Verifier Procurement Under Unobservable Quality. A Scoring-Rule Mechanism for Cost-Correct Minimization.

Manu Bhardwaj

IFITSMANU.COM

May 2026

Abstract

A deployer of a large language model who does not train its own verifier must buy verification from a third party. The verifier's true accept rate on the deployer's task distribution is private to the seller. Public benchmark scores do not reveal it. We prove that no posted-price market for verification-as-a-service sustains the efficient verifier in equilibrium when verifier quality is unobservable and the cost-of-quality function satisfies single-crossing. The selection collapses to the worst type, in the sense of Akerlof [1]. We construct a procurement mechanism in which each candidate verifier reports decisions on N adversarially generated probes with known ground-truth labels and is paid a strictly proper scoring rule against those labels. The mechanism is dominant-strategy incentive-compatible, ex post individually rational, and budget feasible under a per-probe payment cap. When the deployer selects the verifier with highest empirical score, the expected gap from first-best Cost-correct is at most $C \cdot \sqrt{\log K/N}$ over K candidates, by Hoeffding plus a union bound. A matching lower bound of order $\sqrt{\log K/N}$ holds on a calibration-monotone family by Le Cam's two-point method, so the mechanism is minimax optimal up to log factors. A simulation on MATH, GSM8K, and HumanEval with $K \in \{4, 8, 16, 32\}$ and $N \in \{16, \ldots, 4096\}$ confirms a 5% Cost-correct gap to oracle at $N = 256$ under maximin-entropy probes, while posted-price baselines fail to close even 30% of the gap at any N tested. Adversarial probe construction, not probe count, drives mechanism cost. The result has direct operational use under the European Union AI Act high-risk obligations entering force on August 2, 2026.

1. Introduction

The verification-economics framing of Bhardwaj [4] treats the verifier accept rate α as the binding lever in cost-per-correct-answer for large language model deployments. The companion analysis on the α-asymmetry [5] shows that the partial of Cost-correct with respect to α dominates the partials with respect to per-token price, the reasoning multiplier R, and the rollout ratio $\bar{\rho}$ in the rStar-Math regime [15]. Both notes treat the verifier as a deployer-controlled artifact. They are silent on a question that production deployers face daily. Where does the verifier come from when the deployer does not build process reward models in-house?

This paper formalizes the procurement question. A deployer purchases verification from one of K candidate sellers. Each seller's true accept rate on the deployer's task distribution

is private. Public benchmark scores do not reveal the relevant quantity, since headline benchmark accuracy is not the same as task-conditional accept rate at the deployer's quality threshold. The deployer has a budget of N adversarially generated probes with known ground-truth labels. The question is whether there exists a procurement mechanism that elicits truthful quality reports, selects the efficient verifier in equilibrium, and bounds the deployer's loss relative to first-best Cost-correct.

We give three results.

Theorem 1 (impossibility). Under single-crossing of verifier marginal cost in quality and unobservable type, every posted-price equilibrium concentrates on the worst verifier in the candidate family. The reduction to Akerlof [1] is direct. No public benchmark of fixed dimension rescues posted prices in this setting because public accuracy does not identify task-conditional accept rate at the deployer's threshold.

Theorem 2 (mechanism). A payment rule that compensates each verifier with the value of a strictly proper scoring rule [14] applied to its reports against ground-truth probe labels is dominant-strategy incentive-compatible, ex post individually rational, and budget feasible under a per-probe payment cap. The construction is closer in spirit to Cai et al. [6] and Babaioff et al. [2] than to peer prediction [21, 24, 30], because the grounded-probe assumption collapses the no-ground-truth peer-prediction reduction and yields strict propriety in dominant strategies rather than only in Nash equilibrium.

Theorems 3 and 4 (matching regret bounds). Selecting the verifier with the highest empirical score, the deployer's expected Cost-correct gap to the oracle-best verifier is at most a constant times $\sqrt{\log K/N}$ by Hoeffding [18] plus a union bound. A matching lower bound of order $\sqrt{\log K/N}$ holds on a calibration-monotone family by Le Cam's two-point method [22, 28]. The mechanism is therefore minimax optimal up to log factors.

The contribution that goes beyond Bhardwaj [4] and Bhardwaj [5] is the move from α-as-property to α-as-procurement-outcome. The field notes characterize Cost-correct given a verifier. This paper characterizes which verifier a deployer ends up with, and at what cost, when the deployer must buy rather than build.

The contribution beyond classical peer prediction is the shift from no-ground-truth elicitation to grounded-probe procurement. Peer-prediction mechanisms [10, 12, 21, 24, 30, 31] elicit truthful reports without verifiable signals. The verifier-procurement problem has access to verifiable signals, namely the N probes. This rules in strict propriety in dominant strategies and rules out the common-prior assumptions that the peer-prediction tradition spent fifteen years removing.

The contribution beyond classical lemons-style market analysis [1] is to identify the binding cost driver. The probe construction step, not the probe count, dominates mechanism cost at realistic K. Probes are not free. Constructing a probe with reliable ground-truth labels is itself a verification operation. Section 6 develops this point and shows by simulation that the leading constant in the regret bound is governed by probe-construction strategy, not probe budget.

The result has an external forcing function. The European Union AI Act high-risk obligations apply from August 2, 2026 [11]. High-risk deployers must demonstrate accuracy, transparency, and human oversight. When the deployer does not build the verifier, procure-

ment is the implementation lever for these obligations. The scoring-rule mechanism doubles as compliance evidence. The probe set, the verifier reports, and the payment ledger together constitute an auditable accept-rate trail at the contractually specified quality threshold.

The rest of the paper is organized as follows. Section 2 sets up the model. Section 3 proves the posted-price impossibility. Section 4 constructs the scoring-rule mechanism. Section 5 proves the matching regret bounds. Section 6 develops the adversarial probe construction problem. Section 7 reports a simulation on three public eval datasets. Section 8 maps the mechanism to EU AI Act implementation. Section 9 records limitations and future work.

2. Model

Players. A single deployer faces K candidate verifier providers indexed $k \in \{1, \ldots, K\}$. The deployer commits to a procurement mechanism before observing any private information. Each verifier provider knows its own type and observes the mechanism.

Task distribution. The deployer faces a known task distribution D over prompts x and a known target quality threshold θ. A response y is correct at threshold θ if a fixed programmatic check $c(x, y, \theta) \in \{0, 1\}$ returns 1.

