PAYING FOR BETTER DATA IN PREDICTION-POWERED TESTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Automated judgments offer a cheaper alternative to expert labels for evaluating large language models, but their accuracy may depend on a provider's unobserved effort. We study whether paying for better judgments can reduce testing costs. Our design combines random audits with bonuses for agreement with experts and chooses the contract and sample size together. We show that the robust likelihood factor maximizing expected log evidence equals a prediction-powered betting factor with jointly optimized predictions and bet size. The test controls false positives in finite samples under arbitrary reporting and attains the optimal leading sample size under the induced review model. With uncertain review costs and accuracy, we characterize contracts and sample sizes ensuring participation, high effort, and prescribed power throughout specified ranges. A finite-sample cost criterion compares candidate designs with direct gold acquisition. Simulations identify when incentives reduce certified costs, and comparisons with prediction-powered baselines clarify the role of inference efficiency. An LLMBar study using observed judgments and assumed costs illustrates how a weaker accuracy gain can favor direct gold acquisition.

1 INTRODUCTION

Consider a research institute comparing two large language models (LLMs). It samples questions from its intended users and asks which model gives the better answer. An expert panel can adjudicate each comparison under a common rubric, but these annotations are expensive. A cheaper review service can ask an LLM judge for a direct comparison or spend additional computation generating evaluation criteria and a reference answer before judging. The institute observes the judgment, but not which procedure produced it. A fixed payment per judgment gives the service little reason to use the more costly procedure.

The institute can encourage careful review by announcing an audit probability and a bonus for agreement with an expert. Audits are drawn after judgments are submitted. Increasing the audit probability or bonus makes careful review more rewarding. Gold labels therefore supply evidence and, through the prospect of an audit, improve judgments on items that are never audited. Better reports may reduce the number of comparisons needed, but audits and bonuses both cost money. The principal must choose the contract and sample size together.

Prediction-powered inference corrects inexpensive predictions using observed outcomes (Angelopoulos et al., 2023a). Here the predictions are supplied by a rational provider whose choice of procedure depends on the contract. We connect the two problems through an exact identity: the likelihood factor against the closest robust null distribution is an optimized prediction-powered betting factor. Its false-positive guarantee allows arbitrary reporting bias and dependence on earlier records. Its power and cost depend on the review procedure induced by payment. The contract and sample size are fixed before data collection, and the test is applied once at the end.

Uncertainty about the provider's technology makes this choice harder. A contract that just covers the estimated additional cost can fail when careful review costs slightly more or improves accuracy slightly less. The provider then rationally chooses the cheaper procedure, leaving a study designed for accurate reports with insufficient power. We consider one provider and two effort levels, and allow the principal to know only ranges for the cost and accuracy of high effort. One contract and one sample size must meet the power requirement throughout these ranges.

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.

The resulting design answers an acquisition question: when can paying for better judgments reduce the cost of a test with finite-sample false-positive control and prescribed power? We characterize contracts that induce participation and high effort throughout the technology range, and establish the leading sample size and cost. For finite studies, we select a contract and a valid betting factor using a power bound, and give the gold-label price at which a candidate becomes cheaper than direct gold acquisition. The experiments establish a regime with certified savings, distinguish this economic gain from the choice of inference procedure, and use LLMBar judgments to examine how the decision depends on an assumed effort-cost gap.

2 Related Work

Inference with inexpensive predictions builds on missing-data and semi-supervised methods (Robins et al., 1994; Zhang et al., 2019). PPI and PPI++ correct predictions with observed outcomes (Angelopoulos et al., 2023a;b); active inference and model-evaluation methods allocate annotation effort (Zrnic and Candès, 2024; Angelopoulos et al., 2025; Fisch et al., 2024; Boyeau et al., 2025). Prediction-powered e-values extend this correction to testing (Csillag et al., 2025), and R-AutoEval+ provides finite-sample reliability while adapting reliance on an automated evaluator (Park et al., 2025). Other testing methods use proxy statistics (Xu et al., 2025; Kuang et al., 2025) or account for uncertain judge accuracy (Feng et al., 2026). We establish the exact audit-correction form of our factor and compare PPI, PPI++, and prediction-powered betting under the same contract; our economic model makes their input quality endogenous.

Economic models of testing study participation and evidence production (Tetenov, 2016; Bates et al., 2022), including risk sensitivity, private types, and strategic sample sizes (Shi et al., 2024; 2025; Chen et al., 2025). Related work studies strategic data sources, endogenous privacy noise, and delegated experimentation (Cai et al., 2015; Wang et al., 2018; Gerardi and Maestri, 2012; McClellan, 2022). Peer prediction and spot-checking incentivize costly review (Miller et al., 2005; Dasgupta and Ghosh, 2013; Gao et al., 2019; Wang et al., 2020; Zarkoob et al., 2020). For AI-assisted review, Bastani and Cachon (2025) study the cost of sustaining human inspection, and Yin et al. (2026) combine sentinel tasks and incentives with active inference. Our audits adjudicate the same population items that enter a test with prescribed false-positive and power levels.

The statistical arguments use least favorable distributions and information projections (Huber and Strassen, 1973; Grünwald et al., 2024; Lardy et al., 2024; Larsson et al., 2025), and their established connection to testing error exponents (Brandão et al., 2020). We use this structure to analyze the cost of incentives and the effect of uncertain review technology.

3 Problem Formulation

The institute is the principal and the review service is the provider. The principal chooses the contract, sample size, and test; the provider chooses whether to participate, which procedure to use, and what to report.

3.1 Population and Review Procedures

Each item contains a sampled question and two LLM answers. Let $Y_i \in \{0, 1\}$ be the judgment an expert panel would give under a prespecified rubric, including a rule for ties. It equals one when the panel favors the new model. Items are independent draws from a fixed population; the gold judgment is observed only if the principal buys it. The population win rate

$$\mu = \Pr(Y_i = 1) \tag{1}$$

is fixed and unknown to both parties. We test $H_0 : \mu \leq \mu_0$ and require power when $\mu \geq \mu_+ = \mu_0 + \delta$, where $\delta > 0$ is the smallest improvement at which power is required.

The provider is risk-neutral and maximizes expected payment minus review cost. It has no separate reward for favoring either model. Costs and payments add across items; earlier actions do not change later contract terms or the available procedures. Effort $e \in \{L, H\}$ selects a low-cost or high-cost review procedure, with $0 \leq c_L < c_H$. The provider chooses before inspecting the item. Review

produces a private binary assessment U_i, summarizing all its item-specific information about Y_i. The provider does not see the gold judgment before reporting. We assume symmetric errors,

$$\Pr(U_i \neq Y_i \mid Y_i, e) = \eta_e, \qquad 0 < \eta_H < \eta_L < \tfrac{1}{2}, \tag{2}$$

with signals independent of earlier items conditional on the current gold label and effort. We first take both costs and error rates as fixed and known. Section 5 allows the principal to know only ranges for the high-effort procedure. The contract affects which procedure is chosen, not its accuracy or cost.

Both parties know a population range

$$\mu \in [a, 1-a], \qquad \eta_L < a < \mu_0 < \mu_+ < 1 - a. \tag{3}$$

This ensures that even low-effort review predicts the gold judgment better than always selecting either model. The optimal response below holds at every win rate in this range, so no specific belief about μ is needed.

3.2 Payments, Audits, and the Provider's Response

A contract $d = (\pi, b, r)$ specifies the audit probability, base payment, and bonus for agreement with the gold label. A gold label costs $g > 0$, and a known cap M bounds payment per completed item:

$$0 < \pi \leq 1, \qquad b \geq 0, \qquad r > 0, \qquad b + r \leq M. \tag{4}$$

These terms remain fixed for all n items. On each item the provider first chooses participation and effort, then reviews and reports $S_i \in \{0, 1\}$. Only afterward does the principal draw $A_i \sim \mathrm{Bernoulli}(\pi)$, independently of the item, report, and past. An audit reveals Y_i, and payment is

$$T_i = b + A_i r \mathbf{1}\{S_i = Y_i\}. \tag{5}$$

Every sampled item is retained. A refusal is recorded as $S_i = \perp$, receives no payment, and is still audited with probability π. The observed record is $O_i = (S_i, A_i, A_i Y_i)$; the last zero is a placeholder when the gold label is unobserved.

The agreement bonus makes passing on the private assessment optimal. The proof is in Appendix A.1.

Proposition 3.1 (Truthful reporting)**.** *Reporting $S_i = U_i$ uniquely maximizes expected payoff on every accepted item, for either effort and after every history. This response holds for every $\mu \in [a, 1-a]$ without knowing μ.*

Consequently, effort e gives expected utility

$$\mathcal{U}(e; d) = b + \pi r(1 - \eta_e) - c_e. \tag{6}$$

The provider participates if its largest utility is nonnegative and chooses high effort precisely when

$$\pi r(\eta_L - \eta_H) > c_H - c_L. \tag{7}$$

The extra expected bonus must exceed the extra review cost. Participation ties select acceptance and effort ties select low effort. These choices do not depend on μ or earlier records, so a fixed contract induces the same response on each item.

A contract is feasible if it obeys (4) and induces participation. Write $\eta(d)$ for its induced error rate. The participation and effort choices, together with truthful reporting, form the equilibrium response. Under that response the records are independent, and the expected expenditure per item is

$$k(d) = g\pi + b + \pi r\{1 - \eta(d)\}. \tag{8}$$

3.3 Error Requirements and Design

The principal fixes d, n, and a test ϕ before collecting data and applies the test once to all n records. We write $\phi \in [0, 1]$ to allow randomized tests. Power and cost are evaluated at the provider's equilibrium response, with expectation denoted by $\mathbb{E}^d_\mu$.

False-positive control also protects against departures from that response. Let σ specify the joint law of reports and gold labels. Conditional on earlier records, the current gold label must remain Bernoulli with mean μ, and the audit must remain an independent post-report draw with probability π. Within these restrictions, reports and refusals may have arbitrary accuracy, bias, or dependence on earlier records. These laws form the robustness class; denote their expectations by $\mathbb{E}^d_{\mu,\sigma}$. For $\alpha, \beta \in (0,1)$ with $\alpha + \beta < 1$, require

$$\sup_{a \le \mu \le \mu_0} \sup_{\sigma} \mathbb{E}^d_{\mu,\sigma} \phi \le \alpha, \tag{9}$$

$$\inf_{\mu_+ \le \mu \le 1-a} \mathbb{E}^d_{\mu} \phi \ge 1 - \beta. \tag{10}$$

The design minimizes expected expenditure,

$$C^\star = \inf_{n,d,\phi} nk(d), \quad n \in \mathbb{N},\ d \text{ feasible}, \quad \phi \text{ satisfies (9)–(10)}. \tag{11}$$

The gold-only alternative buys labels directly, with no provider payment, and uses the exact one-sided binomial test under the same error requirements, randomizing at the rejection boundary when needed.

4 Prediction-Powered Testing With Audited Reports

4.1 Correcting Predictions With Audits

For a report S, let m_S predict the gold judgment. The audited residual corrects prediction error:

$$W = m_S + \frac{A}{\pi}(Y - m_S), \qquad e_{\lambda,m}(O) = 1 + \lambda(W - \mu_0). \tag{12}$$

Averaging over the independent audit gives $\mathbb{E}[W \mid S, Y] = Y$. Consequently, any fixed $\lambda \ge 0$ and prediction pair $m = (m_0, m_1)$ for which the factors are positive on the completed-record support satisfy $\mathbb{E}[e_{\lambda,m} \mid \text{past}] \le 1$ under every allowed null reporting law. Set the factor to zero for a refusal. Repeated conditional expectation makes their product an *e-value*, a nonnegative statistic with null expectation at most one. Markov's inequality gives the fixed-sample test

$$Z_n = \sum_{i=1}^{n} \log e_{\lambda,m}(O_i), \qquad \phi_n = \mathbf{1}\{Z_n \ge \log(1/\alpha)\}. \tag{13}$$

Here and below $\log$ is logarithm to base e. This construction controls false positives regardless of the reports' accuracy or bias. Predictions, betting coefficients, and sample size are chosen before collecting the study data.

4.2 Choosing the Factor and Sample Size

The contract supplies a model for tuning the factor. Fix a feasible d with induced error $\eta = \eta(d)$. At the target win rate μ_+, let $p_+(o)$ be the record distribution. For a null comparison, use win rate μ_0 and allow both conditional reporting probabilities $x_y = \Pr(S = 1 \mid Y = y)$ to vary; write q_x for this law. There are two unaudited records, $(0,0,0)$ and $(1,0,0)$, and four audited report–gold pairs, $(0,1,0)$, $(1,1,0)$, $(0,1,1)$, and $(1,1,1)$. With $\theta_+ = (1-\mu_+)\eta + \mu_+(1-\eta)$ and $\theta_0 = (1-\mu_0)x_0 + \mu_0 x_1$, their probabilities are as follows.

Write Ω_d for the completed-record support, omitting unaudited records when $\pi = 1$. Minimize the Kullback–Leibler divergence

$$H_d(x) = \sum_{o \in \Omega_d} p_+(o) \log \frac{p_+(o)}{q_x(o)}, \qquad x^\star_d = \arg\min_{x \in (0,1)^2} H_d(x), \qquad J(d) = H_d(x^\star_d), \tag{14}$$

and define

$$e_d(o) = p_+(o)/q_{x^\star_d}(o), \qquad \ell_d(o) = \log e_d(o). \tag{15}$$

The optimization is strictly convex in two variables, with a unique interior minimum and $J(d) > 0$. Table 1 calculates each audited entry by multiplication and each unaudited entry by summing over the unobserved gold label.

Table 1: Record probabilities at the target improvement, p_+, and at the null win rate with reporting probabilities x, q_x.

Record (S, A, AY)	$p_+(o)$	$q_x(o)$
$(0,0,0)$	$(1-\pi)(1-\theta_+)$	$(1-\pi)(1-\theta_0)$
$(1,0,0)$	$(1-\pi)\theta_+$	$(1-\pi)\theta_0$
$(0,1,0)$	$\pi(1-\mu_+)(1-\eta)$	$\pi(1-\mu_0)(1-x_0)$
$(1,1,0)$	$\pi(1-\mu_+)\eta$	$\pi(1-\mu_0)x_0$
$(0,1,1)$	$\pi\mu_+\eta$	$\pi\mu_0(1-x_1)$
$(1,1,1)$	$\pi\mu_+(1-\eta)$	$\pi\mu_0 x_1$

The projection and prediction-correction constructions are exactly the same when tuned for the same target law.

Proposition 4.1 (Optimal prediction-powered factor)**.** *For $0 < \pi < 1$, e_d has the form (12) with $\lambda > 0$ and $0 < m_0 < m_1 < 1$. It uniquely maximizes $\mathbb{E}_{p_+} \log e$ over nonnegative factors valid under every allowed null, up to equality on the completed-record support. At $\pi = 1$, it is the Bernoulli likelihood ratio and m cancels.*

Appendix E.1 proves the identity and optimality. The strict comparison concerns expected log evidence: a different factor can have higher power at a particular finite sample size. We therefore also compare a small prespecified set of positive factors in (12), always including e_d.

For any such factor, put $\ell(o) = \log e(o)$ and let $p_{\mu,\eta}$ be its record law under the induced procedure. Define

$$\psi_{d,e}(s) = \max_{\mu \in \{\mu_+, 1-a\}} \log \sum_{o \in \Omega_d} p_{\mu,\eta}(o) e^{-s\ell(o)}, \qquad s > 0. \tag{16}$$

The maximum is at an endpoint because each moment is affine in μ. A Chernoff bound gives nonrejection probability at most $\exp\{s \log(1/\alpha) + n\psi_{d,e}(s)\}$. We use the smallest positive integer $N(d, e)$ such that

$$\inf_{s>0} \{s \log(1/\alpha) + N(d,e)\psi_{d,e}(s)\} \le \log \beta. \tag{17}$$

The inner optimization is one-dimensional and convex; optimization followed by integer checks determines $N(d, e)$. Discard factors with no finite power bound. The projection factor always supplies one. Numerical checks and an alternative Hoeffding bound appear in Appendix C.

Algorithm 1 Contract and factor selection before data collection

Require: Feasible contract menu, valid positive factor candidates, α, β, δ.
1: **for** each contract d **do**
2: Determine the induced procedure; include its projection factor from (14).
3: For each candidate factor, compute $N(d, e)$ from (17).
4: Retain the factor with the smallest certified cost $N(d, e)k(d)$.
5: **end for**
6: Select the least expensive retained design or direct gold-only testing.
7: Fix the selected design, collect all items, and apply its test once.

For a fixed contract, the smallest sample size of any valid, powerful test is $\log(1/\alpha)/J(d) + O(\sqrt{\log(1/\alpha)})$ as $\alpha \downarrow 0$. Thus the projection factor supplies the optimal leading term, while finite-sample factor selection can improve the implemented power bound. Appendix A.7 gives the known-technology result; the next section extends it to uncertain accuracy.

5 Contracts When Review Technology Is Uncertain

The automated procedure may be well understood while the cost and accuracy of additional checking are less certain. We now keep c_L and η_L known and let the principal know only that

$$(c_H, \eta_H) \in \Theta = [\underline{c}_H, \overline{c}_H] \times [\underline{\eta}_H, \overline{\eta}_H], \qquad c_L < \underline{c}_H \le \overline{c}_H, \quad 0 < \underline{\eta}_H \le \overline{\eta}_H < \eta_L < a. \tag{18}$$

The provider knows its actual costs and error rates. These are fixed throughout the study. The intervals are supplied before design; their construction from preliminary data is outside the model. A point interval recovers the known-technology setting. We compare contracts that induce the same effort level for every technology in Θ, and also allow the principal to buy gold labels directly.

The principal chooses one contract, one sample size, and one test. Participation and power must hold throughout Θ, and the objective is the largest expected expenditure over this set. False-positive control remains (9). In particular, misspecifying a technology range can affect power and cost, but does not invalidate the test's false-positive guarantee.

Put $t = \pi r$. The hardest provider to induce into high effort has the largest cost and error rate, so uniform high effort requires

$$t > \overline{h} := \frac{\overline{c}_H - c_L}{\eta_L - \overline{\eta}_H}, \qquad b \geq [\overline{c}_H - (1 - \overline{\eta}_H)t]_+, \tag{19}$$

where $[z]_+ = \max\{z, 0\}$. The second inequality ensures participation. The effort inequality is strict because effort ties select low effort. These conditions give a direct way to build candidate contracts: choose t just above $\overline{h}$, set the smallest allowed base payment, and retain audit probabilities satisfying $b + t/\pi \leq M$.

