one-sided error at most δ. For uniform observations, conditioning on the vector of stratum sample sizes leaves independent uniform subsets within strata. S4 uses its original source-determined stratum sizes. Thus the tail inversion and union-bound proof in Appendix C applies separately to each fixed allocation. The argument does not require independence between acceptance and error indicators.

Pool	Model	Sample	Equal	Size	Error	Size − Equal	Error − Equal
CSQA	Qwen	U	93.6	91.6	94.3	-2.0	+0.7
CSQA	Ministral	U	98.7	98.7	98.8	+0.0	+0.1
CSQA	Ministral	S4	98.8	98.4	98.8	-0.4	+0.0
OBQA	Qwen	U	8.2	7.1	9.1	-1.1	+0.9
OBQA	Ministral	U	0.5	0.5	0.8	+0.0	+0.3

Table 9: **The global–stratified difference persists under two alternative allocations.** Correct-decision rates are percentages; the last two columns are paired percentage-point changes. U rows apply source-stratified inference to the original uniform masks. S4 retains its original nonuniform sampling design. Each row uses 1,000 masks.

Table 9 shows that the large OpenBookQA reduction remains under both alternatives. The paired transitions distinguish net improvement from samplewise agreement: on OBQA/Qwen, error-weighting gains 29 correct decisions and loses 20, for a net change of nine. Size-weighting on CSQA/Ministral U has one gain and one loss despite its unchanged average.

The discrete mechanism can be inspected on an existing CSQA/Qwen U observation. Its strata have $(N_h, n_h, a_h, e_h) = (136, 71, 71, 5)$ and $(56, 25, 25, 11)$. Equal allocation yields $M_L = 20$ and $K_U = 192$, which satisfies $10M_L > K_U$. Size-weighting changes M_L to 19 and leaves the audit unresolved. The target outcomes are identical. The example uses the first differing mask in the original sequence.

For a single global stratum, size-weighting equals equal allocation. Error-weighting preserves $q(96)$ in all four primary cells while moving the .90 crossings (Table 11). The original equal-allocation curves remain the prospective primary analysis. The numerical materials provide all per-mask decisions and count bounds.

Pool	Model	Sample	Rule	Gains	Losses	Unresolved	Change (pp)	MC SE (pp)
CSQA	Qwen	U	Size	0	20	84	-2.0	0.44
CSQA	Qwen	U	Error	7	0	57	+0.7	0.26
CSQA	Ministral	U	Size	1	1	13	+0.0	0.14
CSQA	Ministral	U	Error	1	0	12	+0.1	0.10
CSQA	Ministral	S4	Size	0	4	16	-0.4	0.20
CSQA	Ministral	S4	Error	0	0	12	+0.0	0.00
OBQA	Qwen	U	Size	0	11	929	-1.1	0.33
OBQA	Qwen	U	Error	29	20	909	+0.9	0.70
OBQA	Ministral	U	Size	0	0	995	+0.0	0.00
OBQA	Ministral	U	Error	3	0	992	+0.3	0.17

Table 10: **Paired gains and losses under identical observations.** Counts are out of 1,000 original masks; gains and losses compare correct decisions with equal allocation. All rows have zero observed false assertions. MC SE is the standard error of the paired correctness difference across masks, not uncertainty over new question pools.

Pool	Model	Common $q(96)$ (%)	Equal $n_{.90}/n^{+}_{.90}$	Error $n_{.90}/n^{+}_{.90}$
CSQA	Qwen	97.51	77/82	72/77
CSQA	Ministral	99.81	59/63	54/59
OBQA	Qwen	63.97	113/122	110/120
OBQA	Ministral	26.83	148/148	146/146

Table 11: **Equal power at one budget can coexist with different threshold crossings.** Exact global-U curves for equal and error-weighted allocations. The $q(96)$ values agree; first and sustained crossings are properties of each fixed test. The crossings describe the fixed statistical rules; execution costs are held unchanged.