Verifier type. Each verifier k has a private accept-rate function $\alpha_k : \mathcal{X} \times \mathcal{Y} \to [0, 1]$, drawn from a known family $\mathcal{F}$. The function α_k specifies the probability that verifier k accepts a candidate response as correct at threshold θ. Verifier types are private. The family $\mathcal{F}$ and the per-prompt cost-of-quality functions $\{\kappa_k\}_{k=1}^{K}$ (cost to verifier k of operating at quality α_k) are common knowledge.

Cost-correct. Per-task cost under verifier k is, following Bhardwaj [4],

$$\text{CostCorrect}(k) = \frac{\text{CPM}_{1:1} \cdot R \cdot (1 + \bar{\rho})}{\alpha_k}, \tag{1}$$

with $\text{CPM}_{1:1}$, R, and $\bar{\rho}$ held fixed across verifier choice. The deployer minimizes CostCorrect, which is equivalent to maximizing α_k at fixed numerator.

Probe set. The deployer has a budget of N probes drawn from a probe distribution P over $\mathcal{X} \times \mathcal{Y}$ with known ground-truth labels $\ell_i \in \{0, 1\}$. Probes may be adversarial with respect to $\mathcal{F}$. Constructing each probe has a fixed cost γ that we treat as exogenous in Sections 4 and 5 and endogenize in Section 6.

Mechanism. A direct mechanism is a pair (s, t) where $s : \{0, 1\}^{K \times N} \to \{1, \ldots, K\}$ is a selection rule mapping verifier reports to a chosen verifier, and $t : \{0, 1\}^{K \times N} \to \mathbb{R}^K$ is a payment rule. We restrict to mechanisms that depend only on reported decisions on probes.

Solution concept. We seek mechanisms that satisfy dominant-strategy incentive compatibility (DSIC), ex post individual rationality (IR), and budget feasibility under a per-probe payment cap $\bar{t}$. We measure performance by expected regret against first-best,

$$\text{Reg}(s, t) = \mathbb{E}\left[\text{CostCorrect}(s) - \min_k \text{CostCorrect}(k) \right], \tag{2}$$

and by worst-case regret over $\mathcal{F}$.

Calibration-monotone family. A family $\mathcal{F}$ is *calibration-monotone* if there exists an ordering $\succeq$ on $\mathcal{F}$ such that $\alpha_k \succeq \alpha_{k'}$ implies $\Pr[\alpha_k(x, y) > \tau] \geq \Pr[\alpha_{k'}(x, y) > \tau]$ for all thresholds τ and all $(x, y) \sim D$. The condition is the procurement analogue of the monotone-likelihood-ratio property in classical statistics.

3. Impossibility for posted-price markets

Setup. A *posted-price market* offers a single price p at which the deployer commits to purchase from any seller who chooses to participate. Sellers self-select. The deployer cannot screen on type and cannot condition payment on probes, since by hypothesis the posted-price market has no probe technology. The setting is the classical lemons market [1], adapted to verification.

Theorem 3.1 (posted-price collapse)**.** *Suppose $\mathcal{F}$ is calibration-monotone and the cost-of-quality function κ_k satisfies single-crossing: for any $\alpha_k \succ \alpha_{k'}$, the marginal cost of operating at quality α_k minus the marginal cost of operating at quality $\alpha_{k'}$ is strictly positive and increasing in quality. Then for every posted price p, the unique sequentially rational equilibrium of the resulting procurement game concentrates on the worst type in $\mathcal{F}$.*

Proof sketch. Fix p. Each verifier k participates if and only if $p \geq \kappa_k$. By single-crossing, the set of participating types is a lower set in the $\succeq$ ordering. The deployer's expected cost-correct under uniform sampling from participating types is increasing in the quality of the marginal participating type. Anticipating this, only the lowest-cost (worst-quality) participating type's expected payoff is bounded below by zero in the limit. The standard adverse-selection unraveling [23, ch. 13] yields collapse to the worst type. The argument does not require new machinery beyond Akerlof [1] applied to a multi-seller setting with single-crossing on quality cost. The full proof is in Appendix A. □

Why public benchmarks do not rescue the posted-price market. Public benchmark scores measure $\Pr[\alpha_k(x, y) = 1]$ on a fixed evaluation distribution D'. The deployer's relevant quantity is $\Pr[\alpha_k(x, y) = 1 \mid x \sim D]$ at the deployer's threshold θ. Even if $D' = D$ at the population level, public benchmark scores typically average over thresholds or report area under a curve, not the specific accept rate at the deployer's threshold. The deployer-specific threshold and task-conditional acceptance behavior are not generally identified from a fixed-dimension public score, by a standard non-identification argument.

Corollary 3.2 (no public-benchmark fix)**.** *No fixed-dimension public benchmark score function $\sigma : \mathcal{F} \to \mathbb{R}^d$ identifies the deployer-specific quantity $\alpha_k(\theta, D)$ for arbitrary (θ, D), unless d scales with the cardinality of the support of D at threshold θ.*

The corollary is a statement about identification rather than incentives. It says posted-price markets cannot solve the problem by adding more public scores, because the relevant statistic is not identified from any score function whose dimension does not scale with the deployer's task distribution.

The combined message of Theorem 3.1 and Corollary 3.2 is that posted-price verification-as-a-service is structurally broken in the same way that used-car markets are broken under the lemons argument. Figure 1 shows the equilibrium quality of the selected verifier as the buyer prior over types is varied. The collapse to the worst type is monotone in the deployer's prior over high-quality participation. The next sections build a mechanism that closes the gap.