5.1 Power and Cost Throughout the Range

For a contract guaranteeing high effort, construct the score using $\eta = \overline{\eta}_H$ in Table 1. Write $\overline{J}(d)$ for the resulting divergence. For finite-sample power, replace the maximum in (16) by a maximum over the four pairs

$$(\mu, \eta_H) \in \{\mu_+, 1 - a\} \times \{\underline{\eta}_H, \overline{\eta}_H\}. \tag{20}$$

The record probabilities are affine in either parameter with the other fixed, so these four calculations cover the entire rectangle. For each candidate factor, Algorithm 1 minimizes $N(d, e)\overline{k}(d)$, where $\overline{k}(d) = g\pi + b + \pi r(1 - \underline{\eta}_H)$ is the largest per-item expenditure. A uniformly low-effort contract uses η_L and its unchanged per-item cost instead.

For the projection factor, the least accurate procedure determines the leading sample size even though its actual accuracy is unknown. The following result is proved in Appendix E.

Theorem 5.1 (Testing under uncertain accuracy)**.** *For a fixed contract inducing high effort throughout Θ, the score constructed at $\overline{\eta}_H$ satisfies*

$$\mathbb{E}_{\mu, \eta_H} \ell_d(O) \geq \overline{J}(d) > 0, \quad \mu \in [\mu_+, 1 - a], \quad \eta_H \in [\underline{\eta}_H, \overline{\eta}_H]. \tag{21}$$

The test remains valid under arbitrary reporting, and the four-corner Chernoff bound ensures power at least $1 - \beta$ throughout the range. For fixed β, δ, d, Θ, the smallest sample size of any test satisfying these requirements obeys

$$n^\star_{\mathrm{unif}}(d) = \frac{\log(1/\alpha)}{\overline{J}(d)} + O\Big(\sqrt{\log(1/\alpha)}\Big), \qquad \alpha \downarrow 0. \tag{22}$$

For a fixed finite menu of contracts that each induce one effort throughout Θ, the minimum worst expected expenditure is

$$\overline{C}^\star_{\mathcal{D}} = \log(1/\alpha) \min_{d \in \mathcal{D}} \frac{\overline{k}(d)}{\overline{J}(d)} + O\Big(\sqrt{\log(1/\alpha)}\Big), \tag{23}$$

using $J(d)$ and $k(d)$ for uniformly low-effort contracts.

The largest payment occurs at the most accurate high-effort procedure, which earns the most agreement bonuses. The leading sample-size requirement is determined by the least accurate one. Both enter the worst expenditure because the principal fixes the same sample size for all technologies.

5.2 Comparing Costs for Small Improvements

More auditing raises expenditure per item but can reduce the number of items needed. To understand this tradeoff, take $\mu_0 = 1/2$ and consider small improvements δ. The word *local* refers to alternative win rates approaching the null boundary as δ decreases.

First consider the uncertainty left in one item after audit correction. At the boundary win rate $1/2$, use conditional-mean predictions in (12) and the least accurate high-effort procedure, with error $\overline{\eta}_H$. The corrected observation has variance

$$\overline{V}(d) = \frac{1}{4} + \left(\frac{1}{\pi} - 1\right) v, \qquad v = \overline{\eta}_H(1 - \overline{\eta}_H). \tag{24}$$

Full auditing gives variance $1/4$, that of the gold label. The extra term measures uncertainty from missing gold labels; it decreases with more auditing or more accurate reports.

Smaller variance means fewer items are needed. Combining Theorem 5.1 with the small-improvement information formula in Appendix B gives

$$\overline{k}(d) n^{\star}_{\mathrm{unif}}(d) \;\approx\; \frac{2\log(1/\alpha)}{\delta^2}\,\overline{F}(d;g), \qquad \overline{F}(d;g) = \overline{k}(d)\overline{V}(d). \tag{25}$$

The factor $2\log(1/\alpha)/\delta^2$ is shared by all contracts. The *local cost coefficient* $\overline{F}$ isolates the contract's contribution: spending per item multiplied by the variance that determines how many items are needed. Precisely, the ratio of the two sides tends to one as $\alpha \downarrow 0$ and then $\delta \downarrow 0$, with β, the contract, and the technology range fixed. Gold-only testing has coefficient $g/4$, so $4\overline{F}/g$ is the corresponding leading cost ratio to gold.

We can now minimize this coefficient over high-effort contracts. The incentive and participation conditions determine two limits: the smallest worst-case provider payment per item, w, and the lower audit fraction imposed by the payment cap, ρ. Feasible contracts can approach payment w at every $\pi > \rho$. Appendix F gives their formulas.

Theorem 5.2 (Local cost of guaranteeing high effort)**.** *Suppose a contract guarantees participation and high effort throughout Θ. The infimum of the local cost coefficient over all such contracts is*

$$\overline{F}^{\star}_H(g) = \inf_{\rho < \pi \le 1} (g\pi + w)\left\{\frac{1}{4} + \left(\frac{1}{\pi} - 1\right) v\right\}. \tag{26}$$

There is an explicit gold-price threshold g_G such that $\overline{F}^{\star}_H(g) < g/4$ exactly when $g > g_G$. Enlarging Θ cannot lower $\overline{F}^{\star}_H(g)$ or g_G.

Appendix F gives the minimizing audit rate and g_G. The next section makes the comparison at the chosen finite error levels.

5.3 A Finite-Sample Cost Criterion

For an actual study, use the certified number of items rather than its local approximation. Fix α, β, δ, the technology range, and a finite list of contract–factor pairs. Their certified sample sizes $N(d, e)$ do not depend on g. Let N_G be the exact gold-only sample size and w_d the largest expected provider payment per item.

A candidate costs $N(d, e)(g\pi_d + w_d)$, whereas direct gold costs gN_G. Savings therefore require fewer than N_G expected audits, with the avoided gold expenditure covering all provider payments.

Hence some candidate is cheaper than gold exactly when

$$g > g_{\mathrm{cert}} := \min_{(d,e):\, N(d,e)\pi_d < N_G} \frac{N(d,e) w_d}{N_G - N(d,e)\pi_d}, \qquad \min \varnothing = +\infty. \tag{27}$$

This threshold is exact for the specified menu and power bound. An infinite g_{cert} rules out savings for this menu, not for all valid tests.

6 Numerical Experiments

6.1 Local Predictions and Finite-Sample Decisions

We use $\alpha = 0.05$, $\beta = 0.2$, $\mu_0 = 0.5$, $\delta = 0.1$, $a = 0.25$, $c_L = 0.02$, nominal $c^0_H = 0.04$, $\eta_L = 0.20$, $M = 1$, and $g = 10$. Nominal high-effort errors are $\eta^0_H = 0.01$ and 0.05. An uncertainty width $\omega \in \{0, 0.01, 0.05, 0.10, 0.20\}$ specifies

$$c_H \in [c_L + (1-\omega)(c^0_H - c_L),\; c_L + (1+\omega)(c^0_H - c_L)], \quad \eta_H \in [\eta^0_H, \eta^0_H + \omega(\eta_L - \eta^0_H)]. \tag{28}$$

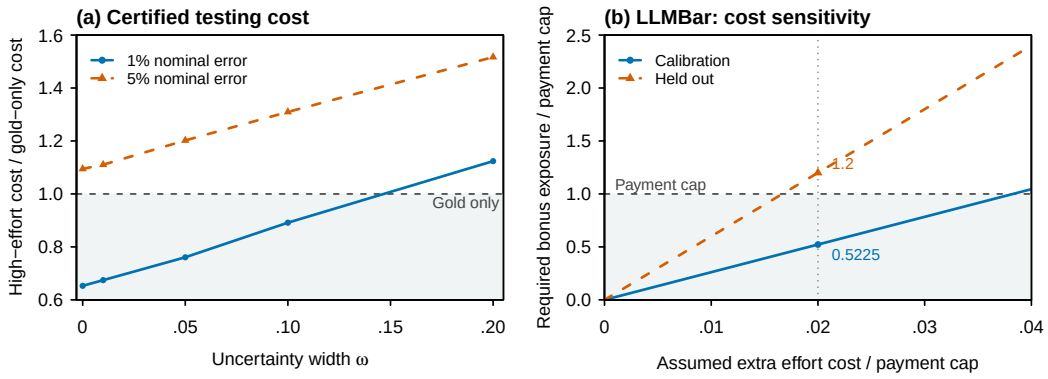

Figure 1: Two acquisition decisions. Left: certified high-effort cost relative to direct gold, across all ten synthetic settings; costs below one give savings. Right: bonus exposure required at balanced prevalence, computed from measured LLMBar accuracy gains while varying the assumed extra effort cost. High effort requires exposure strictly above each curve and at most the cap; this panel does not compare total testing costs. The marked gap 0.02 is a scenario input.

Thus width is measured relative to the extra cost and accuracy gain. We search audit probabilities $0.02, 0.04, \ldots, 1$; high-effort contracts use $\pi r = 1.0001\bar{h}$ and the smallest feasible base payment. Direct gold costs 1,530 for 153 labels and has exact power 0.80075, with no provider payment. All reported simulation frequencies use 20,000 repetitions; the code and appendix give seeds and intervals.

The local gold-cost thresholds range from 0.156 to 0.641, all below $g = 10$, whereas finite-sample savings occur in only part of the grid (Figure 1, left). Replacing the local information approximation with exact KL information raises the leading cost ratio. The remaining increase to the projection-based finite design is larger: its sample size is 3.08–3.73 times $\log(1/\alpha)/\bar{J}$. This measures departure from the small-α limit, not solely looseness of the Chernoff inequality. The observed powers near 0.97 also show conservatism relative to the requested 0.8. The local information approximation overstates exact information by 20.3% in the 5% point-error setting and 60.7% in the 1% setting. Appendix H.1 tabulates both thresholds and all cost comparisons, and verifies the stated order of limits.

Finite-sample factor selection improves the implemented design. Candidates include the projection factor, its betting coefficients scaled by $0.5, 0.75, 1.25, 1.5$ when admissible, and raw-report betting with either a tuned coefficient or half its admissible maximum. Nine configurations select the 0.75-scaled projection factor; one selects the 0.5 scale. The cost criterion (27) selects high effort at nominal error 0.01 and widths through 0.10, saving 34.7%, 32.5%, 23.9%, and 10.9%; it selects gold in the remaining six settings. All selected high-effort designs have certified power at least 0.8; simulated upper-corner power is 0.964–0.969.

Two secondary diagnostics clarify these comparisons. In the projection-only 5% point-error run, the projection design costs 2,113 and has estimated power 0.9655. Matching that estimated power requires 296 direct-gold labels, costing 2,960, so the comparison changes to 28.6% savings. This is a pointwise matched-power comparison, not the nominal 80%-power guarantee. The information required for arbitrary-report validity is also substantial: assuming the modeled report law in the null increases information by factors 2.19 and 4.52 for the known-technology high- and low-effort contracts. Appendix D.7 gives these calculations.

6.2 Prediction-Powered Baselines and LLMBar

We implement the published PPI and PPI++ mean-estimation Wald tests, and prediction-powered e-values with raw, conditional-mean, or jointly tuned predictions. All methods share the same reports and audits at a given budget. Joint tuning is algebraically identical to the projection factor. In the two synthetic point-error settings it improves expected log growth over raw-report betting by 5.8%

Table 2: LLMBar replay under the low-effort contract, $\pi = 0.06$. Empirical budgets target 80% power and use a separate Monte Carlo check; gold uses its exact design. Wald and betting methods have different validity guarantees. Expected costs at $\mu = 0.6$ are conditional on the public calibration data.

Method	Items	Null rejection	Power	Cost
PPI (Wald)	1,600	0.0518	0.8125	992
PPI++ (Wald)	1,300	0.0589	0.8165	806
Joint log-growth betting	2,600	0.0073	0.8223	1,612
Finite-menu betting	2,900	0.0022	0.8089	1,798
Raw-report betting	2,900	0.0080	0.8106	1,798
Direct gold (exact)	153	0.0500	0.8007	1,530

and 7.5%, but does not uniformly improve finite-sample power. The Wald tests have higher power at common budgets and can exceed the nominal size; full comparisons appear in Appendix H.2.

We use all 419 LLMBar items and released GPT-4 Vanilla+Rules and Rules+Metrics+Reference judgments (Zeng et al., 2024). The second procedure generates evaluation criteria and a reference answer before judging. An item split assigns 209 items to calibration and 210 to evaluation, keeping both output orders together. The measured error rates change from 17.22% versus 13.40% in calibration to 13.81% versus 12.14% in evaluation: the extra procedure's accuracy advantage falls from 3.83 to 1.67 percentage points.

The economic inputs are assumed scenarios: $c_L = 0.02$, $c_H = 0.04$, $g = 10$, and $M = 1$. LLMBar supplies no measurements of these costs or of responses to payment. Figure 1 (right) varies the assumed cost gap while holding the empirical accuracy gains fixed. At balanced prevalence, dividing the assumed gap 0.02 by those gains gives required bonus exposures 0.5225 and 1.2. The latter exceeds the cap, ruling out uniformly high effort in this scenario. The calibrated certified-cost comparison selects direct gold. A smaller assumed gap changes the incentive calculation; a total-cost advantage also requires sufficient savings in gold labels. Prospective costs condition on the available public calibration judgments.

We then resample held-out judgments within gold-label strata to impose controlled prevalences 0.5 and 0.6, preserving the observed asymmetric errors. These are benchmark populations, not measured deployment win rates. Contracts and scores use calibration data only. Table 2 compares descriptive budgets chosen for approximately 80% power and checked with independent Monte Carlo replications. The betting tests retain finite-sample null control; PPI and PPI++ have asymptotic guarantees. Their empirical budgets are not uniform power certificates. Appendix H.3 gives the protocol, pilot uncertainty diagnostic, and complete results.

7 Discussion

The contribution is a joint design of data acquisition and testing when prediction quality responds to payment. A useful contract must make better review privately worthwhile, and its reports must save enough gold labels to cover provider payments. Our results connect these two requirements while retaining false-positive control under arbitrary reporting. Under the stated effort model and technology ranges, the synthetic experiments demonstrate that designs in the stated menu can reduce certified testing cost by 10.9–34.7%.

The projection identity gives optimal expected log evidence and leading sample size, but a smaller bet can improve finite-sample power certification. We therefore select factors by certified cost. The higher power of PPI and PPI++ at common budgets motivates tighter finite-sample tests. These comparisons concern inference from a given source; the contract analysis determines which source the provider supplies and whether it is worth purchasing.

On LLMBar, the accuracy gain shrinks on held-out items, making the assumed high-effort cost unsustainable under the payment cap. Deployment requires measuring costs and responses to payments, and accounting for estimation error in both review procedures.

AI USE STATEMENT

Generative AI tools were used to aid and polish writing, retrieve and discover related literature, and assist in proving mathematical claims. This assistance included revising the exposition, identifying potentially relevant references, and developing and checking proof arguments. The authors take responsibility for the final manuscript, including its references, mathematical claims, and proofs.

REPRODUCIBILITY STATEMENT

Sections 3–5 state the assumptions, procedure, and guarantees. The appendices contain the proofs, algorithms, parameter grids, and Monte Carlo intervals. The source package includes executable R scripts, saved numerical results, and the frozen LLMBar labels and judge outputs with their license and checksums. A reproduction guide lists commands for the original experiments, threshold diagnostics, prediction-powered baselines, finite-factor selection, and LLMBar replay. All reported Monte Carlo comparisons were executed with 20,000 repetitions. The LLMBar study reuses public judgments and models costs and responses; it does not measure responses to actual payments.

REFERENCES

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

Anastasios N. Angelopoulos, John C. Duchi, and Tijana Zrnic. PPI++: Efficient prediction-powered inference. *arXiv preprint arXiv:2311.01453*, 2023b. arXiv:2311.01453.

Anastasios N. Angelopoulos, Jacob Eisenstein, Jonathan Berant, Alekh Agarwal, and Adam Fisch. Cost-optimal active AI model evaluation. *arXiv preprint arXiv:2506.07949*, 2025. arXiv:2506.07949.

Hamsa Bastani and Gerard P. Cachon. The human–AI contracting paradox. *The Wharton School Research Paper*, 2025. doi:10.2139/ssrn.5962739.

Stephen Bates, Michael I. Jordan, Michael Sklar, and Jake A. Soloff. Principal-agent hypothesis testing. *arXiv preprint arXiv:2205.06812*, 2022. Revised 2024. arXiv:2205.06812.

Pierre Boyeau, Anastasios Nikolas Angelopoulos, Tianle Li, Nir Yosef, Jitendra Malik, and Michael I. Jordan. AutoEval done right: Using synthetic data for model evaluation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 5276–5290, 2025. Proceedings paper.

Fernando G. S. L. Brandão, Aram W. Harrow, James R. Lee, and Yuval Peres. Adversarial hypothesis testing and a quantum Stein's lemma for restricted measurements. *IEEE Transactions on Information Theory*, 66(8):5037–5054, 2020. doi:10.1109/TIT.2020.2979704.

Yang Cai, Constantinos Daskalakis, and Christos Papadimitriou. Optimum statistical estimation with strategic data sources. In *Proceedings of the 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 280–296, 2015. Proceedings paper.

Yatong Chen, Safwan Hossain, and Yiling Chen. Strategic hypothesis testing. In *Advances in Neural Information Processing Systems*, volume 38, 2025. Proceedings paper.

Daniel Csillag, Claudio Jose Struchiner, and Guilherme Tegoni Goedert. Prediction-powered e-values. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 11493–11514, 2025. Proceedings paper.

Anirban Dasgupta and Arpita Ghosh. Crowdsourced judgement elicitation with endogenous proficiency. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 319–330, 2013. doi:10.1145/2488388.2488417.

Chen Feng, Minghe Shen, Ananth Balashankar, Carsten Gerner-Beuerle, and Miguel R. D. Rodrigues. Noisy but valid: Robust statistical evaluation of LLMs with imperfect judges. In *International Conference on Learning Representations*, 2026. arXiv:2601.20913.

Adam Fisch, Joshua Maynez, R. Alex Hofer, Bhuwan Dhingra, Amir Globerson, and William W. Cohen. Stratified prediction-powered inference for effective hybrid evaluation of language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 111489–111514, 2024. doi:10.52202/079017-3541.

Xi Alice Gao, James R. Wright, and Kevin Leyton-Brown. Incentivizing evaluation with peer prediction and limited access to ground truth. *Artificial Intelligence*, 275:618–638, 2019. doi:10.1016/j.artint.2019.03.004.

Dino Gerardi and Lucas Maestri. A principal–agent model of sequential testing. *Theoretical Economics*, 7(3):425–463, 2012. doi:10.3982/TE914.

Peter Grünwald, Rianne de Heide, and Wouter Koolen. Safe testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(5):1091–1128, 2024. doi:10.1093/jrsssb/qkae011.