F RUNTIME DEFINITIONS AND COST COMPARISONS

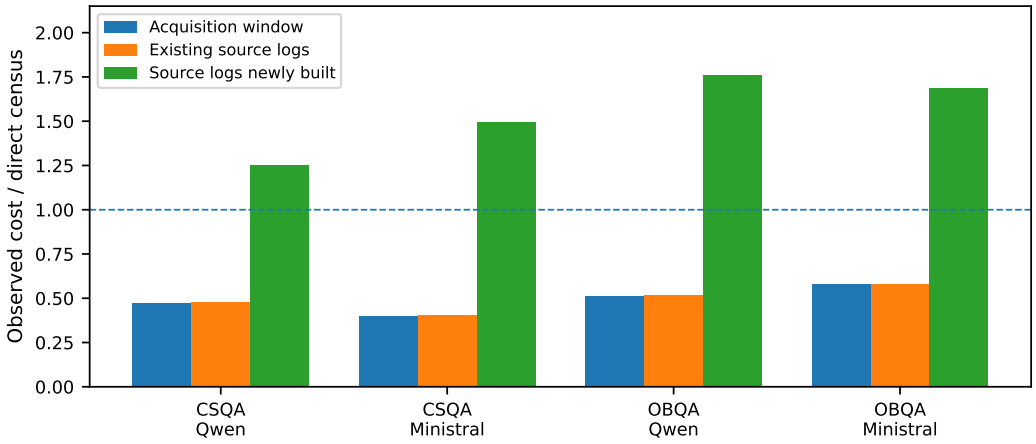

Figure 3: **Asset ownership changes the comparator.** Selected acquisition, existing-log procurement, and newly constructed-source procurement divided by each cell's measured direct-census window. Target initialization is outside the primary resident-model ratio; it is added equally in a separate sensitivity. Decision-service qualification comes from Table 1, not from this bar height.

Existing versus newly constructed source information. The source-built comparison includes 224 source executions per cell, source-model setup, single-budget planning, and the selected target audit. The direct comparator evaluates the 192 target audit questions without constructing source outputs or purchasing a pilot. The pilot inputs explain the difference between 224 and 192. Both

cost accounts condition on an existing historical development bank and exclude experiment-wide preparation; they quantify the stated acquisition designs.

Measured and expected runtime. Eight audit executions and four direct-census executions provide measured wall-clock windows. Primary comparisons begin with a resident target model; an additional sensitivity includes the same measured target initialization on both sides. Single-budget planning is charged separately, and target-pilot generation and updating are charged only to paid-pilot methods.

For uniform sampling of n questions, expected component cost is n/N times the census sum of per-question costs. This expectation is conditional on the fixed trajectories and excludes separately specified setup overhead. Nonuniform designs sum the components they acquire. The measured windows include input preparation, prefill, generation, answer scoring, and output recording. Component sums include the acquired per-question executions but exclude model initialization, planning, and method-specific orchestration outside those components; they are reported as component expectations rather than end-to-end timings.

Shared and research-only costs. Source construction and companion-target observations are charged when required by a method and counted once when shared. Planning and acquisition components are combined only when their timing intervals do not overlap. Full-census construction for evaluation, counterfactual analyses, validation, and figure production are research expenses rather than repeated deployment expenses. These definitions distinguish the information a method must acquire from the cost of evaluating that method experimentally.

F.1 Source-free acquisition and menu comparisons

Source-free U uses its own measured target-execution window and purchases no source logs, pilot, or procurement forecast. The CSQA/Ministral execution times and decision probabilities are paired in the main cost comparison. The source-held and source-built columns in Table 3 then charge only the assets purchased by that pipeline. The two acceptance policies reuse the same generated trajectories, with shared execution charged once.

Two existing offline menu diagnostics answer different questions. The cost-matched gap compares the selected design with original menu entries whose realized expected component cost is no greater. The forecast-budget gap instead admits designs that were feasible under the original predicted costs and budget. Both subtract the selected design's empirical correct-decision rate from the best eligible rate. The cost-matched eligibility set uses realized costs, and both maxima use revealed outcomes: these are offline diagnostics. Table 12 shows the primary cells; all 32 policy–budget cells are retained in the numerical materials.

Pool	Model	Selected	Cost-matched gap (pp)	Forecast-budget gap (pp)
CSQA	Qwen	U	2.50	2.50
CSQA	Ministral	S4	0.00	1.20
OBQA	Qwen	U	0.00	0.00
OBQA	Ministral	U	0.00	0.00

Table 12: **Offline selection gaps within the original procurement menu.** Primary policy and half-cost budget. All rates use the original 1,000 audit randomizations. Cost matching uses expected acquisition components, not a wall-clock ranking with unmeasured overhead set to zero.

G Prediction-assisted baselines

We independently implement the published CollabEval Algorithms 1–3 and covariance expression, using ten-fold cross-fitting, power tuning, and the rank sequence $[1, 2, 4, 8, 16, 16, 16, 16]$, truncated to $[1, 2, 4, 4, 4, 4, 4, 4]$ for the 4×192 matrix. Source rows and sampled entries of the two target rows enter the matrix. Acceptance and accepted-error attributes are analyzed separately before the prescribed risk delta calculation. A fixed anchor mean replaces matrix completion for the simpler