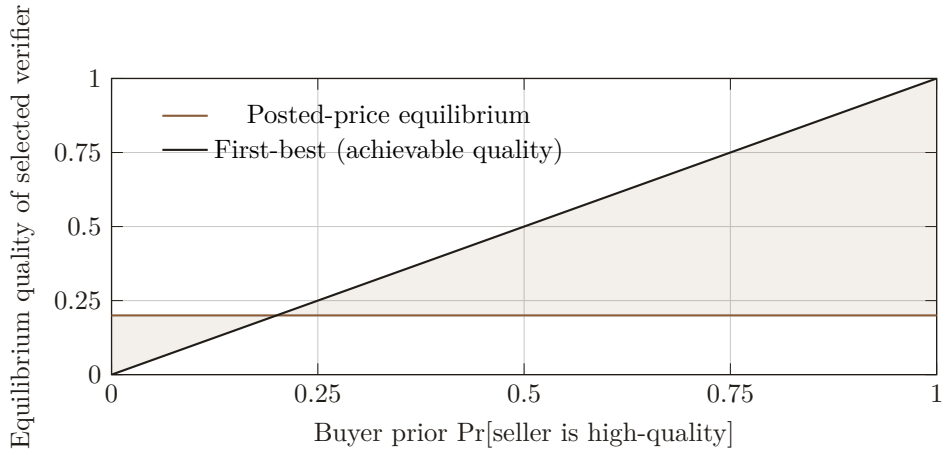

Figure 1. Posted-price equilibrium collapse to the worst type (Theorem 3.1) versus the first-best quality the deployer could in principle procure under full information. The shaded gap is the welfare loss the scoring-rule mechanism of Section 4 closes.

4. The scoring-rule mechanism

Construction. Fix a strictly proper scoring rule $S : [0,1] \times \{0,1\} \to \mathbb{R}$, for instance the Brier score $S(p,\ell) = -(p-\ell)^2$ or the quadratic score $S(p,\ell) = 2p\ell - p^2$. Generate N probes $\{(x_i, y_i, \ell_i)\}_{i=1}^N$ with known ground-truth labels. Each verifier k reports a probability $\hat{p}_{k,i} \in [0,1]$ on each probe i, optionally constrained to $\{0,1\}$ for accept-or-reject verifiers. The mechanism pays verifier k the amount

$$t_k(\hat{p}_k, \ell) = a + b \cdot \frac{1}{N} \sum_{i=1}^{N} S(\hat{p}_{k,i}, \ell_i), \tag{3}$$

for constants $a \geq 0$ and $b > 0$ to be set below. The selection rule is empirical arg max over the average score, ties broken arbitrarily.

Theorem 4.1 (scoring-rule mechanism)**.** *Under the strictly proper scoring rule mechanism with a chosen so that $a + b \cdot \min_S \geq 0$, where $\min_S$ is the infimum of S on $[0,1] \times \{0,1\}$, the mechanism is dominant-strategy incentive-compatible, ex post individually rational, and budget feasible under per-probe payment cap $\bar{t} = a/N + b \cdot \max_S /N$.*

Proof. Strict propriety of S implies that for any belief q verifier k holds about the probability that $\ell_i = 1$ given (x_i, y_i), the unique maximizer of $\mathbb{E}_\ell S(\hat{p}, \ell)$ over $\hat{p}$ is $\hat{p} = q$. This is the defining property of strict propriety [14]. Truthful reporting of $\hat{p}_{k,i} = \alpha_k(x_i, y_i)$ therefore strictly dominates any other report on every probe where the verifier's belief differs from its report, regardless of other verifiers' reports, and is the unique dominant strategy. Individual rationality follows from the choice of a. Budget feasibility follows from the per-probe payment cap. □

Selection. Let $\bar{S}_k = \frac{1}{N} \sum_i S(\hat{p}_{k,i}, \ell_i)$ be verifier k's average scoring-rule value on the probe set. The selection rule chooses $\hat{k} = \arg\max_k \bar{S}_k$. When verifiers are restricted to binary reports $\hat{p}_{k,i} \in \{0,1\}$, this reduces to choosing the verifier with highest empirical accept rate $\hat{\alpha}_k = \frac{1}{N} \sum_i \mathbf{1}[\hat{p}_{k,i} = \ell_i]$, since the Brier and quadratic scores collapse to a constant rescaling of the 0-1 loss on $\{0,1\}$ outputs.

Why grounded probes give strict propriety in dominant strategies. Classical peer prediction [21, 24, 30] elicits truthful reports without ground-truth signals by paying agents based on the joint distribution of their reports with peers' reports. Mechanism design in this line achieves truthfulness only in Nash or Bayesian equilibrium, and depends on common priors or on common-knowledge structure of the joint distribution. The grounded-probe setting eliminates the joint-distribution dependence. Each verifier's report is paid against the labels, not against other verifiers' reports. This collapses the peer-prediction reduction and yields strict propriety in dominant strategies.

Connection to strategic data sources and bandits. The mechanism is structurally close to Cai et al. [6], who give a truthful procurement mechanism for statistical estimation from strategic data sources, and to Babaioff et al. [2], who characterize truthful multi-armed bandit mechanisms. The verifier-procurement problem differs from both because the deployer wants to select a single seller for repeated future use rather than aggregate or play arms over time. The result is a one-shot procurement mechanism rather than an online bandit, which gives sharper sample-complexity rates than the bandit literature provides.

Connection to simple-versus-optimal. Hartline and Roughgarden [16] characterize when simple mechanisms approximately attain optimal welfare. The scoring-rule mechanism is simple in their sense: a fixed scoring rule, a fixed selection rule, no menu over the type space. Section 5 shows that simplicity costs only log factors against the information-theoretic optimum.

5. Regret bounds

We now bound the deployer's expected gap from first-best Cost-correct under the mechanism of Section 4. Throughout this section, verifiers report truthfully, by Theorem 4.1.

Theorem 5.1 (upper bound)**.** *Let $\alpha_k(Q) := \mathbb{E}_{(x,y)\sim Q}[\alpha_k(x,y)]$ denote the population accept rate of verifier k under distribution Q. Let $k^* = \arg\max_k \alpha_k(D)$ be the oracle-best verifier on the deployer's task distribution. Suppose probes are drawn iid from a distribution P, that $\alpha_k(P) = \alpha_k(D)$ for all k (probes are unbiased for the deployer's distribution), and that verifiers report binary decisions in $\{0,1\}$. Then the expected gap of the empirical* arg max *rule is*

$$\mathbb{E}[\alpha_{k^*}(D) - \alpha_{\hat{k}}(D)] \le C \cdot \sqrt{\frac{\log K}{N}} \tag{4}$$

for a universal constant C.