Peter J. Huber and Volker Strassen. Minimax tests and the Neyman–Pearson lemma for capacities. *The Annals of Statistics*, 1(2):251–263, 1973. doi:10.1214/aos/1176342363.

Qi Kuang, Bowen Gang, and Yin Xia. Active hypothesis testing under computational budgets with applications to GWAS and LLM. *arXiv preprint arXiv:2512.01423*, 2025. arXiv:2512.01423v2 (revised April 2026).

Tyron Lardy, Peter Grünwald, and Peter Harremoës. Reverse information projections and optimal e-statistics. *IEEE Transactions on Information Theory*, 70(11):7616–7631, 2024. doi:10.1109/TIT.2024.3444458.

Martin Larsson, Aaditya Ramdas, and Johannes Ruf. The numeraire e-variable and reverse information projection. *The Annals of Statistics*, 53(3):1015–1043, 2025. doi:10.1214/24-AOS2487.

Andrew McClellan. Experimentation and approval mechanisms. *Econometrica*, 90(5):2215–2247, 2022. doi:10.3982/ECTA17021.

Nolan Miller, Paul Resnick, and Richard Zeckhauser. Eliciting informative feedback: The peer-prediction method. *Management Science*, 51(9):1359–1373, 2005. doi:10.1287/mnsc.1050.0379.

Sangwoo Park, Matteo Zecchin, and Osvaldo Simeone. Adaptive prediction-powered AutoEval with reliability and efficiency guarantees. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi:10.52202/085713-5303.

James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994. doi:10.1080/01621459.1994.10476818.

Flora C. Shi, Stephen Bates, and Martin J. Wainwright. Sharp results for hypothesis testing with risk-sensitive agents. *arXiv preprint arXiv:2412.16452*, 2024. arXiv:2412.16452.

Flora C. Shi, Martin J. Wainwright, and Stephen Bates. Instance-adaptive hypothesis tests with heterogeneous agents. *arXiv preprint arXiv:2510.21178*, 2025. arXiv:2510.21178.

Aleksey Tetenov. An economic theory of statistical testing. CeMMAP Working Paper CWP50/16, 2016. doi:10.1920/wp.cem.2016.5016.

Weina Wang, Lei Ying, and Junshan Zhang. The value of privacy: Strategic data subjects, incentive mechanisms, and fundamental limits. *ACM Transactions on Economics and Computation*, 6(2):Article 8, 2018. doi:10.1145/3232863.

Wanyuan Wang, Bo An, and Yichuan Jiang. Optimal spot-checking for improving the evaluation quality of crowdsourcing: Application to peer grading systems. *IEEE Transactions on Computational Social Systems*, 7(4):940–955, 2020. doi:10.1109/TCSS.2020.2998732.

Ziyu Xu, Catherine Wang, Larry Wasserman, Kathryn Roeder, and Aaditya Ramdas. Active multiple testing with proxy p-values and e-values. *arXiv preprint arXiv:2502.05715*, 2025. arXiv:2502.05715.

Qichuan Yin, Ziwei Su, and Shuangning Li. Overcoming the incentive collapse paradox. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. arXiv:2603.27049.

Hedayat Zarkoob, Hu Fu, and Kevin Leyton-Brown. Report-sensitive spot-checking in peer-grading systems. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems*, pages 1593–1601, 2020. Proceedings paper.

Anru Zhang, Lawrence D. Brown, and T. Tony Cai. Semi-supervised inference: General theory and estimation of means. *The Annals of Statistics*, 47(5):2538–2566, 2019. doi:10.1214/18-AOS1756.

Tijana Zrnic and Emmanuel Candès. Active statistical inference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 62993–63010, 2024. Proceedings paper.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. Evaluating large language models at evaluating instruction following. In *International Conference on Learning Representations*, 2024. arXiv:2310.07641.

SUPPLEMENTARY MATERIAL

Appendix A proves the reporting and testing results. Appendix B derives the local cost coefficient and compares low and high effort when review technology is known. Appendix C describes how to compute the test factors and sample-size bounds, and Appendix D provides the known-technology simulation settings and additional results.

Appendix E proves the prediction-powered identity and the testing guarantees under uncertain accuracy. Appendix F characterizes contracts and cost thresholds under uncertain review technology. Appendix G reports reference experiments using the projection factor alone. Appendix H connects the local cost comparison to the finite-sample design and gives the expanded factor-menu results, prediction-powered baseline comparisons, and details of the LLMBar calibration and replay.

A PROOFS FOR REPORTING AND TESTING

A.1 PROOF OF PROPOSITION 3.1

Proof. Fix any $\mu \in [a, 1-a]$, any history, and either effort level e. The fresh-item and signal assumptions give

$$\Pr_{\mu}(Y_i = 1 \mid U_i = 1) = \frac{\mu(1-\eta_e)}{\mu(1-\eta_e) + (1-\mu)\eta_e} > \frac{1}{2},$$

because $\mu \geq a > \eta_e$. Similarly, $\Pr_{\mu}(Y_i = 0 \mid U_i = 0) > 1/2$ because $\mu \leq 1 - a < 1 - \eta_e$. These comparisons also hold conditional on the history. Since $\pi r > 0$, reporting $S_i = U_i$ uniquely maximizes expected payment after either assessment. The review cost is already incurred and does not depend on the report.

Before review, truthful reporting has agreement probability $1-\eta_e$, independent of μ. Even a provider told the true μ therefore cannot earn more than

$$v(d) = \max\{0, \mathcal{U}(L; d), \mathcal{U}(H; d)\}$$

in expected net payoff per item. Choosing a maximizing participation and effort action and then reporting truthfully attains this bound for every μ in the range. With m items remaining, the same policy attains $m\, v(d)$, since later contracts and review costs are unchanged. Current reporting cannot improve this continuation value. Thus the reporting rule also maximizes expected remaining payoff after every accepted review. It is the same rule for every admissible μ and requires no knowledge of its value. In particular, averaging over any uncertainty about μ within the allowed range leaves this optimality unchanged. □

A.2 THE OBSERVED DISTRIBUTION CLASSES

For the testing results, fix a feasible contract d and write $\eta = \eta(d)$ and $\mu_+ = \mu_0 + \delta$. Let p_μ be the distribution of one observed record under the equilibrium response. The alternative class is

$$\mathcal{P}_1(d) = \{p_\mu : \mu \in [\mu_+, 1-a]\}.$$

Let $\mathcal{Q}_0(d)$ contain the one-item observed distributions allowed by the robustness requirement in Section 3.3 with $\mu \in [a, \mu_0]$. All probabilities below refer to one sampled item unless a product distribution is specified.

A useful representation of the null class uses the joint probabilities

$$t_{ys} = \Pr(Y = y, S = s), \qquad y \in \{0, 1\}, \quad s \in \{0, 1, \perp\}.$$

These probabilities are nonnegative, sum to one, and satisfy

$$a \leq \sum_s t_{1s} \leq \mu_0.$$

For $O = (S, A, AY)$, the resulting observed probabilities are

$$q(s, 0, 0) = (1-\pi) \sum_y t_{ys},$$

$$q(s, 1, y) = \pi t_{ys}. \tag{29}$$

The admissible arrays form a closed, bounded set of finite probability vectors. Their linear image $\mathcal{Q}_0(d)$ is therefore compact. It is convex because mixing two admissible arrays preserves every constraint. Conversely, every array satisfying these constraints defines a one-item joint law in the robustness class of Section 3.3, with $\mu = \sum_s t_{1s}$. Repeating this law independently across items gives an admissible null process. The joint probabilities specify the statistical association between reports and gold labels.

At equilibrium there are no refusals, and the corresponding array is

$$t^{\mathrm{eq}}_{ys}(\mu) = \mu^y(1-\mu)^{1-y}\left[(1-\eta(d))\mathbf{1}\{s=y\} + \eta(d)\mathbf{1}\{s=1-y\}\right], \qquad s \in \{0,1\}.$$

This array and its observed law p_μ are affine in μ. Thus $\mathcal{P}_1(d)$ is also compact and convex: it is the line segment joining p_{μ_+} and p_{1-a}. Every distribution in this class is positive on the same support, denoted Ω_d: completed reports with positive audit-protocol probability. In particular, records with $A = 0$ are excluded when $\pi = 1$. All equilibrium probabilities on Ω_d are bounded away from zero for a fixed d.

The two classes are disjoint. Indeed, the observed marginal

$$\Pr(A = 1, AY = 1) = \pi\mu$$

identifies μ because $\pi > 0$. Equal observed laws would therefore imply equal win rates, contrary to the separation between the null and the alternative. There is also a null law positive on Ω_d, obtained by taking $\mu = \mu_0$ and reporting a fair coin independently of Y. Compactness and lower semicontinuity of KL divergence imply that its minimum over $\mathcal{P}_1(d) \times \mathcal{Q}_0(d)$ is attained. This minimum is finite because of the positive null law, and strictly positive because the two classes are disjoint. Any minimizing null law must be positive on Ω_d.

A.3 Reduction to Two Variables

We show that the two-variable optimization in (14) gives the smallest KL divergence over the full distribution classes:

$$J(d) = \min_{\substack{p\in\mathcal{P}_1(d)\\ q\in\mathcal{Q}_0(d)}} D_{\mathrm{KL}}(p\|q) = D_{\mathrm{KL}}(p^\star_d\|q^\star_d). \tag{30}$$

The reduction fixes the alternative win rate at μ_+ and the null win rate at μ_0, then leaves only the two null reporting probabilities to optimize.

First, suppose $\mu > \mu_+$ and put $\lambda = \delta/(\mu-\mu_0) \in (0,1)$. The affine dependence of p_μ on μ and joint convexity of KL divergence give, for any null law q,

$$D_{\mathrm{KL}}\big(p_{\mu_+} \,\big\|\, \lambda q + (1-\lambda)p_{\mu_0}\big) \le \lambda D_{\mathrm{KL}}(p_\mu\|q).$$

The mixture on the left is still a null law: its gold mean is at most μ_0, and it retains the audit protocol. Since the divergence is positive, this improves any finite candidate with $\mu > \mu_+$. Hence $p^\star_d = p_{\mu_+}$.

Next, suppose a candidate null law has gold mean $\nu < \mu_0$. Set $\lambda = (\mu_0-\nu)/(\mu_+-\nu) \in (0,1)$ and mix this law toward p_{μ_+}. The resulting law $q' = (1-\lambda)q + \lambda p_{\mu_+}$ has gold mean μ_0 and satisfies

$$D_{\mathrm{KL}}(p_{\mu_+}\|q') \le (1-\lambda)D_{\mathrm{KL}}(p_{\mu_+}\|q).$$

Thus the minimizing null mean is μ_0. Finally, any refusal mass can be reassigned to report zero, keeping the gold label and audit indicator unchanged. This preserves the null mean and audit protocol, increases some completed-report probabilities, and does not decrease any of them. Since p_{μ_+} is positive on all completed-report records and has no refusal mass, this strictly decreases the divergence whenever refusal mass is present.

Every remaining null law is specified by $x_y = Q(S = 1 \mid Y = y)$, $y \in \{0,1\}$. Its four joint probabilities are

	$S=0$	$S=1$
$Y=0$	$(1-\mu_0)(1-x_0)$	$(1-\mu_0)x_0$
$Y=1$	$\mu_0(1-x_1)$	$\mu_0 x_1$.

Write $\theta_+ = \eta + (1-2\eta)\mu_+$ and $\theta_0 = (1-\mu_0)x_0 + \mu_0 x_1$ for the alternative and null report-one probabilities, as in Section 4. Conditional on an audit, the chain rule for KL divergence separates

the gold-label distribution from the report distribution given that label. Without an audit, only the report is observed. Define Bernoulli KL divergence by

$$D_{\mathrm{Ber}}(u, v) = u \log \frac{u}{v} + (1-u) \log \frac{1-u}{1-v}.$$

The full divergence is therefore

$$\begin{aligned} H_d(x_0, x_1) = \pi D_{\mathrm{Ber}}(\mu_+, \mu_0) + \pi(1-\mu_+) D_{\mathrm{Ber}}(\eta, x_0) \\ + \pi\mu_+ D_{\mathrm{Ber}}(1-\eta, x_1) + (1-\pi) D_{\mathrm{Ber}}(\theta_+, \theta_0). \end{aligned}$$

This proves (30) and the stated score construction. Each of the two audited conditional terms is strictly convex in its own coordinate, since

$$\frac{\partial^2}{\partial z^2} D_{\mathrm{Ber}}(u, z) = \frac{u}{z^2} + \frac{1-u}{(1-z)^2} > 0.$$

Their coefficients are positive, and the remaining term is convex. Moreover, H_d tends to infinity at the boundary of $[0,1]^2$. It consequently has a unique minimizer in $(0,1)^2$. When $\pi = 1$, the minimizer is $x_0^\star = \eta$, $x_1^\star = 1-\eta$, and $J(d) = D_{\mathrm{Ber}}(\mu_+, \mu_0)$.

A.4 A Fixed-Contract Power Bound

Reaching $\log(1/\alpha)$ requires about $\log(1/\alpha)/J(d)$ items at the target improvement. To ensure power, we account for random variation using

$$R_d = \max_{o \in \Omega_d} \ell_d(o) - \min_{o \in \Omega_d} \ell_d(o), \tag{31}$$

the largest possible log likelihood ratio minus the smallest. With $z_d = R_d \sqrt{\log(1/\beta)/2}$, choose

$$N_{\mathrm{Hoef}}(d) = \left\lceil \left\{ \frac{z_d + \sqrt{z_d^2 + 4J(d)\log(1/\alpha)}}{2J(d)} \right\}^2 \right\rceil. \tag{32}$$

This *Hoeffding bound* ensures power by Proposition A.1. Alternatively, the *Chernoff bound* $N_{\mathrm{Cher}}(d)$ uses the full score distribution to reduce the required sample size (Appendix C.4). Both use the same likelihood ratio and rejection threshold.

Under the equilibrium response, the expected log likelihood ratio is at least $J(d)$ for every $\mu \in [\mu_+, 1-a]$. Together with the e-value property, this gives the following guarantees.

Proposition A.1 (False-positive control and power)**.** *For every fixed feasible d and deterministic n, the test (13) using the projection factor e_d, with the extension in Appendix A.5, satisfies (9), including when reports depend on earlier records. Moreover, $0 < R_d < \infty$, and the power requirement (10) holds whenever*

$$nJ(d) - \log(1/\alpha) \geq R_d \sqrt{\frac{n \log(1/\beta)}{2}}. \tag{33}$$

Solving this Hoeffding bound for n gives (32) (Appendix A.5).

A.5 Proof of Proposition A.1

For brevity, write $p^\star = p_d^\star$, $q^\star = q_d^\star$, and $J = J(d)$. Define

$$e_d(o) = \frac{p^\star(o)}{q^\star(o)} \quad (o \in \Omega_d), \qquad e_d(o) = 0 \quad (o \notin \Omega_d).$$

The second convention includes refusal records, for which we take $\log e_d(o) = -\infty$. The ratio is evaluated only on Ω_d, where its denominator is positive.

The equivalence in (30) lets us use the optimality of the pair over the full distribution classes. Convexity means that moving from a minimizing probability vector toward any other member remains feasible. The directional derivative along such a move must therefore be nonnegative. For any

$q \in \mathcal{Q}_0(d)$, differentiate $D_{\mathrm{KL}}(p^\star \| (1-t)q^\star + tq)$ at $t = 0$ from the right. Convexity and optimality imply

$$0 \le -\sum_{o \in \Omega_d} p^\star(o) \frac{q(o) - q^\star(o)}{q^\star(o)} = 1 - \mathbb{E}_q[e_d(O)].$$

Likewise, differentiation in the alternative coordinate gives, for every $p \in \mathcal{P}_1(d)$,

$$\begin{aligned} 0 &\le \sum_{o \in \Omega_d} [p(o) - p^\star(o)] \left[1 + \log \frac{p^\star(o)}{q^\star(o)}\right] \\ &= \mathbb{E}_p[\log e_d(O)] - J. \end{aligned}$$

These calculations are legitimate because $p^\star$ and $q^\star$ are strictly positive on Ω_d and all alternative laws vanish outside it. They yield the two bounds used for validity and power:

$$\begin{aligned} \sup_{Q \in \mathcal{Q}_0(d)} \mathbb{E}_Q[e_d(O)] &\le 1, \\ \inf_{P \in \mathcal{P}_1(d)} \mathbb{E}_P[\log e_d(O)] &\ge J(d). \end{aligned} \tag{34}$$

Fix a null parameter and any report law allowed in Section 3.3. Conditional on the observations preceding item i, the new population item retains its original distribution, and the new audit draw has the prescribed independent probability. The report–gold association may change with the history, but the conditional law of O_i still belongs to $\mathcal{Q}_0(d)$. Therefore

$$\mathbb{E}[e_d(O_i) \mid O_1, \dots, O_{i-1}] \le 1.$$

Repeated conditional expectation gives $\mathbb{E}[\prod_{i=1}^n e_d(O_i)] \le 1$. Markov's inequality proves the Type I error guarantee for (13). The product is evaluated only after the prespecified n items have been observed. If any item is refused, its factor is zero and the test does not reject.

Under any equilibrium alternative, the observations are independent and identically distributed. Let

$$R_d = \max_{o \in \Omega_d} \log e_d(o) - \min_{o \in \Omega_d} \log e_d(o).$$

This quantity is finite and positive. Both minimizing laws put all their mass on Ω_d, as shown in Appendix A.3. A constant likelihood ratio on this support would therefore equal one and imply $J = 0$, a contradiction. Hoeffding's inequality, together with $\mathbb{E}_p \log e_d \ge J$, yields

$$\Pr_p \left\{ \sum_{i=1}^n \log e_d(O_i) < \log(1/\alpha) \right\} \le \exp\left\{ -\frac{2\{nJ - \log(1/\alpha)\}^2}{nR_d^2} \right\}$$

whenever $nJ > \log(1/\alpha)$. The sample-size condition (33) makes the right-hand side at most β, uniformly over the equilibrium alternative.

A.6 Minimum Sample Size and Cost

Let $n^\star(d)$ be the smallest fixed sample size for any test satisfying both error requirements under d, and let $C^\star_{\mathcal{D}}$ be the minimum expected cost in (11) over contracts in $\mathcal{D}$.