Pool	Model	Estimator	MSE $\times 10^3$	Coverage (%)	Decisive (%)	Undef.
CSQA	Qwen	CollabEval	0.457	99.8	93.7	0
CSQA	Qwen	Anchor mean	0.493	99.7	90.3	0
CSQA	Qwen	Classic IID	0.859	98.9	72.3	0
CSQA	Ministral	CollabEval	0.662	99.9	99.5	0
CSQA	Ministral	Anchor mean	0.658	99.6	99.2	0
CSQA	Ministral	Classic IID	1.024	99.2	93.5	0
OBQA	Qwen	CollabEval	0.141	95.0	93.8	0
OBQA	Qwen	Anchor mean	0.143	95.7	92.5	0
OBQA	Qwen	Classic IID	0.210	96.0	65.9	0
OBQA	Ministral	CollabEval	0.178	97.0	44.1	0
OBQA	Ministral	Anchor mean	0.159	96.9	56.0	0
OBQA	Ministral	Classic IID	0.278	97.1	27.2	0

Table 13: **Estimation and decisive returns with asymptotic intervals.** Primary policy, $\beta = .50$, 1,000 shared masks. MSE is for risk; coverage is the empirical risk-interval coverage. "Decisive" uses the corresponding asymptotic decision rule. Undefined intervals remain in the denominator. Their asymptotic guarantee is distinct from that of the exact finite-population audit.

CrossPPI baseline. Both use the same observed target entries without an additional target training split.

At the main budget all risk intervals are defined, but quarter-budget cases include negative delta variance and unidentifiable risk intervals. For example, OBQA/Qwen has 57 undefined risk intervals for the CollabEval core and 61 for anchor mean out of 1,000; OBQA/Ministral has 29 and 77. These cases remain undefined and are included in interval-coverage denominators. All 96 model/policy/budget/algorithm summaries are retained in the accompanying data.

The comparison distinguishes own-target observations, companion-target observations, source anchors, and shared computation. Joint matrix completion requires observed entries from both target rows; anchor-only estimation does not. Shared costs are counted once. The independent CollabEval implementation evaluates the published algorithms on a four-row matrix. Its comparison reports statistical performance and the information purchased by each method.

G.1 Prediction-assisted decision probabilities and component costs

Table 14: Matched half-pool estimator comparison. All time columns are expected component minutes, not live end-to-end windows. Held adds own-target, companion-target and measured statistical CPU computation; Built also adds source anchors. CollabEval's two model rows share the same joint cost, which must be counted once, not summed. The reported q uses each estimator's asymptotic analysis.

Pool	Model	Estimator	q (%)	Own	Companion	Held	Built
CSQA	Qwen	CollabEval	93.7	77.1	114.6	191.7	518.6
CSQA	Ministral	CollabEval	99.5	114.6	77.1	191.7	518.6
CSQA	Qwen	Anchor mean	90.3	77.1	0.0	77.1	404.0
CSQA	Ministral	Anchor mean	99.2	114.6	0.0	114.6	441.5
CSQA	Qwen	IID	72.3	77.1	0.0	77.1	77.1
CSQA	Ministral	IID	93.5	114.6	0.0	114.6	114.6
OBQA	Qwen	CollabEval	93.8	52.4	81.4	133.9	390.7
OBQA	Ministral	CollabEval	44.1	81.4	52.4	133.9	390.7
OBQA	Qwen	Anchor mean	92.5	52.4	0.0	52.4	309.3
OBQA	Ministral	Anchor mean	56.0	81.4	0.0	81.4	338.3
OBQA	Qwen	IID	65.9	52.4	0.0	52.4	52.4
OBQA	Ministral	IID	27.2	81.4	0.0	81.4	81.4

CollabEval's anchors cost 326.9 component minutes on CSQA and 256.9 on OBQA, shared across its two target rows. Anchor-mean evaluation purchases the same anchors but no companion target; IID uses no anchors. Statistical CPU work is included at its measured duration before rounding the

displayed minutes. These component accounts exclude initialization and unmeasured orchestration and therefore support no end-to-end speedup claim.

Table 15: Original selected-window cost decomposition. Target, Held and Built retain their original ratios; U only is source-free acquisition. All times are minutes.

Pool	Model	Chosen	Census	U only	Target	Held	Built
CSQA	Qwen	73.0	154.5	0.472	0.472	0.475	1.249
CSQA	Ministral	91.5	229.4	0.488	0.399	0.401	1.493
OBQA	Qwen	53.7	105.2	0.511	0.511	0.515	1.758
OBQA	Ministral	94.3	163.2	0.578	0.578	0.581	1.689

H REPRODUCIBILITY RESOURCES

The numerical supplement contains question-level acceptance and error indicators, original observation masks, forecast-world counts, measured-window summaries, and exact uniform-power code. Its documented `quick` and `full` commands reconstruct the supplied tables and decision curves from these records. Complete token trajectories, original timing histories, and the model-execution runner are retained in the separate experimental archive and are not distributed in this numerical supplement. The model, prompt, sampling, and timing specifications above define the conditions for a fresh model-level replication.