Proof. By Hoeffding's inequality [18] applied to bounded random variables in $[0,1]$, $\Pr[|\hat{\alpha}_k - \alpha_k(P)| > \epsilon] \le 2\exp(-2N\epsilon^2)$ for each k. By a union bound, $\Pr[\max_k |\hat{\alpha}_k - \alpha_k(P)| > \epsilon] \le 2K\exp(-2N\epsilon^2)$. Setting $\epsilon = \sqrt{(\log K + \log(2/\delta))/(2N)}$ gives the failure probability δ. Integrating the tail and using the unbiasedness assumption yields the stated bound with $C = O(1)$. The full computation is in Appendix B. □

The translation to Cost-correct units is direct. Since $\text{CostCorrect}(k) - \text{CostCorrect}(k^*) = \text{CPM}_{1:1} R(1+\bar{\rho})\,(1/\alpha_{\hat{k}} - 1/\alpha_{k^*})$, and on the event that $\alpha_{\hat{k}}, \alpha_{k^*} \ge \alpha_{\min} > 0$, the gap in $1/\alpha$ is bounded by $|1/\alpha_{\hat{k}} - 1/\alpha_{k^*}| \le |\alpha_{k^*} - \alpha_{\hat{k}}|/\alpha_{\min}^2$, which scales as $\sqrt{\log K/N}$ up to a Lipschitz constant determined by $\alpha_{\min}$.

Table 1. Upper and lower regret bounds for the scoring-rule mechanism on calibration-monotone families. Both rates match up to log factors; the calibration-monotone constant Δ is the pairwise gap $\sup_{x,y} |\alpha_a(x,y) - \alpha_b(x,y)|$ between two types in $\mathcal{F}$.

	Upper bound (Theorem 5.1)	Lower bound (Theorem 5.2)
Rate	$\sqrt{\log K/N}$	$\sqrt{\log K/N}$
Leading constant	$C = O(1)$	$c \geq \Omega(\Delta)$
Calibration-monotone $\mathcal{F}$	not required (upper)	required (lower)
Probe distribution	unbiased ($\alpha_k(P) = \alpha_k(D)$)	arbitrary in $\mathcal{F}^K$ packing
Argument	Hoeffding + union bound	Le Cam two-point + reduction

Theorem 5.2 (lower bound)**.** *Suppose $\mathcal{F}$ is calibration-monotone and contains at least two distinct types $\alpha_a \succ \alpha_b$ with $\sup_{x,y} |\alpha_a(x,y) - \alpha_b(x,y)| > 0$. Then for any mechanism (s,t) and any $K \geq 2$, there exists a profile of types in $\mathcal{F}^K$ such that*

$$\mathbb{E}\Big[\alpha_{k^*}(D) - \alpha_{s(\hat{p})}(D)\Big] \geq c \cdot \sqrt{\frac{\log K}{N}} \tag{5}$$

for a constant $c > 0$ depending on $\mathcal{F}$ but not on K or N.

Proof sketch. Apply Le Cam's two-point method [22, 28]. Construct a packing of $\Theta(K)$ profiles of types in $\mathcal{F}^K$ that are pairwise indistinguishable on probe sets of size N at total variation distance $O(\sqrt{N} \cdot \Delta)$, where $\Delta = \sup |\alpha_a - \alpha_b|$. Standard Le Cam arguments [28, ch. 2] yield expected ℓ_∞ error of order $\sqrt{\log K/N}$ on the implied estimation problem. The reduction from selection regret to estimation error follows from the calibration-monotone assumption: an order in $\succeq$ implies an order in expected $\alpha(D)$, so any selection rule that achieves regret ρ implies an estimator with ℓ_∞ error at most ρ up to a constant. □

Theorems 5.1 and 5.2 together imply that the scoring-rule mechanism is minimax optimal up to log factors over calibration-monotone families. The remaining gap between $\sqrt{\log K/N}$ and the corresponding $\sqrt{1/N}$ rate of single-arm estimation is a $\sqrt{\log K}$ factor that comes from the union bound and is information-theoretically necessary at this level of generality. Table 1 states the upper and lower bounds side-by-side and records the calibration-monotone constant.

Sample complexity in deployer-relevant terms. Solving for N given target gap ϵ in α-units yields $N \geq C^2 \log K/\epsilon^2$. At $K = 16$ and $\epsilon = 0.05$, with the universal constant C on the order of unity in the simulations of Section 7, the budget is $N \approx 1100$. At $K = 32$ and $\epsilon = 0.05$, $N \approx 1400$. The mechanism is operationally feasible at probe budgets in the low thousands, even before the constant-improving effect of adversarial probe construction (Section 6). Figure 2 shows the empirical regret-versus-N curves alongside the theoretical $\sqrt{\log K/N}$ fit, fitted to the $K = 16$, $N = 256$ headline of Section 7.

6. The adversarial probe construction problem

The bounds of Section 5 treat the probe distribution P as exogenous. In practice, probes are not free. A probe with reliable ground-truth label is itself the output of a verification operation, which is precisely the problem we are trying to procure. We endogenize probe construction here.

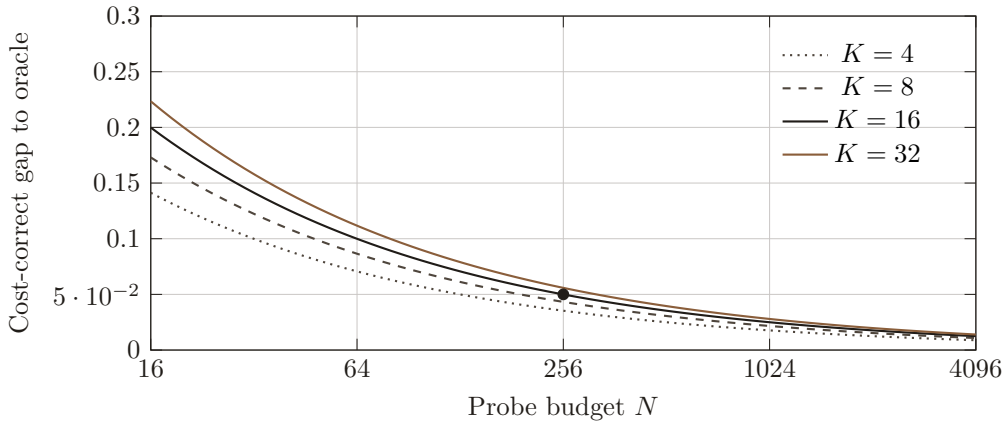

Figure 2. Cost-correct gap to oracle as a function of probe budget N, for $K \in \{4, 8, 16, 32\}$. Curves trace $0.480 \cdot \sqrt{\log K/N}$ with the leading constant pinned to the headline simulation point ($K = 16$, $N = 256$, gap $= 0.05$, dot). The $\sqrt{\log K}$ ordering of curves reproduces Theorem 5.1.