Theorem A.2 (Minimum sample size and cost)**.** *Fix β, δ, and the model inputs. As $\alpha \downarrow 0$, for every fixed feasible contract,*

$$n^\star(d) = \frac{\log(1/\alpha)}{J(d)} + O\left(\sqrt{\log(1/\alpha)}\right). \tag{35}$$

For a fixed nonempty list $\mathcal{D} = \{d_1, \dots, d_K\}$ of feasible contracts,

$$C^\star_{\mathcal{D}}(\alpha, \beta, \delta) = \log(1/\alpha) \min_{1 \le j \le K} \frac{k(d_j)}{J(d_j)} + O\left(\sqrt{\log(1/\alpha)}\right). \tag{36}$$

Both the number K of candidate contracts and their terms are held fixed as α decreases.

The ratio $k(d)/J(d)$ measures expenditure per unit of information. The proof and an explicit Hoeffding sample-size bound appear in Appendices A.7 and A.4.

A.7 Proof of Theorem A.2

Write $L = \log(1/\alpha)$ and let $n^\star(d)$ be the smallest deterministic sample size for which some test satisfies (9) and (10) under d. Solving (33) shows that a valid choice is

$$n = \left\lceil \left\{ \frac{z_d + \sqrt{z_d^2 + 4J(d)L}}{2J(d)} \right\}^2 \right\rceil, \qquad z_d = R_d \sqrt{\frac{\log(1/\beta)}{2}}.$$

In particular, $n^\star(d) \le L/J(d) + O(\sqrt{L})$ for fixed d, β, and δ.

For the converse, consider any admissible test $\phi \in [0,1]$ based on n observations. The product law $(p_d^\star)^{\otimes n}$ is an equilibrium alternative, and $(q_d^\star)^{\otimes n}$ belongs to the stipulated robust null class, as shown in Appendix A.2. Put

$$Z_n = \sum_{i=1}^{n} \log \frac{p_d^\star(O_i)}{q_d^\star(O_i)}.$$

Under $(p_d^\star)^{\otimes n}$, its expectation is $nJ(d)$ and its summands have range length R_d. For every real t, change of measure gives

$$\begin{aligned} 1 - \beta &\le \mathbb{E}_{(p_d^\star)^{\otimes n}}[\phi] \\ &\le \Pr_{(p_d^\star)^{\otimes n}}(Z_n > t) + e^t \mathbb{E}_{(q_d^\star)^{\otimes n}}[\phi] \\ &\le \Pr_{(p_d^\star)^{\otimes n}}(Z_n > t) + \alpha e^t. \end{aligned}$$

Choose $\gamma = (1-\beta)/2$ and $t = nJ(d) + R_d\sqrt{n\log(1/\gamma)/2}$. Hoeffding's inequality bounds the probability by γ. Consequently, every admissible test must satisfy

$$nJ(d) + R_d \sqrt{\frac{n \log(1/\gamma)}{2}} \ge L + \log \gamma.$$

This implies $n \ge L/J(d) - O(\sqrt{L})$. Combining the two bounds establishes

$$n^\star(d) = \frac{L}{J(d)} + O(\sqrt{L}).$$

The lower bound applies to every test satisfying the stated guarantees, including randomized tests.

For a fixed finite nonempty menu $\mathcal{D}$ of feasible contracts, the constants in these bounds can be chosen uniformly over $d \in \mathcal{D}$. Since the contract and sample size are deterministic choices before data collection,

$$\begin{aligned} C_{\mathcal{D}}^\star(\alpha, \beta, \delta) &= \min_{d \in \mathcal{D}} k(d) n^\star(d) \\ &= L \min_{d \in \mathcal{D}} \frac{k(d)}{J(d)} + O(\sqrt{L}). \end{aligned}$$

Every $k(d)$ and $J(d)$ is positive, so this implies the first-order cost formula in the theorem. Finiteness of the menu is essential to this proof's uniform remainder; the argument does not characterize the unrestricted infimum over all feasible contracts.

B Local Information And The Value Of Effort

B.1 When Does Higher Effort Pay?

To compare low and high effort, take $\mu_0 = 1/2$ and write $J_\delta(d)$ to show dependence on the improvement δ.

Proposition B.1 (Information for a small improvement)**.** *For every fixed feasible contract d, as $\delta \downarrow 0$,*

$$J_\delta(d) = \frac{\delta^2}{2V(d)} + o(\delta^2), \qquad V(d) = \frac{1}{4} + \left(\frac{1}{\pi} - 1\right) \eta(d)\{1 - \eta(d)\}. \tag{37}$$

The variance interpretation used in Section 5.2 follows directly from the audit correction. At $\mu = 1/2$, with $\eta = \eta(d)$, the conditional-mean predictions are $m_0 = \eta$ and $m_1 = 1 - \eta$. Write $W = m_S + A(Y - m_S)/\pi$. Conditional on (S, Y), the audit is independent, so $\mathbb{E}W = 1/2$ and

$$\begin{aligned}\mathrm{Var}(W) &= \mathrm{Var}\{\mathbb{E}(Y \mid S)\} + \frac{1}{\pi}\mathbb{E}\{\mathrm{Var}(Y \mid S)\} \\ &= \left(\frac{1}{2} - \eta\right)^2 + \frac{\eta(1-\eta)}{\pi} = V(d).\end{aligned}$$

Taking $\eta = \overline{\eta}_H$ gives $\overline{V}(d)$ in the main text. These are predictions at the null boundary; the finite-improvement projection factor is still fitted at its target alternative.

Full auditing gives $V(d) = 1/4$; fewer audits increase uncertainty according to report accuracy. Taking $\alpha \downarrow 0$ and then $\delta \downarrow 0$, Theorem A.2 makes leading cost proportional to $F(d; g) = k(d)V(d)$, compared with $F_G(g) = g/4$ for gold-only testing. The contract is fixed in both limits.

Let $F_L^\star(g)$ and $F_H^\star(g)$ be the infima of $F(d; g)$ over feasible contracts inducing low and high effort, respectively. Write

$$h = \frac{c_H - c_L}{\eta_L - \eta_H}, \qquad \rho_H = \frac{h}{M}, \qquad v_e = \eta_e(1 - \eta_e).$$

High effort requires $\pi r > h$. Since $r \le M$, it also requires $\pi > \rho_H$: the payment cap puts a lower bound on the audit rate.

Proposition B.2 (When higher effort reduces cost)**.** *Suppose $0 < c_L < (1 - \eta_L)h$ and $M > h$. As $g \to \infty$,*

$$F_L^\star(g)/g \longrightarrow v_L, \qquad F_H^\star(g)/g \longrightarrow v_H + (1/4 - v_H)\rho_H. \tag{38}$$

Consequently, if

$$\rho_H < \frac{v_L - v_H}{1/4 - v_H}, \tag{39}$$

then $F_H^\star(g) < \min\{F_L^\star(g), F_G(g)\}$ for all sufficiently large g. Under the reverse strict inequality, $F_L^\star(g) < F_H^\star(g)$ for all sufficiently large g.

The assumptions ensure high-effort participation. For each fixed g, feasible contracts approaching the two infima give a fixed finite menu covered by Theorem A.2 (Appendix B). The accuracy gain must outweigh the minimum audit fraction imposed by the payment cap. This comparison takes $\alpha \downarrow 0$, then $\delta \downarrow 0$, and finally $g \to \infty$; Section 6 examines finite studies.

For each fixed contract, the small-improvement cost comparison in Section B.1 is the iterated limit

$$\lim_{\delta \downarrow 0} \delta^2 \lim_{\alpha \downarrow 0} \frac{k(d)n^\star(d)}{\log(1/\alpha)} = 2k(d)V(d). \tag{40}$$

The sample-size theorem gives the inner limit, and Proposition B.1 gives the outer limit.

To relate the effort-specific infima to a finite contract menu, fix g and suppose both effort classes are feasible. For every $\varepsilon > 0$, the definition of an infimum provides feasible contracts d_L, d_H with $F_e^\star(g) \le F(d_e; g) < F_e^\star(g) + \varepsilon$ for $e = L, H$. Keep $\mathcal{D} = \{d_L, d_H\}$ fixed as α and δ decrease. Theorem A.2 and the local expansion give

$$\lim_{\delta \downarrow 0} \delta^2 \lim_{\alpha \downarrow 0} \frac{C_{\mathcal{D}}^\star(\alpha, \beta, \delta)}{\log(1/\alpha)} = 2 \min_{d \in \mathcal{D}} F(d; g).$$

Thus finite feasible choices approximate the effort-specific local costs arbitrarily closely. Any strict gap between these infima, or between one of them and $g/4$, is preserved by sufficiently close feasible choices. This argument fixes the menu before taking the statistical limits and does not require an infimum to be attained.

Proof of Proposition B.1. Fix a contract, write π and η for its audit probability and equilibrium error rate, and let p_μ be its observed-data law. The map $\mu \mapsto p_\mu$ is affine. If $\mu > 1/2 + \delta$ and q is any null law, put $\lambda = \delta/(\mu - 1/2)$. Mixing both laws with $p_{1/2}$ gives

$$D\big(p_{1/2+\delta} \,\|\, \lambda q + (1 - \lambda)p_{1/2}\big) \le \lambda D(p_\mu \| q),$$

and the second argument is still a null law. Hence the composite alternative can be restricted to $\mu = 1/2 + \delta$. Likewise, mixing a null law with gold mean strictly below $1/2$ toward $p_{1/2+\delta}$ until its gold mean equals $1/2$ cannot increase the divergence. We may therefore impose equality in the null mean constraint. Refusal reports cause no additional difficulty: mapping every refusal to a binary report preserves the gold mean and audit protocol, leaves $p_{1/2+\delta}$ unchanged, and cannot increase divergence.

It remains to expand the divergence between two binary-report laws. Let $f^0_{ys} = \Pr^d_{1/2}(Y = y, S = s)$, so $f^0_{ys} = (1-\eta)/2$ for $y = s$ and $\eta/2$ otherwise. For a perturbation $\xi = (\xi_{ys})$ satisfying $\sum_{y,s} \xi_{ys} = 0$, write $\xi_{\cdot s} = \sum_y \xi_{ys}$ and $f^0_{\cdot s} = \sum_y f^0_{ys}$. The quadratic term of observed-data KL divergence is one half of

$$G(\xi, \xi) = \pi \sum_{y,s} \frac{\xi^2_{ys}}{f^0_{ys}} + (1-\pi) \sum_s \frac{\xi^2_{\cdot s}}{f^0_{\cdot s}}.$$

Here the first term comes from audited pairs and the second from unaudited reports; when $\pi = 1$, the latter term is absent.

Set $m_s = \mathbb{E}^d_{1/2}[Y \mid S = s] = \eta + (1-2\eta)s$ and define

$$z_{ys} = f^0_{ys} \left(m_s - \frac{1}{2} + \frac{y - m_s}{\pi} \right).$$

Direct substitution gives $\sum_{y,s} z_{ys} = 0$, $G(\xi, z) = \sum_{y,s} y \xi_{ys}$, and

$$\begin{aligned} G(z, z) &= \operatorname{Var}(m_S) + \pi^{-1} \mathbb{E} \operatorname{Var}(Y \mid S) \\ &= \frac{1}{4} + (\pi^{-1} - 1)\eta(1-\eta) = V(d). \end{aligned}$$

Consequently, the Cauchy–Schwarz inequality yields

$$\inf_{\substack{\sum \xi_{ys} = 0 \\ \sum y \xi_{ys} = 1}} G(\xi, \xi) = \frac{1}{V(d)},$$

with equality at $\xi = z/V(d)$.

All four entries of f^0 are positive. Since $\pi > 0$, the observed-data map identifies the underlying joint law. A minimizer of the divergence therefore converges to f^0 as $\delta \downarrow 0$. The preceding quadratic form is positive definite on the space of zero-sum perturbations. The feasible choice whose difference from the alternative joint law is $\delta z / V(d)$ gives an $O(\delta^2)$ upper bound, so the minimizing perturbation is $O(\delta)$. Taylor expansion is thus uniform in its neighborhood. The gold means of the alternative and null differ by exactly δ, and the quadratic minimization above gives

$$J_\delta(d) = \frac{\delta^2}{2V(d)} + o(\delta^2).$$

This expansion keeps the contract, and in particular π, fixed. □

Proof of Proposition B.2. Put $h = (c_H - c_L)/(\eta_L - \eta_H)$ and $t = \pi r$. High effort requires $t > h$. Under $0 < c_L < (1 - \eta_L)h$, we have

$$c_H = c_L + (\eta_L - \eta_H)h < (1 - \eta_H)h = t_H.$$

Hence a zero base payment and any $t > h$ induce participation as well as high effort. The payment cap implies $t \le \pi M$, so such a contract exists exactly when $\pi > \rho_H = h/M$. Every high-effort contract has expected payment $b + (1 - \eta_H)t > t_H$. Conversely, at any fixed $\pi > \rho_H$, choosing $b = 0$ and $t \downarrow h$ approaches t_H while satisfying the cap. Thus t_H is the infimum payment at every feasible high-effort audit probability.

For low effort, participation requires expected payment at least c_L. This amount is attained at any $\pi \in (0, 1]$ by choosing $r > 0$ sufficiently small and setting $b = c_L - \pi r(1 - \eta_L)$. The base is nonnegative, the effort inequality selects low effort, and $b + r < M$ for small enough r, because the assumptions imply $c_L < h < M$. Write $t_L = c_L$.

With $v_e = \eta_e(1-\eta_e)$ and $a_e = 1/4 - v_e$, the local cost coefficient at the payment infimum is

$$(t_e + g\pi)(a_e + v_e/\pi) = t_e a_e + g v_e + g a_e \pi + t_e v_e/\pi.$$

Its unconstrained minimum occurs at $\pi_e^0 = \sqrt{t_e v_e/(g a_e)}$. For low effort the minimizing audit probability is $\min\{1, \pi_L^0\}$; for high effort the infimum is given by restricting π_H^0 to the interval $[\rho_H, 1]$. An endpoint ρ_H is approached from above. In every high-effort case the strict incentive constraint means that the payment infimum is also approached rather than attained.

In particular, the two minimum local costs can be written as

$$\begin{aligned} F_L^\star(g) &= \inf_{0<\pi\le 1} (c_L + g\pi)\left[\frac{1}{4} + (\pi^{-1} - 1)v_L\right], \\ F_H^\star(g) &= \inf_{\rho_H<\pi\le 1} (t_H + g\pi)\left[\frac{1}{4} + (\pi^{-1} - 1)v_H\right]. \end{aligned} \tag{41}$$

As $g \to \infty$, the low-effort minimum satisfies

$$F_L^\star = g v_L + 2\sqrt{c_L v_L g a_L} + c_L a_L,$$

whereas the high-effort infimum eventually satisfies

$$F_H^\star = g(v_H + a_H \rho_H) + t_H(a_H + v_H/\rho_H).$$

Dividing by g gives the stated limits. Because $v_L < 1/4$, the strict inequality $v_H + a_H\rho_H < v_L$ makes high effort cheaper than both low effort and the gold-only coefficient $g/4$ for all sufficiently large g. The reverse strict inequality makes low effort cheaper than high effort for all sufficiently large g. Each comparison involving a strict gap is achieved by feasible high-effort contracts sufficiently close to their infimum. These statements concern the local coefficients, without exchanging the small-effect limit with optimization over audit probabilities. For the same fixed-contract limits, the expected number of audits is proportional to

$$\pi V_e(\pi) = v_e + (1/4 - v_e)\pi.$$

The high-effort audit fraction approaches ρ_H, whereas the low-effort audit fraction approaches zero as gold labels become expensive. The corresponding audit-count coefficients tend to $v_H + (1/4 - v_H)\rho_H$ and v_L, respectively. Condition (39) makes the high-effort coefficient smaller. □

C Computational Details

C.1 Solving the Two-Variable Problem

Fix a feasible contract d and write $\eta = \eta(d)$. The decomposition in Appendix A.3 gives simple derivatives for H_d. Define

$$b_u(z) = \frac{z-u}{z(1-z)}, \qquad h_u(z) = \frac{u}{z^2} + \frac{1-u}{(1-z)^2}.$$

These are the first and second derivatives of $D_{\mathrm{Ber}}(u, z)$ with respect to z. With θ_+ and θ_0 as in Section 4, the gradient is

$$\begin{aligned} \partial_0 H_d &= \pi(1-\mu_+) b_\eta(x_0) + (1-\pi)(1-\mu_0) b_{\theta_+}(\theta_0), \\ \partial_1 H_d &= \pi\mu_+ b_{1-\eta}(x_1) + (1-\pi)\mu_0 b_{\theta_+}(\theta_0), \end{aligned}$$

and the Hessian is

$$\nabla^2 H_d = \pi \begin{pmatrix} (1-\mu_+)h_\eta(x_0) & 0 \\ 0 & \mu_+ h_{1-\eta}(x_1) \end{pmatrix} + (1-\pi) h_{\theta_+}(\theta_0) \begin{pmatrix} 1-\mu_0 \\ \mu_0 \end{pmatrix} \begin{pmatrix} 1-\mu_0 & \mu_0 \end{pmatrix}.$$

A damped Newton iteration starts at $x^{(0)} = (\eta, 1-\eta)$. At each step, solve the 2×2 linear system $\nabla^2 H_d(x)\Delta = \nabla H_d(x)$ and propose $x - \gamma\Delta$, initially with $\gamma = 1$. Halve γ until the proposed point lies in $(0,1)^2$ and

$$H_d(x - \gamma\Delta) \le H_d(x) - \frac{\gamma}{4} \nabla H_d(x)^\mathsf{T} \Delta.$$

Then accept the step. One stopping rule is $\|\nabla H_d(x)\|_\infty \le \varepsilon$ for a chosen numerical tolerance $\varepsilon > 0$. Every Hessian is positive definite, so the Newton direction decreases the objective for a sufficiently small step. When $\pi = 1$, the starting point is already optimal. The resulting probability table gives candidate scores $f_d(o) = p_+(o)/q_x(o)$. Their validity and power are checked by the following calculations, rather than inferred from the stopping tolerance.

C.2 Checking the Null Expectation of a Computed Score

The two-variable problem in (14) can be solved numerically before data collection. An approximate optimizer need not give scores satisfying (34) exactly. The following calculations check and, when necessary, rescale any candidate score. They require only maxima and evaluations at two win rates, without a second optimization problem.