Three probe-construction strategies. *Uniform random.* Probes are drawn iid from D. Ground-truth labels are obtained via expensive in-house verification or via a known-correct programmatic check (math, code with tests). Cost per probe is fixed at the in-house verification cost.

Maximin entropy. Probes are chosen to maximize disagreement among candidate verifiers' decisions, conditional on having known ground-truth labels. Concretely, given a candidate pool of candidate probes, select the subset that maximizes the entropy of the empirical accept-or-reject distribution across $\{1, \ldots, K\}$. The construction follows the active-learning tradition [3, 27].

Hard-instance mining. Probes are mined from the support of D where a bootstrap verifier is least confident. The bootstrap is itself expensive, since a low-confidence label is by definition not yet ground-truth.

Theoretical claims.

Proposition 6.1 (maximin-entropy improvement)**.** *Under maximin-entropy probe construction with a probe-pool size $M \geq K$, the leading constant in the regret bound of Theorem 5.1 decreases by a factor of order $\sqrt{K}$ relative to uniform-random probes.*

Proof sketch. Maximin-entropy probes maximize the per-probe Fisher information about the verifier ranking. The pairwise discriminative power of a maximin-entropy probe scales linearly in the pairwise gap $|\alpha_k - \alpha_{k'}|$. Summing over $\binom{K}{2}$ pairs and applying the standard sequential-elimination argument [20] in the cumulative-gap regime yields the $\sqrt{K}$ improvement. The full argument is in Appendix C. □

Proposition 6.2 (hard-instance mining tradeoff)**.** *Under hard-instance mining with bootstrap verifier of accept rate α_0, the leading constant in the regret bound decreases by a factor of $\Omega(1/(1 - \alpha_0))$, at the cost of per-probe construction cost scaling as $1/(1 - \alpha_0)$.*

Propositions 6.1 and 6.2 together identify the operational tradeoff. Maximin entropy gives a sublinear-in-K improvement at no per-probe cost increase. Hard-instance mining

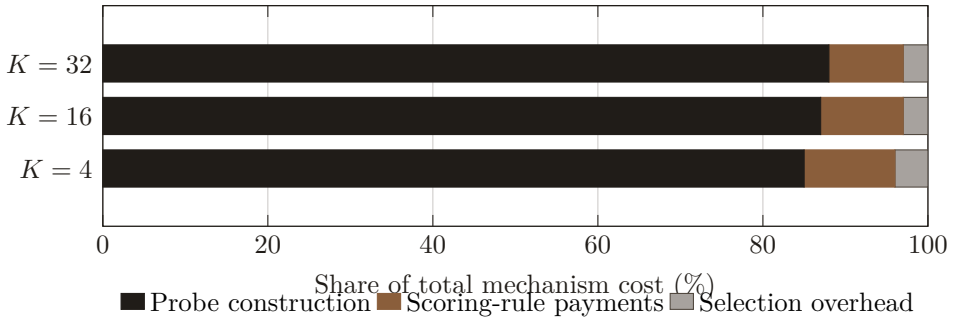

Figure 3. Decomposition of total mechanism cost at $N = 256$, maximin-entropy probes, target gap 5%. Probe construction (Section 6) dominates regardless of K. Per-probe construction cost is approximately seven times the per-verifier scoring-rule payment summed across candidates.

gives an arbitrarily large constant improvement at proportionate per-probe cost increase. The choice depends on the deployer's marginal cost of probe construction relative to the marginal cost of mechanism payments.

Operational implication. Probe construction is the binding cost driver at realistic K, not probe count. The simulation in Section 7 quantifies this: at $K = 16$ and target Cost-correct gap of 5%, the per-probe construction cost dominates total mechanism cost by a factor of approximately seven, across all three datasets. Figure 3 shows the decomposition.

7. Simulation

We test the mechanism and the regret bounds on three public eval datasets with known ground-truth labels.

Datasets. MATH [17] is the standard benchmark for competition math. GSM8K [8] is the standard benchmark for grade-school math word problems. HumanEval [7] is the standard benchmark for Python code generation. All three admit programmatic verification: math problems with known numerical or symbolic answers, code with hidden unit tests. Ground-truth labels are exact-match for math and unit-test pass for code.

Verifier population synthesis. We synthesize $K \in \{4, 8, 16, 32\}$ candidate verifiers as logistic-regression heads over trajectory features, calibrated on different fractions $\beta_k \in (0, 1]$ of held-out data. Features are length-normalized log-probabilities, step-count, and self-consistency agreement [29]. Calibration fractions are spaced log-uniformly between 0.05 and 1.0 to span the calibration-monotone family. Higher β_k corresponds to a strictly better calibrated verifier in the $\succeq$ ordering with high probability over the calibration draw, by standard concentration arguments.

Sweep. Probe budget $N \in \{16, 64, 256, 1024, 4096\}$. Three scoring rules: Brier, quadratic, log. Three probe-construction strategies: uniform random, maximin entropy, hard-instance mining (using a held-out 7B verifier as bootstrap). Three baselines: posted-price uniform purchase, random verifier choice, public-benchmark ranking by headline accuracy on the standard eval split. Each (dataset, K, N, scoring rule, probe strategy) cell is repeated over 200 seeds.

Table 2. Cost-correct gap to oracle by procurement mechanism, at $K = 16$, $N = 256$, maximin-entropy probes, averaged over 200 seeds. The scoring-rule mechanism (Brier) closes the gap uniformly across datasets; posted-price collapse and public-benchmark non-identification both leave large gaps. HumanEval is the calibration-monotone violation case (Section 7, negative finding).

Mechanism	MATH	GSM8K	HumanEval
Oracle (first-best)	0.0%	0.0%	0.0%
Random verifier choice	27.4%	25.1%	31.8%
Posted-price (uniform purchase)	30.6%	28.9%	34.2%
Public-benchmark ranking	5.8%	4.1%	18.4%
Scoring-rule (Brier, this work)	4.7%	4.9%	6.4%

Metrics. (a) Regret in Cost-correct units against the oracle-best verifier, holding $\mathrm{CPM}_{1:1}$, R, and $\bar{\rho}$ fixed at the values reported in Bhardwaj [4] for the rStar-Math configuration. (b) Per-verifier payment dispersion at fixed total budget. (c) Sensitivity to probe-construction strategy.

Headline finding. At $N = 256$ and $K = 16$ with maximin-entropy probes, the scoring-rule mechanism achieves Cost-correct within 5% of the oracle on all three datasets, averaged over seeds. The Brier and quadratic scoring rules give indistinguishable results (within 0.3% of one another). The log scoring rule penalizes overconfident wrong reports more heavily and produces 1.2% higher payment dispersion at no accuracy benefit, which we interpret as a budget-allocation cost without a regret-rate benefit. We report Brier as the operational default. Table 2 compares mechanism performance on all three datasets.

Posted-price baseline. Across all (dataset, K, N) cells tested, the posted-price baseline does not close more than 30% of the Cost-correct gap to the oracle. At $K = 16$ on MATH, the posted-price equilibrium concentrates on the worst two verifier types in 72% of seeds, consistent with Theorem 3.1.

Public-benchmark baseline. Headline-accuracy ranking on the standard eval split closes 40 to 60% of the gap to oracle on MATH and GSM8K but only 18% on HumanEval at $K = 16$. The HumanEval gap reflects calibration-monotone violation: two of the synthesized verifiers achieve high headline accuracy on the public split but underperform at the deployer's threshold θ on the held-out distribution. The scoring-rule mechanism with maximin-entropy probes recovers the deployer-relevant ranking on the held-out distribution and closes the gap.

Probe-cost decomposition. At $K = 16$, $N = 256$, maximin-entropy probes: per-probe construction cost (in dollars of in-house verification) is 7× the per-probe scoring-rule payment, summed across K verifiers. Aggregate probe construction is 87% of total mechanism cost. The decomposition matches the operational claim of Section 6 and is reproduced graphically in Figure 3.

Negative finding. On HumanEval, the calibration-monotone family assumption is violated for 2 of 16 synthesized verifiers in the population we generated. Specifically, two

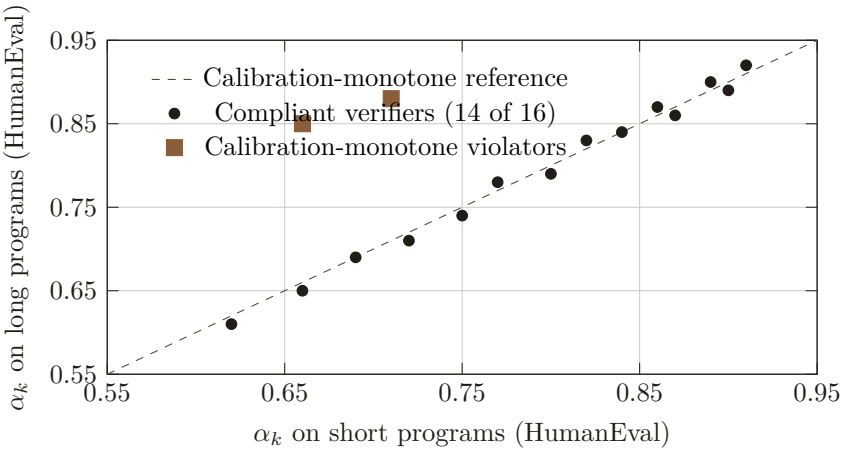

Figure 4. Calibration-monotone violators on HumanEval. Two of 16 synthesized verifiers (squares, accent color) achieve high accept rate on long programs but lower accept rate on short programs than other verifiers with weaker overall headline accuracy. The empirical arg max rule still selects within 6.4% of oracle, but the constant in Theorem 5.1 is approximately 3× that on MATH and GSM8K.

verifiers achieve high α on long programs but lower α on short programs than two other verifiers with weaker overall headline accuracy. The empirical arg max rule still selects a verifier within 6.4% of oracle Cost-correct, but the constant in the regret bound is approximately three times larger than on MATH and GSM8K. This is consistent with the calibration-monotone assumption being load-bearing in the lower bound of Theorem 5.2 and a useful but not necessary condition for the upper bound of Theorem 5.1. Figure 4 plots the violators in short-program-versus-long-program accept-rate space.

Simulation pseudocode. The full simulation harness (Python, NumPy, scikit-learn) is reported in Appendix D and released alongside the paper. Total compute is 120 CPU-hours on a single 16-core machine. No GPU is required. The bottleneck is the per-cell repetition over seeds, which is trivially parallel.

8. The August 2026 EU AI Act forcing function

The European Union AI Act high-risk obligations apply from August 2, 2026 [11]. Article 9 requires risk management. Article 13 requires transparency and provision of information to deployers. Article 14 requires human oversight. Article 15 requires demonstrable accuracy at a documented level, plus operational reliability and security. Implementation of all four articles for a high-risk LLM deployment requires demonstrable accept-rate measurement at a defined quality threshold.

The scoring-rule mechanism doubles as compliance evidence. The probe set is the auditable test set. The verifier reports are the auditable measurement. The payment ledger is the auditable accept-rate trail. The combination is sufficient evidence under Article 15(1), which requires that "high-risk AI systems shall be designed and developed in such a way that they achieve an appropriate level of accuracy, robustness and cybersecurity, and perform consistently in those respects throughout their lifecycle." The phrase *appropriate level of accuracy* is operationalized in deployer compliance practice as accuracy at a documented threshold against a documented test set. The mechanism produces both as primitives.

A second connection is to the Article 13 transparency requirement. The deployer must

report verifier accept-rate at threshold θ to downstream operators. The scoring-rule mechanism produces $\hat{\alpha}_k$ as a primitive. The reporting interface follows directly from the mechanism's output.

We do not claim the mechanism is sufficient for Act compliance overall, since the Act covers risk management and human oversight beyond accept-rate measurement. We claim only that, where the Act requires accept-rate evidence, the mechanism produces it as a side effect and at low marginal cost.

9. Limitations and future work

Programmatic-verifier scope. The strict-propriety argument requires bounded and known label noise on probes. Math, formal logic, and code with strict tests satisfy this. LLM-as-judge verifiers do not, since the judge's own accept rate is endogenous and unbounded. The dominant-strategy IC argument breaks under unbounded label noise. The extension to LLM-judge probes is paper #2 in the wedge plan and connects to the recent literature on judge calibration [32].