Fix a feasible contract and let $f_d(o)$ be a candidate multiplicative score, finite and positive on the completed-record support Ω_d and zero on refusal records. For $y \in \{0, 1\}$, define

$$C_y = \max_{s \in \{0,1\}} \{(1-\pi) f_d(s, 0, 0) + \pi f_d(s, 1, y)\}.$$

Terms for unaudited records are omitted when $\pi = 1$. Conditional on gold outcome y, the expression inside the maximum is the expected score from submitting report s. Any distribution of binary reports has expectation at most C_y; a refusal gives zero and cannot increase this maximum. Consequently, at a fixed null win rate μ, the largest expected score is $(1-\mu)C_0 + \mu C_1$. This value is attained by choosing a maximizing report for each gold outcome in the null joint distribution. Since it is affine in μ, the largest value over $[a, \mu_0]$ occurs at an endpoint. Thus

$$\begin{aligned} c_d &= \max\{(1-a)C_0 + aC_1,\ (1-\mu_0)C_0 + \mu_0 C_1\} \\ &= \sup_{Q \in \mathcal{Q}_0(d)} \mathbb{E}_Q[f_d(O)]. \end{aligned} \tag{42}$$

In particular, $c_d > 0$ and the normalized score $e_d^{\mathrm{num}} = f_d / c_d$ has null expectation at most one. The product of these scores therefore gives the same false-positive guarantee as in Proposition A.1, including when reports depend on earlier observations.

C.3 Computing a Power Bound

For the normalized score, let L_y denote its expected log value conditional on gold outcome y under the equilibrium response:

$$\begin{aligned} L_y = \sum_{s=0}^{1} &\big[(1-\eta)\mathbf{1}\{s = y\} + \eta \mathbf{1}\{s \neq y\}\big] \\ &\times \{(1-\pi)\log e_d^{\mathrm{num}}(s, 0, 0) + \pi \log e_d^{\mathrm{num}}(s, 1, y)\}, \qquad \eta = \eta(d). \end{aligned}$$

Again omit unaudited terms when $\pi = 1$. At win rate μ, the expected log score is $(1-\mu)L_0 + \mu L_1$. Its smallest value on the alternative interval also occurs at an endpoint:

$$\begin{aligned} m(d) &= \min\{(1-\mu_+)L_0 + \mu_+ L_1,\ aL_0 + (1-a)L_1\} \\ &= \inf_{P \in \mathcal{P}_1(d)} \mathbb{E}_P[\log e_d^{\mathrm{num}}(O)]. \end{aligned} \tag{43}$$

Let R_d^{num} be the largest minus the smallest log score on Ω_d. Normalization subtracts the same number $\log c_d$ from every completed-record log score, so this range is also the range of $\log f_d$.

If $m(d) > 0$, replace $J(d)$ in (32) by $m(d)$ and use R_d^{num} for the log-ratio range. The proof of Proposition A.1 then gives the desired power for the normalized test. If $m(d) \leq 0$, these calculations do not certify power; the candidate score must be improved or discarded. For the exact optimal score, (34) and equality under the minimizing pair give $c_d = 1$ and $m(d) = J(d)$.

Finite-precision calculations can use an upper bound $\bar{c}_d \geq c_d$ in the normalization, a lower bound on the resulting expected log score, and an upper bound on its range. The formulas above reduce these bounds to arithmetic and logarithm evaluations; outward rounding can account for their numerical error. The bound on the sample size follows with these conservative values, without requiring a certified optimizer. The test is defined using the resulting fixed scores and threshold.

The asymptotic cost statements concern the exact scores. If the computed minimizers converge to the exact ones and the evaluation bounds are tightened as $\alpha \downarrow 0$, the normalization converges to one, the minimum expected log score to $J(d)$, and the range to R_d. The same leading cost coefficient is then retained for every contract in the fixed finite menu.

C.4 A Chernoff Refinement of the Sample-Size Bound

The six-record distribution also gives a sample-size refinement using the full distribution of the log score. Fix a feasible contract and a normalized score e_d^{num} from the preceding sections, with $m(d) > 0$. Write $\ell_d^{\text{num}} = \log e_d^{\text{num}}$ on Ω_d. The test rejects if the final sum of log scores is at least $\log(1/\alpha)$. For $s > 0$, define

$$\mathcal{M}_d(s) = \max_{\mu \in \{\mu_+, 1-a\}} \sum_{o \in \Omega_d} p_\mu(o) \exp\{-s\ell_d^{\text{num}}(o)\}. \tag{44}$$

The probabilities p_μ describe the equilibrium response; the score function remains fixed in both endpoint evaluations. For any chosen s with $\mathcal{M}_d(s) < 1$, a sufficient sample size is

$$N_{\text{Cher}}(d; s) = \left\lceil \frac{\log(1/\beta) + s\log(1/\alpha)}{-\log \mathcal{M}_d(s)} \right\rceil. \tag{45}$$

The best bound of this form is

$$N_{\text{Cher}}(d) = \left\lceil \inf_{\substack{s>0 \\ \mathcal{M}_d(s)<1}} \frac{\log(1/\beta) + s\log(1/\alpha)}{-\log \mathcal{M}_d(s)} \right\rceil. \tag{46}$$

This is an optional refinement of the range-based bound in (32).

To verify the guarantee, put $L = \log(1/\alpha)$. For any alternative win rate μ, exponential Markov inequality and independence give

$$\Pr_\mu^d \left\{ \sum_{i=1}^n \ell_d^{\text{num}}(O_i) < L \right\} \le e^{sL} \left[\sum_{o \in \Omega_d} p_\mu(o) e^{-s\ell_d^{\text{num}}(o)} \right]^n \le e^{sL} \mathcal{M}_d(s)^n.$$

The last inequality holds throughout $[\mu_+, 1-a]$ because the quantity in brackets is affine in μ and attains its maximum at an endpoint. Equation (45) makes the last bound at most β. The normalization still supplies false-positive control under every law in the robustness class.

An admissible s exists because $\mathcal{M}_d(0) = 1$ and its right derivative at zero is $-m(d) < 0$. Moreover, a normalized score with positive expected log value must have some completed-record score below one: otherwise a null law positive on Ω_d would have expected score greater than one. Every alternative law is positive on this support, so $\mathcal{M}_d(s)$ tends to infinity as $s \to \infty$. Convexity of $\mathcal{M}_d$ makes the admissible set a bounded interval, and the ratio in (46) tends to infinity at both endpoints. Its infimum is attained, which justifies taking the ceiling as written.

Implementation requires only one admissible value of s; the best of finitely many evaluated values also gives a sufficient sample size. The scalar optimization improves the bound but is not needed for its validity. With numerical evaluations, use an upper bound on $\mathcal{M}_d(s)$ below one and check, after rounding n up, that

$$s\log(1/\alpha) + n\log \mathcal{M}_d(s) \le \log \beta$$

using conservative evaluation bounds. This separates checking the power guarantee from locating the minimizing s.

For exact scores, Hoeffding's lemma gives $\log \mathcal{M}_d(s) \le -sJ(d) + s^2 R_d^2/8$. Set $n = N_{\text{Hoef}}(d)$ from (32) and $s = 4\{nJ(d) - L\}/(nR_d^2) > 0$. Then

$$sL + n\log \mathcal{M}_d(s) \le -\frac{2\{nJ(d) - L\}^2}{nR_d^2} \le \log \beta.$$

Thus the optimized bound in (46) satisfies $n^*(d) \le N_{\text{Cher}}(d) \le N_{\text{Hoef}}(d)$. Theorem A.2 gives $N_{\text{Cher}}(d) = L/J(d) + O(\sqrt{L})$, so Algorithm 1 has the same leading cost under either rule for a fixed finite menu. This conclusion concerns the optimized bound; any admissible fixed s still guarantees power.

D Simulation Details

This section retains the known-technology reference experiments using the projection factor alone. Appendix H adds the prediction-powered comparisons and the finite-factor menu used in the Algorithm 1.

D.1 Known-Technology Experiment Results

We examine expected cost under the provider's equilibrium response and false-positive control when reports depart from this model.

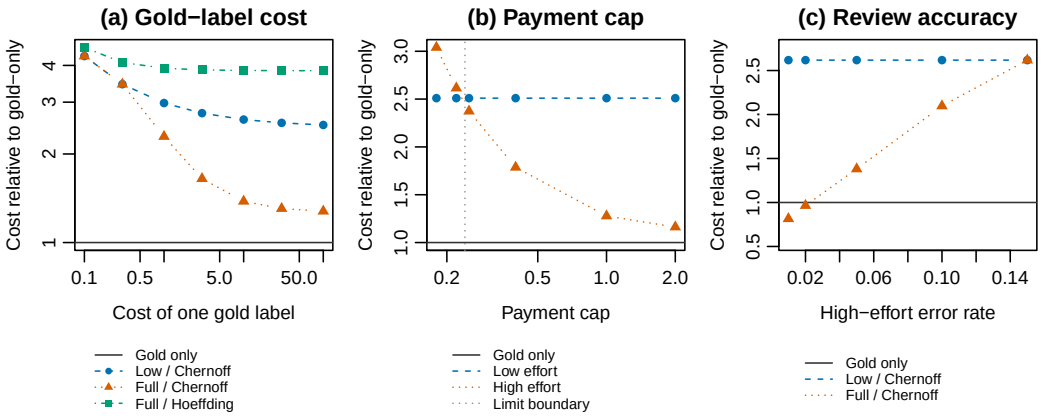

Figure 2: Expected cost relative to exact gold-only testing at the nominal error requirements. Full menu allows both effort levels. (a) Gold-label cost g; (b) payment cap at $g = 100$ (vertical line: limiting boundary $M = 0.24$); (c) high-effort error at $g = 10$. Panels (b)–(c) use Chernoff sample sizes.

D.1.1 Simulation Design

We require false-positive probability at most 5% for win rates at most 0.5, and power at least 80% at 0.6 or higher. The allowed win rates are $[0.25, 0.75]$. Unless varied, low- and high-effort reviews have error probabilities 0.2 and 0.05 and cost 0.02 and 0.04 per item, respectively. The payment cap is $M = 1$. One gold label costs g in these same arbitrary units; the baseline is $g = 10$, and the payment-cap comparison uses $g = 100$.

At each audit probability in $\{0.02, 0.04, \ldots, 1\}$, we construct a low-effort contract and, when feasible, a high-effort contract. The *full menu* contains both types. Appendix D specifies the payments and parameter grids.

We apply Algorithm 1 restricted to the projection factor, using its Chernoff rule and the Hoeffding rule in Appendix A.4 to the full menu, and Chernoff to the low-effort menu. Both rules use the same statistic and rejection threshold: Hoeffding uses the score range, whereas Chernoff uses its full distribution. The gold-only benchmark purchases labels directly, with no provider payment, and uses the smallest sample size for an exact randomized binomial test meeting both error requirements.

The simulator draws independent Bernoulli(μ) gold labels, flips each with the chosen effort's error probability to generate a report, and draws independent audits with probability π. For each design chosen before simulation, we ran 20,000 independent studies. Rejection frequencies estimate rejection probabilities; 95% Wilson binomial confidence intervals quantify Monte Carlo uncertainty (Appendix D). Expected audit costs and payments are calculated and checked against simulated averages.

D.1.2 Cost and the Value of Effort

At $g = 10$, the full-menu Chernoff design costs 2,113, compared with 4,006 for low effort and 1,530 for gold-only testing (Figure 2). These designs use about 194, 388, and 153 gold labels, respectively. The full-menu Chernoff contract audits each of 1,384 items with probability 0.14. Hoeffding selects a different contract: low effort with full auditing and 587 items. The local approximation in (37) overstates $J(d)$ by 20% here, and the finite-sample bounds and rejection threshold remain conservative: empirical power is 0.965 against the requested 0.8.

Table 3: False-positive frequencies for three tests and power of the robust test. Model uses the modeled null response; Reports treats reports as gold. Intervals appear in Appendix D.

	False positives ($\mu = 0.5$)			Power
Reporting rule	Robust	Model	Reports	($\mu = 0.6$)
Equilibrium response	.0044	.0000	.0506	.9655
Lower accuracy	.0003	.0001	.0502	.3057
Favorable bias	.0000	1.0000	1.0000	.1066
History dependent	.0031	.5153	.5538	.9492

At $g = 100$, high effort becomes cheaper than low effort between $M = 0.22$ and 0.25, near the limiting boundary $M = 0.24$ in Proposition B.2. At $g = 10$, reducing η_H to 0.01 lowers cost to 1,247, 18.5% below gold-only testing, with empirical power 0.970. This design uses about 115 gold labels, compared with 153 for gold-only testing. Higher accuracy supplies more information and requires a smaller bonus.

D.1.3 Reports That Depart from the Model

For the same $g = 10$ contract and 1,384 items, we compare the equilibrium response with three alternatives: reports with error rate 0.2, a rule changing 40% of negative assessments to positive reports, and a rule introducing this bias midway through the study after unfavorable audited judgments. These rules use private assessments and past audits, without observing the current gold outcome.

Table 3 compares our test with a likelihood ratio using the modeled null response and a binomial test treating reports as gold labels. Bias makes the latter two tests reject frequently under the null. Our test's false-positive frequencies remain below 0.05 in these settings, but its power falls under inaccurate or persistently biased reporting. Power and cost depend on the response model; validity relies on the audit protocol.

Appendix D.7 gives a comparison at matched estimated power and quantifies the information cost of robust validity.

D.2 Model Inputs and Candidate Contracts

Unless varied below, the inputs are

$$a = 0.25, \quad \mu_0 = 0.5, \quad \delta = 0.1, \quad \eta_L = 0.2, \quad \eta_H = 0.05,$$
$$c_L = 0.02, \quad c_H = 0.04, \quad M = 1, \quad g = 10,$$

with $\alpha = 0.05$ and $\beta = 0.2$. Costs are expressed in common arbitrary units. For each parameter setting, the audit grid is $\{0.02, 0.04, \ldots, 1\}$. At each grid point π, we construct a low-effort contract

$$r = 10^{-6}, \qquad b = c_L - \pi r(1 - \eta_L),$$

whose expected payment equals c_L, and a high-effort contract

$$b = 0, \qquad r = \frac{(1 + 10^{-4})h}{\pi}, \qquad h = \frac{c_H - c_L}{\eta_L - \eta_H}.$$

The small positive margin in the second contract makes high effort strictly preferable. We retain only contracts that satisfy the payment cap and induce the intended effort and participation. For all settings reported here, the low-effort base payment is positive. The code also reduces its small bonus when needed for other user-supplied inputs. The comparisons optimize over this fixed grid and these two payment specifications, rather than the continuous class of all contracts.

D.3 Sample Sizes and the Gold-Only Benchmark

For each retained contract, we compute the likelihood ratio by the two-variable optimization in Section 4.1. The numerical ratio is normalized using the analytic null supremum in Appendix C.2, multiplied by $1 + 10^{-12}$. Calculations use double precision; this buffer is a numerical precaution, and the implementation does not use interval arithmetic.

We compare the original Hoeffding sample-size bound with the Chernoff refinement in Appendix C.4. The two calculations select sample sizes for the same fixed-sample test and rejection threshold. The Chernoff calculation uses the largest exponential moment over the two alternative endpoints, minimizes the resulting bound over its scalar parameter, and checks neighboring integer sample sizes. We minimize expected cost separately for low-effort contracts with the Chernoff bound, the full menu with the Chernoff bound, and the full menu with the Hoeffding bound. Thus changing the bound can change both the selected contract and sample size.

The gold-only benchmark uses an exact randomized binomial test. For $K_n = \sum_{i=1}^{n} Y_i$, choose a cutoff c_n satisfying

$$\Pr_{\mu_0}(K_n > c_n) \le \alpha \le \Pr_{\mu_0}(K_n \ge c_n), \qquad \gamma_n = \frac{\alpha - \Pr_{\mu_0}(K_n > c_n)}{\Pr_{\mu_0}(K_n = c_n)}.$$

The test rejects above c_n and rejects with independent probability γ_n at equality. We choose the smallest n for which its power at $\mu_0 + \delta$ is at least $1 - \beta$. Monotonicity gives the requirements throughout both hypotheses. For the stated inputs, $n = 153$, $c_n = 87$, and $\gamma_n = 0.8174396855$. Its expected cost is $153g$.

D.4 Parameter Comparisons

The gold-cost comparison uses $g \in \{0.1, 0.3, 1, 3, 10, 30, 100\}$ and compares the four designs above. For each selected design, rejection probabilities are evaluated at $\mu \in \{0.5, 0.6, 0.65, 0.75\}$. The payment-cap comparison sets $g = 100$ and varies $M \in \{0.18, 0.22, 0.25, 0.4, 1, 2\}$, selecting low- and high-effort contracts separately with the Chernoff bound. The accuracy comparison uses the baseline $g = 10$ and $M = 1$ and varies $\eta_H \in \{0.01, 0.02, 0.05, 0.1, 0.15\}$, comparing the low-effort and full menus with the Chernoff bound. All other inputs remain fixed in each comparison.

A further diagnostic holds fixed the baseline contract selected from the full menu with the Chernoff bound. Its power at $\mu = 0.6$ is evaluated at the distinct sample sizes $\lceil f N_{\mathrm{Cher}} \rceil$, for $f \in \{0.2, 0.35, 0.5, 0.65, 0.8, 1\}$, together with its Hoeffding sample size N_{Hoef}. The ratio function remains unchanged. This diagnostic does not enter contract or sample-size selection.

D.5 Reporting Departures and Comparator Tests

The robustness experiment fixes the baseline full-menu Chernoff design and applies every test to the same simulated records. We consider $\mu = 0.5$ and $\mu = 0.6$ under four reporting rules:

1. *Equilibrium response.* Reports follow the symmetric error channel with error $\eta = \eta(d)$.
2. *Lower accuracy.* Reports instead follow the symmetric channel with error $\eta_L = 0.2$.
3. *Favorable bias.* An equilibrium report is replaced by one with probability 0.4, independently across items. Consequently, $x_0 = \eta + 0.4(1 - \eta)$ and $x_1 = 1 - \eta + 0.4\eta$.
4. *History-dependent bias.* The first $\lfloor n/2 \rfloor$ items follow the equilibrium response. If at most half of their audited gold labels favor the new system, including the case of no audits, the remaining items use the favorable-bias rule. Otherwise, they retain the equilibrium response.

The last rule uses only expert judgments revealed by earlier audits. It does not use an item's gold label before submitting that item's report.

The model-based comparator uses the likelihood ratio between win rates $\mu_0 + \delta$ and μ_0 while imposing the assumed symmetric report error $\eta(d)$ under both. It uses the same threshold $1/\alpha$. The second comparator applies the randomized binomial rule above to $\sum_i S_i$, treating the reports as gold labels. Both use the robust design's contract and sample size; their sample sizes are not optimized separately. Under reporting departures, this experiment assesses false-positive control and the resulting power, rather than imposing the equilibrium power requirement on those departures.

D.6 Monte Carlo Calculation and Uncertainty

Each rejection probability is estimated from $20{,}000$ independent studies. For a fixed reporting rule, the six record counts are drawn jointly from their multinomial distribution, which is equivalent to

generating the independent items individually. For the history-dependent rule, first-half counts are generated first and determine the conditional distribution used for the second half. Gold-only studies use binomial draws and independent uniform draws for boundary randomization.