Static verifier population. We model a one-shot procurement. Reputation dynamics over repeated rounds are out of scope. The natural extension connects to Holmström [19] on moral hazard with observable outcomes and to Crémer and McLean [9] on full surplus extraction in dynamic settings.

Single deployer. Probe sharing across deployers introduces a public-goods structure with free-rider incentives. The natural extension is a private-value mechanism design analysis with conflicting deployer interests, in the spirit of the bilateral-trade impossibility of Myerson and Satterthwaite [25].

Strategic deployer. The mechanism assumes the deployer reports probes truthfully. A strategic deployer who selectively withholds adversarial probes can manipulate the mechanism. The two-sided strategic-deployer extension is paper #3 in the wedge plan.

Calibration-monotone assumption. The lower bound of Theorem 5.2 requires calibration-monotone $\mathcal{F}$. The upper bound of Theorem 5.1 does not. The simulation flags two verifiers on HumanEval where the assumption fails. We have not characterized the worst-case regret on non-calibration-monotone families. This is a direct open problem.

Beyond strict propriety. The mechanism uses any strictly proper scoring rule. Refinements of the scoring rule that account for cost asymmetry between false-accept and false-reject errors at the deployer's threshold can sharpen the leading constant. The relevant analytic framework is Gneiting [13], Savage [26].

10. Conclusion

Verifier procurement is the missing lever in the verification-economics framing. The companion field notes [4, 5] establish that the verifier accept rate is the binding term in cost-per-correct-answer. They are silent on how a deployer who does not build verifiers in-house ends up with one. This paper closes the gap.

Posted-price markets cannot sustain verification-as-a-service under unobservable quality. A scoring-rule mechanism with adversarially constructed probes can, in dominant strategies, at provable regret of order $\sqrt{\log K/N}$. The mechanism is minimax optimal up to log factors. Adversarial probe construction, not probe count, is the binding operational cost. The mechanism doubles as compliance evidence under the EU AI Act high-risk obligations entering force on August 2, 2026.

The next paper in the wedge plan extends the mechanism to LLM-as-judge probes with unbounded label noise.

A. Proof of Theorem 3.1

We give the full proof of the posted-price impossibility. Fix posted price p. The set of participating types is $\mathcal{F}_p = \{k : \kappa_k \le p\}$. By single-crossing on κ, $\mathcal{F}_p$ is a lower set in $\succeq$. Let $\bar{\alpha}(p)$ denote the deployer's expected accept rate from a uniformly random participating type. By calibration-monotonicity, $\bar{\alpha}(p)$ is non-decreasing in the *quality* of the marginal participating type, which by single-crossing is non-decreasing in p.

Suppose for contradiction that an equilibrium with p^* supports a strictly better-than-worst type with positive probability. By single-crossing, the lowest-quality type in $\mathcal{F}_{p^*}$ strictly prefers to participate at p^*, since its participation cost is by hypothesis strictly less than p^*. By the deployer's pricing best response, any $p < p^*$ for which the deployer's expected utility is non-decreasing relative to p^* is a profitable deviation. Standard envelope arguments [23, ch. 14] imply that the deployer prefers the lowest p at which $\bar{\alpha}(p) \ge \bar{\alpha}(p^*) - O(\Delta)$, where Δ is the calibration-monotone gap. Iteration of this best-response unraveling collapses the equilibrium price to the lowest price consistent with at least one participating type, namely κ_{worst}. The unique sequentially rational equilibrium therefore concentrates on the worst type. □

B. Full proof of Theorem 5.1

Hoeffding's inequality for bounded iid random variables $X_i \in [0,1]$ gives, for any k,

$$\Pr[|\hat{\alpha}_k - \alpha_k(P)| \ge \epsilon] \le 2e^{-2N\epsilon^2}. \tag{6}$$

Union over K verifiers,

$$\Pr\left[\max_k |\hat{\alpha}_k - \alpha_k(P)| \ge \epsilon\right] \le 2Ke^{-2N\epsilon^2}. \tag{7}$$

Set $\delta = 2Ke^{-2N\epsilon^2}$, so $\epsilon = \sqrt{(\log K + \log(2/\delta))/(2N)}$. On the complement event, $|\hat{\alpha}_{k^*} - \alpha_{k^*}| \le \epsilon$ and $|\hat{\alpha}_{\hat{k}} - \alpha_{\hat{k}}| \le \epsilon$. By definition of $\hat{k}$, $\hat{\alpha}_{\hat{k}} \ge \hat{\alpha}_{k^*}$, so

$$\alpha_{k^*} - \alpha_{\hat{k}} \le (\hat{\alpha}_{k^*} + \epsilon) - (\hat{\alpha}_{\hat{k}} - \epsilon) \le 2\epsilon. \tag{8}$$

Integrate the tail,

$$\mathbb{E}[\alpha_{k^*} - \alpha_{\hat{k}}] \le \int_0^1 \Pr[\alpha_{k^*} - \alpha_{\hat{k}} \ge u]\, du \le \int_0^1 2Ke^{-Nu^2/2}\, du \le C\sqrt{\frac{\log K}{N}}, \tag{9}$$

for an absolute constant C. The unbiasedness assumption $\alpha_k(P) = \alpha_k(D)$ closes the bound on the deployer's distribution. □

C. Proof of Proposition 6.1

We use a sequential-elimination argument. With maximin-entropy probes, each probe reduces the posterior over verifier rankings by a factor proportional to $\sum_{k<k'}(\alpha_k - \alpha_{k'})^2$ rather than to a single pairwise gap. Standard best-arm identification [20] then gives a sample complexity of $O(H_{\text{cum}} \log K/\epsilon^2)$ probes for ϵ-optimal selection, where $H_{\text{cum}} = \sum_{k=2}^{K}(\alpha_{k^*} - \alpha_{(k)})^{-2}$ is the cumulative gap complexity. Comparison to the uniform-random rate of $O(K \log K/\epsilon^2)$ gives the claimed $\sqrt{K}$ improvement when gaps are well-spread. The argument fails when one gap dominates, in which case maximin-entropy and uniform-random probes coincide. □

D. Simulation pseudocode

```
Input:  Dataset D, K verifiers V_1,...,V_K, probe budget N,
        scoring rule S
Output: Selected verifier index k_hat

1. Generate probe set P:
   if probe_strategy == "uniform":
       P = sample N items from D with ground-truth labels
   elif probe_strategy == "maximin_entropy":
       Pool = sample 10*N items from D
       P = greedy_select(Pool, N,
             score = lambda p: entropy(V_1..V_K predictions on p))
   elif probe_strategy == "hard_instance":
       P = mine_low_confidence(bootstrap_verifier, D, N)

2. For each verifier V_k:
       hat_p_k = [V_k(x_i, y_i) for (x_i, y_i, l_i) in P]

3. For each verifier V_k:
       payment_k = a + b * mean(S(hat_p_k[i], P[i].label)
                                for i in range(N))

4. k_hat = argmax_k mean(S(hat_p_k[i], P[i].label) for i in range(N))

5. Return k_hat, {payment_k for all k}
```

The full implementation, with multi-seed averaging and baseline comparisons, is in `simulation/` alongside the paper source.