Reported rejection intervals are pointwise 95% Wilson intervals; Table 4 gives them for every robustness comparison. Cost curves use the deterministic expected cost $nk(d)$ and expected number of audits $n\pi$. We also record Monte Carlo means of realized audits, payments, and total cost as checks of these expectations. Under the response model, realized payment is $nb + rN_{\text{agree}}$, where N_{agree} is the number of audited agreements.

These earlier experiments use seed 20260908 and the same random-number settings as the uncertainty experiments. The run used R 4.5.1 on macOS. The accompanying code supplies all parameter grids, fitted contracts, numerical diagnostics, and simulation summaries.

Table 4: Rejection percentages and pointwise 95% Wilson intervals from $20{,}000$ studies per row and win rate. All tests use the baseline full-menu Chernoff design.

Reporting rule	Test	$\mu = 0.5$	$\mu = 0.6$
Equilibrium response	Robust LR	0.435 [0.353, 0.536]	96.550 [96.288, 96.794]
	Model-based LR	0.000 [0.000, 0.019]	99.875 [99.816, 99.915]
	Reports as gold	5.060 [4.765, 5.372]	100.000 [99.981, 100.000]
Lower accuracy	Robust LR	0.025 [0.011, 0.059]	30.565 [29.930, 31.207]
	Model-based LR	0.005 [0.001, 0.028]	85.510 [85.015, 85.991]
	Reports as gold	5.020 [4.726, 5.331]	99.770 [99.693, 99.828]
Favorable bias	Robust LR	0.000 [0.000, 0.019]	10.660 [10.240, 11.095]
	Model-based LR	100.000 [99.981, 100.000]	100.000 [99.981, 100.000]
	Reports as gold	100.000 [99.981, 100.000]	100.000 [99.981, 100.000]
History-dependent bias	Robust LR	0.310 [0.242, 0.397]	94.920 [94.607, 95.216]
	Model-based LR	51.530 [50.837, 52.222]	99.900 [99.846, 99.935]
	Reports as gold	55.375 [54.685, 56.063]	100.000 [99.981, 100.000]

D.7 Additional Cost and Information Comparisons

We hold the baseline full-menu Chernoff design fixed and use its power estimate $\widehat{p} = 19310/20000 = 0.9655$ from Table 4. Recomputing the exact gold-only benchmark with power target $\widehat{p}$ and level 0.05 gives 296 labels: their power is 0.965671, whereas 295 labels give 0.965420. At $g = 10$, the cost is $2{,}960$, compared with $2{,}112.92$ for the contracted design, a reduction of 28.6%.

The 95% Wilson interval for the estimated power is $[0.962880, 0.967941]$. Applying the same calculation at its endpoints gives 291–302 gold labels, costs of $2{,}910$–$3{,}020$, and savings of 27.4%–30.0%. This compares estimated power: null rejection probability remains 0.05 for gold-only testing, versus an empirical 0.00435 for the contracted design. Figure 2 retains the nominal benchmark.

For the information comparison, let p_0 be the record distribution at μ_0 under the equilibrium response. The information against this null is

$$J_{\text{model}}(d) = \sum_{o \in \Omega_d} p_+(o) \log \frac{p_+(o)}{p_0(o)}.$$

The ratio $J_{\text{model}}(d)/J(d)$ measures the loss of information when the null allows arbitrary reporting. Since leading sample size is proportional to inverse information, it also compares the leading sample-size coefficients under the two validity requirements, keeping the contract fixed. For the baseline high-effort contract, $J_{\text{model}} = 0.0168272$ and $J = 0.00767236$, giving 2.19; for the selected low-effort contract, they are 0.00799247 and 0.00176896, giving 4.52.

For the high-effort contract, the local approximation $\delta^2/\{2V(d)\} = 0.00922874$ exceeds $J(d)$ by 20.3%. The local cost ratio $F(d; g)/(g/4)$ is 0.331, compared with finite-sample cost ratio 1.381. The difference reflects both the local approximation and the conservative finite-sample construction, whose power exceeds the nominal gold-only target.

E TESTING WITH UNCERTAIN ACCURACY

We prove Theorem 5.1 and (23). All contracts in this section are fixed. Write $p_{\mu,\eta}$ for the six-record probability vector under symmetric report error η, and write $\mathcal{Q}_0(d)$ for the robust null class defined in Appendix A.2. For clarity, define

$$J(\eta) = \min_{q \in \mathcal{Q}_0(d)} D_{\mathrm{KL}}(p_{\mu_+,\eta} \| q).$$

The reduction in Appendix A.3 applies for every fixed $0 < \eta < 1/2$: the minimizing null has gold mean μ_0, no refusals, and a unique pair of reporting probabilities in $(0,1)^2$. Its probability vector is denoted $q^\star_\eta$. Put $\eta_* = \overline{\eta}_H$, $p_* = p_{\mu_+,\eta_*}$, $q_* = q^\star_{\eta_*}$, and $e = p_*/q_*$. The ratio is defined on the completed-record support and is zero on refusal records, as before.

E.1 AN EXACT AUDIT-CORRECTION IDENTITY

Suppose first that $0 < \pi < 1$. Use the notation $e_{s0} = e(s,0,0)$ and $e_{s1y} = e(s,1,y)$. Differentiating the divergence with respect to each null reporting probability at its interior optimum gives

$$(1-\pi)e_{00} + \pi e_{01y} = (1-\pi)e_{10} + \pi e_{11y} = C_y, \qquad y = 0, 1. \tag{47}$$

Indeed, changing the probability of report one conditional on $Y = y$ transfers mass between its two unaudited report cells and between its two audited cells, giving precisely this equality. Since $\mathbb{E}_{q_*} e = 1$,

$$(1-\mu_0)C_0 + \mu_0 C_1 = 1.$$

Let $\lambda = C_1 - C_0$. The equalities above hold independently of the report probabilities, so

$$\mathbb{E}_{p_*} e = 1 + (\mu_+ - \mu_0)\lambda.$$

Moreover, $\mathbb{E}_{p_*} \log e = J(\eta_*) > 0$. Jensen's inequality implies $\mathbb{E}_{p_*} e > 1$, and hence $\lambda > 0$. The constants C_y are positive; in particular, $C_0 = 1 - \lambda\mu_0 > 0$.

Define

$$m_s = \mu_0 + \frac{e_{s0} - 1}{\lambda}.$$

Solving (47) for the audited factors gives

$$e_{s0} = 1 + \lambda(m_s - \mu_0), \qquad e_{s1y} = 1 + \lambda\{m_s + (y - m_s)/\pi - \mu_0\},$$

which proves (12) on completed records. We next show $0 < m_0 < m_1 < 1$.

Let $z_s = P_*(Y = 1 \mid S = s)$, and temporarily write $v_0 = e_{s10}$, and $v_1 = e_{s11}$. The likelihood-ratio identities for the audited and unaudited probabilities imply

$$\frac{1}{e_{s0}} = \frac{1 - z_s}{v_0} + \frac{z_s}{v_1}.$$

Substituting the factor formulas and simplifying yields

$$z_s = m_s + \frac{\lambda m_s(1 - m_s)}{\pi e_{s0}}. \tag{48}$$

All factors are positive and $z_s \in (0,1)$. If $m_s \le 0$, (48) would give $z_s \le 0$. If $m_s \ge 1$, then

$$z_s - 1 = (m_s - 1)\frac{v_0}{e_{s0}} \ge 0.$$

Both contradict $z_s \in (0,1)$. Thus $m_s \in (0,1)$.

For fixed λ, define the conditional expected log factor

$$\begin{aligned} f(m; z) = {} & (1-\pi)\log\{1 + \lambda(m - \mu_0)\} \\ & + \pi(1-z)\log\{1 + \lambda[m - m/\pi - \mu_0]\} \\ & + \pi z \log\{1 + \lambda[m + (1-m)/\pi - \mu_0]\}. \end{aligned}$$

Its domain is the interval on which all three factors are positive. It is strictly concave in m. The reciprocal identity above gives $f_m(m_s; z_s) = 0$. Also

$$\frac{\partial f_m(m; z)}{\partial z} = (1 - \pi)\lambda \left(\frac{1}{v_0} - \frac{1}{v_1} \right) > 0.$$

The root of $f_m(\cdot; z)$ therefore increases strictly with z. The symmetric error model with $\eta_* < 1/2$ has $z_1 > z_0$. It follows that $m_1 > m_0$.

When $\pi = 1$, the ratio is independent of the report and equals

$$e(O) = \left(\frac{\mu_+}{\mu_0} \right)^Y \left(\frac{1 - \mu_+}{1 - \mu_0} \right)^{1-Y} = 1 + \frac{\delta}{\mu_0(1 - \mu_0)}(Y - \mu_0).$$

Thus (12) still holds, with m_S canceling from the expression.

The identity explains the relation to corrected prediction methods. For every fixed pair (m_0, m_1) and every binary-report law, independence of the audit gives

$$\mathbb{E}\left[m_S + \frac{A}{\pi}(Y - m_S) \right] = \mathbb{E}[Y] = \mu.$$

Whenever the factor in (12) is nonnegative for every possible record, its expectation is at most one under the null. The projection chooses m_0, m_1 and λ to maximize expected log evidence at its target distribution. To see the optimality directly, let f be any other nonnegative factor valid for every null. Since q_* is a null law,

$$\mathbb{E}_{p_*} \log \frac{f}{e} \le \log \mathbb{E}_{p_*} \frac{f}{e} = \log \mathbb{E}_{q_*} f \le 0.$$

Equality requires f/e to be constant on the completed-record support by strict concavity of the logarithm; equality in the null bound then forces that constant to be one. This proves the uniqueness assertion in Proposition 4.1. The two constructions therefore coincide when the audit-correction parameters are fully optimized for the same target distribution and contract.

E.2 The Least Accurate Procedure Determines Mean Information

First, $J(\eta)$ is nonincreasing in η. For $0 < \eta_1 < \eta_2 < 1/2$, independently flip each completed report with probability

$$\tau = \frac{\eta_2 - \eta_1}{1 - 2\eta_1}.$$

This randomization changes p_{μ_+, η_1} into p_{μ_+, η_2}. It leaves the gold mean and audit protocol unchanged, and maps every member of $\mathcal{Q}_0(d)$ to another member. The data-processing inequality for KL divergence therefore gives

$$J(\eta_2) \le D_{\mathrm{KL}}(T p_{\mu_+, \eta_1} \| T q^\star_{\eta_1}) \le J(\eta_1),$$

where T denotes the report randomization. The map J is also convex, because $p_{\mu_+, \eta}$ is affine in η, the null class is convex, and KL divergence is jointly convex.

For $0 < \pi < 1$, the positive definite Hessian of the two-variable null minimization and the implicit function theorem make $q^\star_\eta$ differentiable in η. Differentiating its optimal value, with the derivative through the null minimizer equal to zero, gives

$$J'(\eta_*) = \sum_o \frac{\partial p_{\mu_+, \eta_*}(o)}{\partial \eta} \log e(o) \le 0.$$

The inequality follows from monotonicity. Affine dependence on η then gives the exact relation

$$\mathbb{E}_{\mu_+, \eta} \log e = J(\eta_*) + (\eta - \eta_*) J'(\eta_*) \ge J(\eta_*), \qquad 0 < \eta \le \eta_*. \tag{49}$$

For $\pi = 1$, the factor depends only on Y, so the same conclusion holds with equality for every η.

We must also check larger win rates for the same fixed factor. Write $\Delta_\mu(\eta)$ for the slope in

$$\mathbb{E}_{\mu, \eta} \log e = \mathbb{E}_{\mu_+, \eta} \log e + (\mu - \mu_+)\Delta_\mu(\eta).$$

This slope is affine in η. The fixed-accuracy projection property in (34) gives $\Delta_\mu(\eta_*) \geq 0$. At $\eta = 0$, the report equals the gold label, so

$$\Delta_\mu(0) = (1-\pi)\log\frac{e_{10}}{e_{00}} + \pi\log\frac{e_{111}}{e_{010}} > 0.$$

Indeed, $m_1 > m_0$ makes the first ratio greater than one, and

$$e_{111} - e_{010} = \frac{\lambda}{\pi}\{1-(1-\pi)(m_1-m_0)\} > 0$$

makes the second ratio greater than one. Thus $\Delta_\mu(\eta) \geq 0$ for every $\eta \in [0, \eta_*]$. Full auditing gives the same result directly from the Bernoulli likelihood ratio. Combining this fact with (49) proves (21).

E.3 Uniform Error Bounds and Minimum Cost

Validity follows from the null projection property exactly as in Appendix A.5; it makes no assumption about the actual report error. Under every equilibrium technology in the rectangle, records are independent and have finite log-score range R_d. Equation (21) and Hoeffding's inequality therefore give the stated uniform power guarantee and an upper bound $\log(1/\alpha)/\overline{J}(d) + O(\sqrt{\log(1/\alpha)})$.

For completeness, let $\mathcal{V}$ be the four parameter pairs listed in the main text and define, for $t > 0$,

$$M_d(t) = \max_{(\mu,\eta)\in\mathcal{V}} \sum_o p_{\mu,\eta}(o) e^{-t\ell_d(o)}.$$

For fixed t, the expectation inside the maximum is affine in μ for fixed η and affine in η for fixed μ. Its maximum on the rectangle is therefore attained at a corner. Markov's inequality applied to $\exp(-tZ_n)$ gives

$$\sup_{\mu,\eta} \Pr_{\mu,\eta}\{Z_n < \log(1/\alpha)\} \leq \inf_{t>0} \exp\{t\log(1/\alpha)\} M_d(t)^n.$$

All four log-score means are at least $\overline{J}(d) > 0$, so $\log M_d(t) < 0$ for sufficiently small positive t and the bound provides a finite sample size. Hoeffding's moment bound shows that its optimized sample size is no larger than the Hoeffding choice.

For the lower bound, any uniformly powerful test must have power at least $1-\beta$ at (μ_+, η_*), and any robustly valid test must have size at most α under the independent null law q_*. The change-of-measure argument in Appendix A.7, applied to this pair, yields $n \geq \log(1/\alpha)/\overline{J}(d) - O(\sqrt{\log(1/\alpha)})$. This proves (22) for all tests, not only the proposed likelihood-ratio test, and also establishes the leading term for the Chernoff sample size.

For each fixed contract, the same deterministic n is used at every technology. Its worst expected cost is exactly $n\overline{k}(d)$. Hence a finite menu has optimal worst cost $\min_{d\in\mathcal{D}} \overline{k}(d) n^\star_{\text{unif}}(d)$. The remainder in (22) is uniform on a fixed finite menu, proving (23). For a uniformly high-effort contract,

$$\overline{k}(d) = g\pi + b + \pi r(1 - \underline{\eta}_H),$$

because the most accurate permitted review earns the most agreement bonuses. The information coefficient instead uses $\overline{\eta}_H$. These two extremes need not coincide: uniform power fixes a sufficiently large common n, after which that study must also be affordable at the accuracy yielding the largest payment.

Finally, when $\mu_0 = 1/2$, the existing fixed-accuracy local expansion with $\eta = \overline{\eta}_H$ gives

$$\overline{J}_\delta(d) = \frac{\delta^2}{2\overline{V}(d)} + o(\delta^2), \qquad \overline{V}(d) = \frac{1}{4} + \left(\frac{1}{\pi} - 1\right)\overline{\eta}_H(1-\overline{\eta}_H).$$

Thus the local worst-cost coefficient is $\overline{k}(d)\overline{V}(d)$, with the limits taken in the same order as in (40).

F CONTRACTS UNDER UNCERTAIN TECHNOLOGY

We give the payment limits and explicit formulas in Theorem 5.2. Recall that $\overline{h} = (\overline{c}_H - c_L)/(\eta_L - \overline{\eta}_H)$ is the bonus exposure required to induce high effort at every technology. At $t = \overline{h}$, define the limiting base payment, largest expected payment, and minimum audit fraction by

$$b_\star = [c_L - (1-\eta_L)\overline{h}]_+, \qquad w = b_\star + (1-\underline{\eta}_H)\overline{h}, \qquad \rho = \frac{\overline{h}}{M - b_\star}. \tag{50}$$

The last quantity is used when $M > b_\star + \overline{h}$. Often the agreement bonus already covers review costs, giving $b_\star = 0$.

A contract inducing participation and high effort throughout Θ exists if and only if $M > b_\star + \overline{h}$. When this holds, the feasible audit probabilities are $(\rho, 1]$, and w is approached as bonus exposure decreases to $\overline{h}$ from above. Set $a_H = 1/4 - v$. The infimum in (26) has the explicit value

$$\overline{F}^\star_H(g) = (g\pi_H + w)(a_H + v/\pi_H), \qquad \pi_H = \min\left\{1, \max\left\{\rho, \sqrt{\frac{wv}{ga_H}}\right\}\right\}. \tag{51}$$

When $\pi_H = \rho$, the value is an infimum approached from above. The gold-price threshold is

$$\pi_G = \max\left\{\rho, \frac{\sqrt{v}}{1/2 + \sqrt{v}}\right\}, \qquad g_G = \frac{w(a_H + v/\pi_G)}{a_H(1-\pi_G)}. \tag{52}$$

These formulas optimize the local coefficient. They do not exchange optimization over contracts with the statistical limits.

We prove the high-effort result and give the low-effort comparison. All local results in this section take $\mu_0 = 1/2$. For low effort, put $v_L = \eta_L(1-\eta_L)$ and $a_L = 1/4 - v_L$. Under high-effort feasibility, $M > c_L$. When $c_L > 0$, the optimal local coefficient among contracts inducing low effort throughout Θ is

$$F^\star_L(g) = (g\pi_L + c_L)(a_L + v_L/\pi_L), \qquad \pi_L = \min\{1, \sqrt{c_L v_L/(ga_L)}\}. \tag{53}$$

For $c_L = 0$ its infimum is gv_L. Under high-effort feasibility, high effort is cheaper than low effort for all sufficiently large g if

$$\rho < \frac{v_L - v}{1/4 - v}. \tag{54}$$

The reverse strict inequality favors low effort; the exact finite-g local formulas determine the comparison at equality.

F.1 PROOF OF THEOREM 5.2

We first characterize contracts inducing a common effort throughout the uncertainty set. Write $t = \pi r$. Truthful reporting remains optimal for every technology because both signal error rates remain below a. The difference between the high- and low-effort payoffs is

$$t(\eta_L - \eta_H) - (c_H - c_L).$$

Thus high effort is strictly preferred at every technology if and only if $t > \overline{h}$. The corner $(\overline{c}_H, \overline{\eta}_H)$ also gives the smallest high-effort participation payoff. Consequently, participation throughout Θ requires and is implied by

$$b \geq b_{\min}(t) := \left[\overline{c}_H - (1-\overline{\eta}_H)t\right]_+.$$

At $t = \overline{h}$,

$$b_{\min}(\overline{h}) = \left[c_L - (1-\eta_L)\overline{h}\right]_+ = b_\star.$$

At a fixed audit probability π, a value $t > \overline{h}$ can satisfy the payment cap if and only if $b_{\min}(t) + t/\pi \leq M$. The function on the left is continuous and strictly increasing: its two possible slopes are $1/\pi$ and $1/\pi - 1 + \overline{\eta}_H$, both positive. Such a value therefore exists if and only if

$$b_\star + \overline{h}/\pi < M.$$

There is some feasible audit probability in $(0, 1]$ exactly when $M > b_\star + \overline{h}$. Under this condition, the feasible audit probabilities are precisely $\pi \in (\rho, 1]$.

For any contract inducing high effort throughout Θ, the largest expected payment is

$$\sup_{\theta \in \Theta} \{b + t(1 - \eta_H)\} = b + t(1 - \underline{\eta}_H).$$

The minimum possible value at fixed t is $W(t) = b_{\min}(t) + t(1 - \underline{\eta}_H)$. This function is continuous and nondecreasing. Its slopes are $\overline{\eta}_H - \underline{\eta}_H \geq 0$ when participation binds and $1 - \underline{\eta}_H > 0$ otherwise. It follows that its infimum over $t > \overline{h}$ is $W(\overline{h}) = w$. At every fixed $\pi > \rho$, this value is approached by taking $t \downarrow \overline{h}$ and $b = b_{\min}(t)$. The cap holds for all sufficiently close choices.

Under the uniform power requirement, the common sample size has local information coefficient

$$V_+(\pi) = \frac{1}{4} + (\pi^{-1} - 1)\overline{\eta}_H(1 - \overline{\eta}_H) = a_H + v/\pi.$$

In particular, the same sample size is used at every technology. The worst expected total cost is exactly the common sample size times the largest expected per-item expenditure. The technology giving the largest payment need not be the one that determines the sample size. Therefore the high-effort local cost infimum is

$$\overline{F}^\star_H(g) = \inf_{\rho < \pi \leq 1} (g\pi + w)(a_H + v/\pi). \tag{55}$$

Expanding the objective gives

$$gv + wa_H + ga_H\pi + wv/\pi.$$

Its derivative is $ga_H - wv/\pi^2$, and its second derivative is $2wv/\pi^3 > 0$. Restricting its unique unconstrained minimizer to the closure $[\rho, 1]$ proves (51). The value at the lower endpoint is an infimum over feasible probabilities approaching that endpoint.

For the comparison with gold-only testing, full auditing costs $(g + w)/4 > g/4$. For any $\pi \in (\rho, 1)$,

$$(g\pi + w)(a_H + v/\pi) < g/4 \quad \Longleftrightarrow \quad g > \frac{w(a_H + v/\pi)}{a_H(1 - \pi)}.$$

Apart from the positive constant w/a_H, the expression on the right is

$$f(\pi) = \frac{a_H\pi + v}{\pi(1 - \pi)}.$$

Its derivative has the sign of $a_H\pi^2 + 2v\pi - v$. The unique zero in $(0, 1)$ is $\sqrt{v}/(1/2 + \sqrt{v})$, because $a_H + v = 1/4$. The minimum over the closure of the feasible interval is consequently at π_G, giving (52). Strict savings are possible exactly when $g > g_G$, including when the threshold is achieved only as $\pi \downarrow \rho$.

To verify monotonicity, consider nested rectangles with the same low-effort technology. Any contract that induces participation and high effort over the larger set does so over the smaller set. At each such contract, enlarging the set cannot decrease the largest expected payment or the largest high-effort error rate. Since $\eta(1 - \eta)$ is increasing for $\eta < 1/2$, it cannot decrease $V_+(\pi)$ either. The feasible set of contracts shrinks and each remaining local worst cost weakly increases. Thus its infimum cannot decrease. The characterization of strict gold-only savings then implies that g_G cannot decrease. If the larger rectangle admits no uniformly high-effort contract, both quantities can be taken as $+\infty$. □

F.2 Low Effort and the Large-Gold-Cost Comparison

Low effort is selected throughout Θ if and only if

$$t \leq \underline{h} := \frac{c_H - c_L}{\eta_L - \underline{\eta}_H}.$$

The weak inequality follows from the convention of choosing low effort at a tie. Participation then requires an expected payment of at least c_L. When $c_L > 0$, this amount is attained at any fixed

$\pi \in (0, 1]$ by taking $t > 0$ sufficiently small and setting $b = c_L - (1-\eta_L)t$. The base is nonnegative and $t \leq \underline{h}$. The cap also holds: high-effort feasibility implies $M > c_L$, because $b_\star + \overline{h} > c_L$ whether $b_\star = 0$ or $b_\star > 0$. Taking t sufficiently small makes $b + t/\pi < M$.

The same one-variable minimization as above, with payment c_L and error rate η_L, proves (53). When $c_L = 0$, setting $b = 0$ and taking $t \downarrow 0$ approaches zero payment while preserving low effort. Subsequently taking $\pi \downarrow 0$ gives the infimum gv_L.

For $c_L > 0$, the low-effort optimizer is interior for all sufficiently large g, and

$$F_L^\star(g) = gv_L + 2\sqrt{c_L v_L g a_L} + c_L a_L.$$

The high-effort unconstrained optimizer tends to zero, so it lies below ρ for all sufficiently large g. Equation (55) then yields

$$\overline{F}_H^\star(g) = g(v + a_H\rho) + w(a_H + v/\rho).$$

Hence

$$\frac{F_L^\star(g)}{g} \longrightarrow v_L, \qquad \frac{\overline{F}_H^\star(g)}{g} \longrightarrow v + (1/4 - v)\rho.$$

These limits also hold when $c_L = 0$. The strict inequality (54) makes high effort cheaper than low effort and gold-only testing for all sufficiently large g. The reverse strict inequality makes low effort cheaper than high effort. No conclusion at equality is needed: the finite-g formulas determine that case directly.

For a point rectangle $\Theta = \{(c_H, \eta_H)\}$ satisfying the earlier assumption $c_L < (1 - \eta_L)h$, one has $b_\star = 0$, $w = (1 - \eta_H)h$, and $\rho = h/M$. Thus the original known-technology comparison is recovered. More generally, the difference between (51) and its value for a point rectangle is the exact additional local cost of requiring the same contract and sample size to work throughout Θ, within the class of contracts that induce high effort at every technology.

For each fixed g and each positive tolerance, feasible contracts can approximate these infima to that tolerance. Keeping such a finite collection fixed while taking the statistical limits gives the same relation to the finite-menu result as in the known-technology case. In particular, none of the arguments exchanges the small-effect limit with optimization over audit probabilities.

G Uncertainty Experiment Details

These reference results use the projection factor alone. They reproduce the projection-only uncertainty comparison and provide the starting point for the finite-sample diagnostics. Appendix H.1 reports the expanded factor menu used in the main text.

G.1 Designs, Actual Responses, and Simulation

The projection-only reference experiments are reproduced by `Rscript simulations/run_uncertainty.R`, run from the source package directory. The script uses the unchanged `core.R` routines for record probabilities, the null projection, sample-size bounds, and simulation. The uncertainty experiment has its own results directory and does not overwrite the known-technology outputs. It uses 20,000 repetitions, seed 20260916, and the R random-number settings `Mersenne-Twister`, `Inversion`, and `Rejection`. The simulations were executed in R 4.5.1; session information is saved with the results.

For each nominal error and uncertainty level in (28), candidate audit probabilities are $0.02, 0.04, \ldots, 1$. High-effort candidates use $t = 1.0001\overline{h}$, $b = [\overline{c}_H - (1 - \overline{\eta}_H)t]_+$, and $r = t/\pi$, and are retained when the cap permits them. Low-effort candidates use a positive bonus no larger than 10^{-6}, small enough to satisfy $t \leq (\underline{c}_H - c_L)/(\eta_L - \underline{\eta}_H)$, and $b = c_L - t(1 - \eta_L) \geq 0$. We verify effort and participation at every technology corner. The rectangular constraints then ensure the same response throughout the set.

The high-effort score uses $\overline{\eta}_H$. Its sample size uses all four distributions in (20), including both ends of the alternative win-rate interval. For each candidate, we verify that the computed Chernoff bound is at most 0.2 and that the preceding integer sample size fails this bound. The selected contract

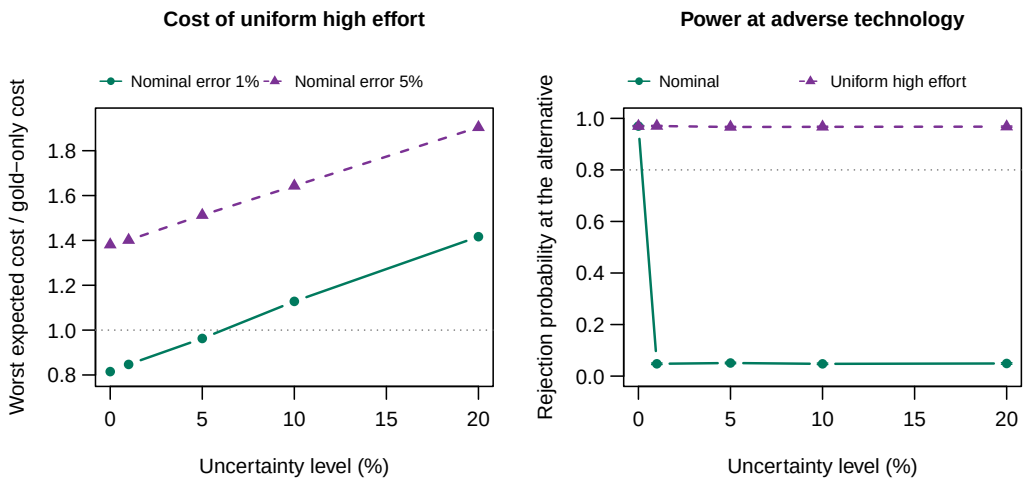

Figure 3: Projection-only uncertainty reference experiment. The horizontal axis gives the relative uncertainty width ω as a percentage.

minimizes the largest expected cost in the candidate menu. A second selection includes the exact gold-only design as an outside option. These are minima over the stated audit grid and payment rule; the finite-sample experiment does not optimize over every possible contract or every valid test.

To evaluate sensitivity, the nominal contract retains its score and sample size from $\omega = 0$. At each actual technology corner we recompute the provider's participation and effort by its utilities. A completed record is simulated from the resulting error rate, even if it differs from the one used to construct the test. Refusal would give zero rejection probability, no payment, and independently generated audits; there were no refusals in the reported settings. The information benchmark labeled oracle knows the actual technology before choosing a contract, sample size, or direct gold acquisition from the same grid. It is not an implementable design under the principal's information.

Independent record counts are drawn from the six-cell multinomial law, which is equivalent to drawing all gold labels, reports, and audits individually. This permits efficient simulation without changing the experiment. We simulate rejection probabilities at $\mu = 0.5$ and 0.6 for all technology corners. The four-corner Chernoff calculation, rather than this finite simulation grid, gives the uniform power guarantee. All costs include expected expert expenditure and provider payment; direct gold testing includes only expert expenditure.

G.2 Costs and Power

Table 5 gives the full numerical comparison behind Figure 3. The high-effort cost is the maximum over the rectangle, attained at its smallest error rate. The reported power is evaluated at the upper cost and error endpoints. The entries therefore answer different questions about the same fixed design.

At nominal error 0.01 and $\omega = 0.05$, the chosen contract has $(\pi, b, r) = (0.12, 0, 0.9696260)$ and uses 1,120 items. Its expected number of gold labels is 134.4, giving expert cost 1,344; the largest expected provider payment is 129.0146. The total 1,473.0146 is below the direct-gold cost 1,530. At the upper cost/error corner, its power estimate is 0.96650 with 95% Monte Carlo interval $[0.96392, 0.96891]$. At $\omega = 0.01$ the corresponding power is 0.97030 with interval $[0.96786, 0.97256]$, whereas the unchanged nominal contract has power 0.04770 with interval $[0.04483, 0.05074]$. These intervals use the Wilson formula in Appendix D; they describe simulation error in rejection frequencies, not uncertainty about the provider's technology.

All ten uniformly high-effort designs have Chernoff-certified power at least 0.8, and their smallest simulated power across the technology corners is 0.9618. The exact gold-only power is 0.800748;

Table 5: Uniform high-effort designs and the effect of ignoring uncertainty. High and Nominal are empirical powers at the upper cost/error corner. Choice includes the gold-only outside option; H denotes high effort.

η_H^0	ω	Audit π	n	Worst cost	High	Nominal	Choice
0.01	0.00	0.12	956	1,246.84	0.9698	0.9700	H
0.01	0.01	0.12	992	1,295.88	0.9703	0.0477	H
0.01	0.05	0.12	1,120	1,473.01	0.9665	0.0508	H
0.01	0.10	0.14	1,130	1,725.94	0.9671	0.0475	Gold
0.01	0.20	0.16	1,234	2,167.31	0.9678	0.0490	Gold
0.05	0.00	0.14	1,384	2,112.92	0.9659	0.9671	Gold
0.05	0.01	0.14	1,402	2,143.99	0.9657	0.3120	Gold
0.05	0.05	0.16	1,330	2,314.22	0.9618	0.3075	Gold
0.05	0.10	0.18	1,286	2,513.91	0.9645	0.3089	Gold
0.05	0.20	0.22	1,219	2,913.43	0.9623	0.3090	Gold

simulated frequencies for that design can fluctuate below 0.8. Design feasibility uses the analytic guarantee, not whether a Monte Carlo point estimate exceeds the target.

The oracle comparison shows the cost of choosing one design before learning the technology. With nominal error 0.01 and $\omega = 0.05$, its costs at the four corners range from 1,241.85 to 1,471.78, compared with worst cost 1,473.01 for the common high-effort design. At $\omega = 0.10$, the oracle uses a high-effort contract costing 1,197.71 at the most favorable corner and direct gold costing 1,530 at the least favorable corner; the common design selects direct gold. At nominal error 0.05, the oracle selects gold at every evaluated corner. These comparisons use the same contract grid and power bounds, with additional information available only to the oracle.

G.3 Verification and Saved Results

For each selected score, the script computes the largest possible null expectation analytically, maximizing over both reporting choices for each gold outcome and over the null win-rate endpoints. The computed scores use the normalization in Appendix C.2 with a 10^{-12} buffer. The largest numerical null expectation is no greater than one. The script also reconstructs the factor from its audit-correction representation; the largest discrepancy among the ten selected designs is 2.1×10^{-11}. These floating-point checks support the implementation of the proved guarantees.

The directory `simulations/results_uncertainty` contains: `designs.csv` (contracts, sample sizes, bounds, and cost components), `sensitivity.csv` (all corner simulations, including the oracle comparison), `recommendations.csv` (the design selected with gold as an option), `upper_corner_power.csv` (the power curves), `diagnostics.csv` (numerical checks), and `parameters.csv` (inputs and seed). The execution report records the actual run settings and session information. The projection-only uncertainty figure is generated from these same results.

H Additional Experiments and Finite-Sample Comparisons

H.1 From Local Costs to the Implemented Design

For a fixed contract and factor, the finite price criterion follows from

$$N(d,e)(g\pi_d + w_d) < gN_G \quad \Longleftrightarrow \quad g\{N_G - N(d,e)\pi_d\} > N(d,e)w_d. \tag{56}$$

For each uncertainty setting, let d_C be the contract selected using the projection factor and its four-corner Chernoff bound, with size N_C. Write $L = \log(1/\alpha)$ and $J_G = D_{\mathrm{Ber}}(\mu_+, \mu_0)$. The leading cost ratios and the actual certified ratio are

$$R^\star_{\mathrm{loc}} = \frac{\overline{F}^\star_H(g)}{g/4}, \qquad R_{\mathrm{loc}}(d_C) = \frac{\overline{k}(d_C)\overline{V}(d_C)}{g/4},$$
$$R_{\mathrm{KL}}(d_C) = \frac{\overline{k}(d_C)J_G}{g\overline{J}(d_C)}, \qquad R_{\mathrm{cert}} = \frac{N_C\overline{k}(d_C)}{gN_G}.$$

They give the exact multiplicative decomposition

$$R_{\mathrm{cert}} = R^\star_{\mathrm{loc}} \frac{R_{\mathrm{loc}}(d_C)}{R^\star_{\mathrm{loc}}} \frac{R_{\mathrm{KL}}(d_C)}{R_{\mathrm{loc}}(d_C)} \frac{N_C\overline{J}(d_C)}{L} \frac{L}{J_G N_G}. \tag{57}$$

The second factor is 1.004–1.088; it includes the difference between the local and finite-sample contract choices, as well as the audit grid and strict-incentive margin. The third is 1.075–1.618 and measures the change from local to exact information, including the gold denominator. The fourth is 3.078–3.730 and the fifth is 0.9724. The fourth factor includes the finite power requirement and the conservatism of the power bound. It is not a measurement of Chernoff looseness alone. Their product reproduces every finite cost ratio.

The theorem takes $\alpha \downarrow 0$ before $\delta \downarrow 0$. Table 6 pairs each of six decreasing α values with its finite cost ratio. They approach the exact-information ratio 0.4007. The exact-information ratios

Table 6: Six error levels and their corresponding finite cost ratios at the fixed 5% point-error contract and $\delta = 0.1$.

α	0.05	10^{-4}	10^{-12}	10^{-40}	10^{-100}	10^{-200}
Cost ratio	1.381	0.749	0.551	0.467	0.438	0.425

at $\delta = 0.1, 0.05, 0.02, 0.01$ are $0.4007, 0.3510, 0.3342, 0.3317$, approaching the local ratio 0.33085. Thus the finite results agree with the stated limits, while those limits need not approximate the design at $\alpha = 0.05, \delta = 0.1$.

Table 7: Complete local-to-finite comparison. Local is the optimized local coefficient, KL uses exact information at the projection design, and Projection uses its finite certificate. Menu adds finite-sample factor selection. The gold price is 10; g_G concerns local limits and g_{cert} the expanded finite menu.