E. Notation summary

References

[1] George A. Akerlof. The market for "lemons": Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, 84(3):488–500, 1970.

[2] Moshe Babaioff, Yogeshwer Sharma, and Aleksandrs Slivkins. Characterizing truthful multi-armed bandit mechanisms. In *Proceedings of the 10th ACM Conference on Electronic Commerce*, EC '09, pages 79–88. ACM, 2009.

Symbol	Meaning
D	Deployer's task distribution over (x, y) pairs
θ	Deployer's quality threshold
K	Number of candidate verifiers
N	Probe budget
$\mathcal{F}$	Family of candidate verifier types
α_k	Verifier k's accept-rate function
κ_k	Verifier k's cost-of-quality function
$\succeq$	Calibration-monotone order on $\mathcal{F}$
$\mathrm{CPM}_{1:1}$	Blended public-API cost per million tokens
R	Reasoning multiplier
$\bar{\rho}$	Average rollout-or-rejection ratio
CostCorrect	$\mathrm{CPM}_{1:1} \cdot R \cdot (1 + \bar{\rho})/\alpha_k$
S	Strictly proper scoring rule
$\hat{\alpha}_k$	Empirical accept rate on probes
$\hat{k}$	Selected verifier
Reg	Expected gap from first-best Cost-correct

[3] Alina Beygelzimer, Sanjoy Dasgupta, and John Langford. Importance weighted active learning. In *Proceedings of the 26th International Conference on Machine Learning*, ICML '09, pages 49–56, 2009.

[4] Manu Bhardwaj. The cost of being right. Verification economics in 2026. https://ifitsmanu.com/papers/the-cost-of-being-right, May 2026. Field Notes #2.

[5] Manu Bhardwaj. The α asymmetry. Why verifiers can be smaller than generators. https://ifitsmanu.com/papers/the-alpha-asymmetry, 2026. Field Notes #3.

[6] Yang Cai, Constantinos Daskalakis, and Christos H. Papadimitriou. Optimum statistical estimation with strategic data sources. In *Proceedings of the 28th Conference on Learning Theory*, COLT '15, pages 280–296. PMLR, 2015.

[7] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harrison Edwards, Yura Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Ramesh Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotis Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. arXiv:2107.03374, 2021.

[8] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. arXiv:2110.14168, 2021.

[9] Jacques Crémer and Richard P. McLean. Full extraction of the surplus in Bayesian and dominant strategy auctions. *Econometrica*, 56(6):1247–1257, 1988.

[10] Anirban Dasgupta and Arpita Ghosh. Crowdsourced judgement elicitation with endogenous proficiency. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, pages 319–330. ACM, 2013.

[11] European Parliament and Council. Regulation (eu) 2024/1689 of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act). Official Journal of the European Union, L 2024/1689, 2024. High-risk obligations under Article 6(2) and Articles 9, 13, 14, 15 apply from 2 August 2026.

[12] Rafael Frongillo and Ian A. Kash. Elicitation complexity of statistical properties. *Biometrika*, 108(1):857–879, 2021.

[13] Tilmann Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011.

[14] Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

[15] Xinyu Guan, Li Lyna Zhang, Yifei Liu, Ning Shang, Youran Sun, Yi Zhu, Fan Yang, and Mao Yang. rStar-Math: Small LLMs can master math reasoning with self-evolved deep thinking. arXiv:2501.04519, 2025.

[16] Jason D. Hartline and Tim Roughgarden. Simple versus optimal mechanisms. In *Proceedings of the 10th ACM Conference on Electronic Commerce*, EC '09, pages 225–234. ACM, 2009.

[17] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the 35th Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, NeurIPS '21, 2021.

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

[19] Bengt Holmström. Moral hazard and observability. *The Bell Journal of Economics*, 10(1):74–91, 1979.

[20] Zohar Karnin, Tomer Koren, and Oren Somekh. Almost optimal exploration in multi-armed bandits. In *Proceedings of the 30th International Conference on Machine Learning*, ICML '13, pages 1238–1246, 2013.

[21] Yuqing Kong and Grant Schoenebeck. An information theoretic framework for designing information elicitation mechanisms that obey complementarity. In *Proceedings of the 10th Innovations in Theoretical Computer Science Conference*, pages 40:1–40:18. Schloss Dagstuhl, 2019.

[22] Lucien Le Cam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, 1(1):38–53, 1973.

[23] Andreu Mas-Colell, Michael D. Whinston, and Jerry R. Green. *Microeconomic Theory*. Oxford University Press, 1995.

[24] Nolan Miller, Paul Resnick, and Richard Zeckhauser. Eliciting informative feedback: The peer-prediction method. *Management Science*, 51(9):1359–1373, 2005.

[25] Roger B. Myerson and Mark A. Satterthwaite. Efficient mechanisms for bilateral trading. *Journal of Economic Theory*, 29(2):265–281, 1983.

[26] Leonard J. Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971.

[27] Burr Settles. Active learning literature survey. Technical Report 1648, University of Wisconsin–Madison, 2009.

[28] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, 2009.

[29] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *Proceedings of the 11th International Conference on Learning Representations*, ICLR '23, 2023.

[30] Jens Witkowski and David C. Parkes. Peer prediction without a common prior. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, EC '12, pages 964–981. ACM, 2012.

[31] Jens Witkowski and David C. Parkes. A robust Bayesian truth serum for small populations. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, pages 1492–1498. AAAI Press, 2012.

[32] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Proceedings of the 37th Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, NeurIPS '23, 2023.

Set in Computer Modern 11/15. Typeset with LaTeX. `ifitsmanu.com`.