η_H^0	ω	Local	KL	Projection	Menu	g_G	g_{cert}	Menu n
0.01	0.00	0.155	0.272	0.815	0.653	0.156	1.29	766
0.01	0.01	0.164	0.279	0.847	0.675	0.165	1.43	790
0.01	0.05	0.202	0.305	0.963	0.761	0.203	2.18	885
0.01	0.10	0.249	0.341	1.128	0.891	0.256	4.07	893
0.01	0.20	0.343	0.408	1.417	1.124	0.390	∞	979
0.05	0.00	0.326	0.401	1.381	1.095	0.322	∞	1097
0.05	0.01	0.334	0.406	1.401	1.110	0.334	∞	1111
0.05	0.05	0.364	0.431	1.513	1.202	0.384	∞	1057
0.05	0.10	0.401	0.462	1.643	1.310	0.456	∞	1025
0.05	0.20	0.477	0.525	1.904	1.517	0.641	∞	971

The expanded menu retains the reference contracts and projection factors. Nine settings choose the projection betting coefficient multiplied by 0.75; the largest 5% error range chooses 0.5. Each factor

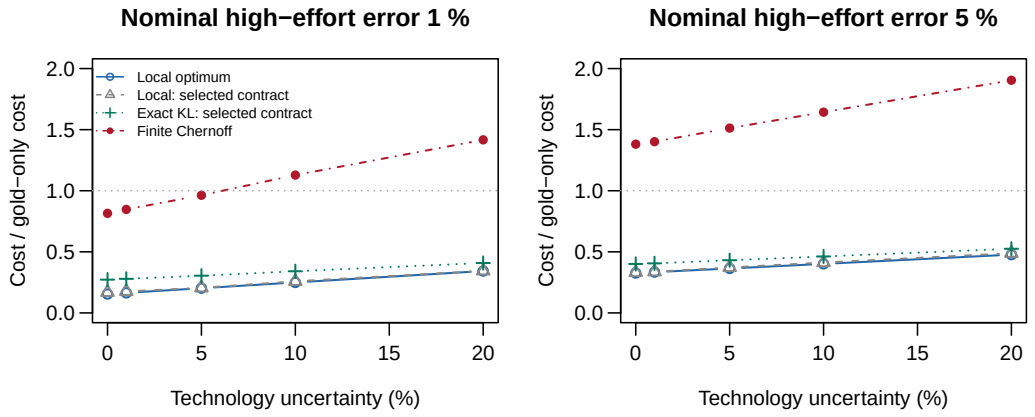

Figure 4: Local, exact-information, and finite projection-only costs. The rightmost comparison uses the actual finite-sample power bound.

is checked for positivity and null expectation at most one, and its power bound uses all four alternative corners. The selected high-effort audit probabilities are $0.12, 0.12, 0.12, 0.14, 0.16$ for nominal error 0.01, and $0.14, 0.14, 0.16, 0.18, 0.22$ for nominal error 0.05. The finite price thresholds for the first four 1% settings are $1.2851, 1.4256, 2.1783, 4.0655$; all others are infinite. Hence the finite criterion chooses high effort in four settings and direct gold in six. The largest null frequency among the 36 simulated checks is 0.008, and upper-corner powers are 0.964–0.969. These frequencies supplement the analytic guarantee.

Run `Rscript simulations/revision_diagnostics.R` for the decomposition and ordered-limit calculation, and `Rscript simulations/revision_score_menu.R` for factor selection. Their results directories contain the complete candidate menus, bounds, selected designs, and Monte Carlo intervals.

H.2 Prediction-Powered Baselines

We compare the tests using the same reports and audits, under the contracts selected at $\eta_H = 0.01$ and 0.05. Their audit probabilities are 0.12 and 0.14. An audited item belongs to the labeled group $\mathcal{L}$ and an unaudited item belongs to the unlabeled group $\mathcal{N}$. Write $n_{\mathcal{L}}$ and $n_{\mathcal{N}}$ for their sizes. Conditional on these sizes, the groups are independent samples from the same population.

The PPI estimator is

$$\widehat{\mu}_{\mathrm{PPI}} = \overline{Y}_{\mathcal{L}} + \overline{S}_{\mathcal{N}} - \overline{S}_{\mathcal{L}}, \qquad \widehat{\mathrm{se}}^2_{\mathrm{PPI}} = \frac{\widehat{\mathrm{var}}_{\mathcal{L}}(Y - S)}{n_{\mathcal{L}}} + \frac{\widehat{\mathrm{var}}_{\mathcal{N}}(S)}{n_{\mathcal{N}}}.$$

PPI++ uses

$$\widehat{\mu}_{\mathrm{PPI++}} = \overline{Y}_{\mathcal{L}} + \widehat{\kappa}(\overline{S}_{\mathcal{N}} - \overline{S}_{\mathcal{L}}), \qquad \widehat{\kappa} = \frac{\widehat{\mathrm{cov}}_{\mathcal{L}}(Y, S)}{(1 + n_{\mathcal{L}}/n_{\mathcal{N}})\widehat{\mathrm{var}}_{\mathcal{L}\cup\mathcal{N}}(S)}.$$

Its squared standard error is $\widehat{\mathrm{var}}_{\mathcal{L}}(Y - \widehat{\kappa}S)/n_{\mathcal{L}} + \widehat{\kappa}^2\widehat{\mathrm{var}}_{\mathcal{L}\cup\mathcal{N}}(S)/n_{\mathcal{N}}$. These formulas follow the published mean-estimation algorithms. Empirical variances and covariances use their population divisors; the PPI++ weight is unrestricted, as allowed for mean estimation. Both procedures reject when $\widehat{\mu} - \mu_0 > z_{1-\alpha}\widehat{\mathrm{se}}$, where $z_{1-\alpha}$ is the standard normal quantile. They use no additional labels. These are asymptotic Wald tests; the simulations do not give them a uniform finite-sample validity guarantee.

For the prediction-powered e-value comparison, we apply the audit correction to $1 + \lambda(Y - \mu_0)$ and use three choices: the raw prediction $m_S = S$ with λ optimized for expected log growth; the same prediction with $\lambda = \{2(\mu_0 + 1/\pi - 1)\}^{-1}$, half its maximal allowed value; and the nominal conditional mean $m_S = \mathbb{E}_{\mu_+,\eta_H}[Y \mid S]$ with λ optimized. The six resulting factors are checked for nonnegativity and null expectation at most one. Jointly optimizing m_0, m_1, λ gives precisely our

Table 8: Comparison at the same 400-item budget. Each entry is false-positive frequency / power, estimated from 20,000 studies. Wald and e-value methods have different validity guarantees.

Test	$\eta_H = 0.01$	$\eta_H = 0.05$
PPI (Wald)	.0509 / .9473	.0522 / .8132
PPI++ (Wald)	.0593 / .9663	.0559 / .8743
Optimally tuned PP e-value	.0053 / .7090	.0061 / .5295
Raw prediction, tuned bet	.0029 / .7145	.0051 / .5000
Raw prediction, fixed bet	.0000 / .1584	.0008 / .2972
Nominal conditional mean, tuned bet	.0027 / .7109	.0057 / .5137
Finite score menu	.0015 / .6037	.0028 / .4511

Table 9: Expected log growth and Chernoff-certified sample size for the prediction-powered e-values. Lower expected log growth can still give a smaller finite-sample certificate.

	$\eta_H = 0.01$		$\eta_H = 0.05$	
Score	Log growth	n	Log growth	n
Optimally tuned	.009645	956	.007672	1,384
Raw prediction, tuned bet	.009118	977	.007139	1,452
Raw prediction, fixed bet	.005715	876	.005889	1,128
Nominal conditional mean, tuned bet	.009168	972	.007345	1,415
Finite score menu	.008232	766	.006955	1,097

projected likelihood ratio; it is the same test, rather than a distinct baseline. We also use the finite score menu in Algorithm 1: the projected factor e, the two raw-prediction factors, and $1 + \gamma(e - 1)$ for $\gamma \in \{0.5, 0.75, 1, 1.25, 1.5\}$ on completed records, keeping the refusal factor at zero. We discard factors that are not positive on the observation support or lack positive expected log growth at an alternative corner. All remaining factors have null expectation at most one. For each contract we choose the factor with the smallest Chernoff-certified sample size before collecting any data. The projected score is always a candidate.

In Table 8, the Wald tests have substantially higher power, while PPI++ exceeds the nominal false-positive level in both settings. The e-tests are more conservative here; this comparison does not establish a universal cost of finite-sample validity. In particular, expected-log optimality does not order rejection probabilities at a given sample size.

Relative to the raw prediction with its bet optimized, joint tuning increases expected log growth by 5.8% and 7.5%. The corresponding Chernoff sample sizes fall by 2.1% and 4.7%. However, a smaller fixed bet gives an even smaller certificate (Table 9). The score optimization therefore has the expected-log and small-error interpretation proved in the theory; we do not claim finite-sample sample-size optimality. For the smaller bet, certified expected costs are 1,142.50 and 1,722.09, respectively, versus 1,246.84 and 2,112.92 for the expected-log-optimal score. The finite score menu selects $\gamma = 0.75$ for both contracts, reducing the certified expected costs further to 999.03 and 1,674.77. These are improvements within the stated candidate menu, not claims of global finite-sample optimality. A smaller certificate also need not increase power at every smaller budget, as the 400-item comparison shows.

The accompanying code also reports power curves, lower-accuracy and biased-report experiments, and a separate empirical comparison at approximately 80% power. For the latter, we select an item count on the grid $100, 150, \ldots, 1600$ using one simulation run, then evaluate it using independent simulated studies. The chosen budgets are descriptive estimates, rather than power certificates. All probabilities use 20,000 repetitions, with random seed 20260917. The count-based implementation of PPI and PPI++ is checked against their direct sample formulas.

H.3 LLMBAR DATA, CALIBRATION, AND EMPIRICAL REPLAY

We downloaded the authors' public LLMBar repository at commit `900616bff90b6c6c8e1681f7d079250637c55992`. The accompanying files preserve the upstream MIT license, URLs, and SHA-256 hashes. We use all five subsets and the released GPT-4 `Vanilla` and `Metrics_Reference` judgments. The first is called Vanilla+Rules in the source paper; the second is Rules+Metrics+Reference. The item text and gold label match across procedures. Every order-specific judgment is binary, so no observations are removed.

An item's SHA-256 digest is computed from its instruction and original candidate outputs. The first 209 items in digest order form the calibration set, and the remaining 210 form the evaluation set. The split does not use gold labels or judgments. Both output orders remain in the same split. In each order, Y indicates that the first presented output is preferred according to the gold label, and S is the corresponding judge decision. Thus each split contains equally many gold-zero and gold-one records. Table 10 reports the empirical errors, including their asymmetry.

Table 10: LLMBar error rates. Each calibration item contributes two orders, but the item is the unit of splitting and resampling for the pilot uncertainty diagnostic.

Split	Procedure	Overall error	$\Pr(S = 1 \mid Y = 0)$	$\Pr(S = 0 \mid Y = 1)$
Calibration	Vanilla+Rules	0.17225	0.18182	0.16268
Calibration	Metrics+Reference	0.13397	0.16268	0.10526
Evaluation	Vanilla+Rules	0.13810	0.15238	0.12381
Evaluation	Metrics+Reference	0.12143	0.14286	0.10000

All contract and score choices use calibration data only. The costs $c_L = 0.02$, $c_H = 0.04$, $g = 10$, and $M = 1$ are assumed normalized values. They are not measured API prices or private costs. With the joint log-growth score, the cheapest certified low-effort design on the original audit grid has $\pi = 0.06$, $n = 5864$, and cost 3635.68. The cheapest nominal high-effort design has $\pi = 0.54$, $r = 0.9676894$, $b = 0$, $n = 806$, and cost 4717.15. Selecting a score from the finite menu preserves these contract choices and reduces the certified sizes to 4637 and 631, at costs 2874.94 and 3692.96. Both select half the joint score's betting fraction. Every design exceeds the direct-gold cost 1530. Applying either contract to the held-out overall error rates predicts low effort. The observed balanced-prevalence error advantage is only 0.01667; at that prevalence, high effort would require $\pi r > 1.2$, exceeding $M = 1$. A paired bootstrap of calibration items gives a 95% interval $[0.00957, 0.06938]$ for the error advantage; its lower endpoint would require exposure 2.09. This is a diagnostic bootstrap interval, not a finite-sample confidence guarantee. Moreover, both errors are estimated in this experiment, whereas the uncertainty theorem takes the low-effort error as known.

For the empirical replay, draw Y at the specified prevalence and then sample uniformly, with replacement, from held-out order-specific records having that label. Independently audit with the contract's probability. This preserves the two conditional error rates in Table 10; the six-cell multinomial simulation is exactly equivalent. The prevalences 0.5 and 0.6 define controlled benchmark populations, rather than measured win rates of a deployed model pair. The released judgments are reused; no LLM queries, human annotation, or payment experiment is performed. Reported costs concern a prospective test conditional on these available calibration labels; they do not charge for collecting a new pilot. Under the asymmetric empirical law the error advantage is $1/105 + \mu/70$. Thus exposure 1.2 is required at $\mu = 0.5$, which rules out a contract inducing high effort uniformly across the testing range when $M = 1$. More generally, at balanced prevalence the point-estimate incentive condition is $(c_H - c_L)/M < 0.0382775$ for calibration and $(c_H - c_L)/M < 0.0166667$ for evaluation. These crossings hold the measured gains fixed while varying an assumed cost; they are not estimates of private costs. Over $\mu \in [0.25, 0.75]$, the smallest held-out advantage is $11/840 = 0.0130952$, so uniformly inducing high effort at gap 0.02 would require exposure exceeding $0.02/(11/840) = 1.52727$. Participation and the payment cap must also hold; satisfying an incentive constraint alone does not establish a total-cost saving. The figure is generated by `Rscript simulations/revision_interpretation_figure.R`, which reads the saved designs and empirical gains without resimulating judgments.

Under the two fitted contracts, low effort remains optimal for every $\mu \in [0.25, 0.75]$: the larger exposure is 0.52255225, and its maximum gain from high effort is less than $0.01058 < c_H - c_L$.

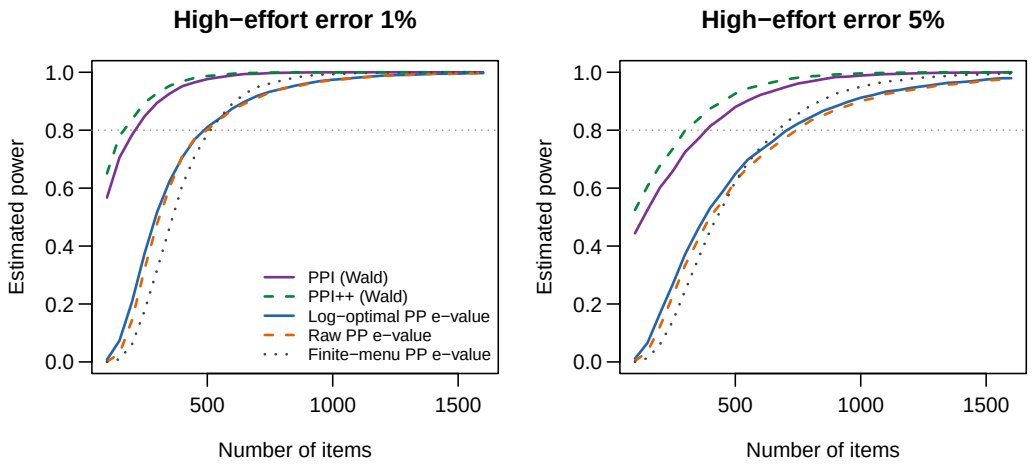

Figure 5: Synthetic power comparisons under the same contracts and item budgets. Expected-log optimality does not order finite-sample power.

Table 11 compares the methods under the low-effort contract. Scores are fitted on calibration data. Descriptive budgets are selected on the grid $400, 500, \ldots, 4000$ using 20,000 replications, requiring the lower endpoint of the Monte Carlo interval for power to exceed 0.8. A separate 20,000-replication batch evaluates each selected budget. These are empirical comparisons, not uniform power certificates. PPI and PPI++ use their published asymptotic Wald constructions. All methods use the same sampled records within each budget comparison. The seed is 20260918.

Table 11: LLMBar empirical replay under the low-effort contract. Direct gold costs 1530 at its exact 80%-power design. PPI and PPI++ have asymptotic validity; the betting methods have the finite-sample null guarantee.

Method	Items	Null rejection	Power	Expected cost
PPI	1600	0.05180	0.81245	992
PPI++	1300	0.05890	0.81650	806
Joint log-growth betting	2600	0.00735	0.82225	1612
Finite-menu betting	2900	0.00225	0.80895	1798
Raw report, tuned bet	2900	0.00795	0.81065	1798
Bayes prediction, tuned bet	2700	0.00765	0.81555	1674
Direct gold (exact)	153	0.05000	0.80075	1530

The nominal high-effort contract also induces low effort on these held-out distributions. At comparable empirical power its higher audit and payment costs make it less attractive for the betting methods: the joint score costs 2487.05 at $n = 425$ and power 0.81045, while the finite-menu choice has power 0.81395 at the same budget. These expected costs are evaluated at $\mu = 0.6$. Full tables and power curves for both contracts accompany the code. In a separate fitted symmetric-channel stress test, the joint-score certified sample sizes retain high power after the predicted effort changes: 0.98775 for the low-effort design and 0.96135 for the nominal high-effort design. Unlike the controlled sensitivity example, this real-data calibration does not exhibit a power collapse because held-out low-effort accuracy is close to the nominal high-effort accuracy.

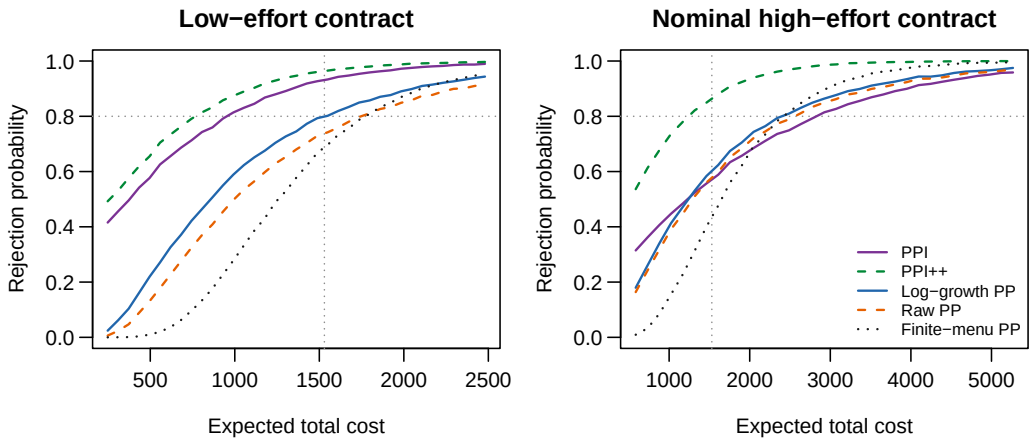

Figure 6: LLMBar empirical replay. Inference comparisons use held-out judgments, assumed costs, and the response predicted by the contracts.