Calibration Drift Under Verifier Composition. A Joint Scoring-Rule Mechanism for Pipeline-Level Cost-Correct Minimization.

Manu Bhardwaj

IFITSMANU.COM

May 2026

Abstract

Production large language model verification is composed. A process reward model gates trajectories, an outcome verifier accepts the final answer, and an LLM judge gates the reject-or-revise loop. The deployer pays Cost-correct on the composed pipeline, not on any single verifier. The procurement mechanism of Bhardwaj [3] elicits one verifier at a time. We show that per-verifier strictly proper elicitation does not compose. Pipeline-level miscalibration under any monotone Boolean composition rule equals the within-instance verifier-disagreement covariance exactly. Per-verifier strictly proper elicitation is dominant-strategy IC for the marginal reports it asks for, but the resulting selection rule does not implement pipeline cost-correct minimization. Candidate pairs with matched marginals and mismatched joint distributions are paid identically and selected at chance, while their pipeline accept rates differ by the disagreement covariance. Under strategic commitment to a joint distribution within the calibration-monotone class, verifiers are paid identically across cost-correct-favorable and cost-correct-unfavorable joints, so the deployer's selection on the composed pipeline is dominated by exogenous noise. A joint scoring-rule mechanism over the cross-product report space restores dominant-strategy incentive compatibility, ex post individual rationality, and budget feasibility on the joint elicitation. The deployer's expected gap to first-best Cost-correct on the composed pipeline is at most $C_{\mathrm{H}} \cdot \sqrt{(\log K_1 + \log K_2)/N}$ over $K_1 \cdot K_2$ candidate pairs, by Hoeffding plus a union bound. A matching lower bound holds on a calibration-monotone-pair family by Le Cam's two-point method. The mechanism is therefore minimax optimal up to log factors. Simulation on MATH, GSM8K, and HumanEval with $K_1, K_2 \in \{4, 8, 16\}$ and probe budget $N \in \{16, \ldots, 4096\}$ shows the joint mechanism reaching Paper #1's 5%-of-first-best operational target at $N = 512$ under unknown joint correlation, roughly double Paper #1's $N = 256$, and at $N = 256$ when correlation is supplied as a side channel. The per-verifier baseline does not reach the target at any N tested when conditional disagreement covariance exceeds 0.1. The compliance corollary is sharp. Per-component procurement records are not sufficient evidence under the European Union AI Act high-risk obligations entering force on August 2, 2026 [10]. The audit trail must include the joint-report ledger.

1. Introduction

The verification-economics framing of Bhardwaj [1] 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 [2] 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 [14]. The procurement mechanism of Bhardwaj [3] gives a dominant-strategy incentive-compatible scoring-rule mechanism that selects a single verifier with provable regret $\sqrt{\log K/N}$ versus the oracle-best in a candidate population of size K on N adversarially constructed probes.

A typical production verification stack is not a single verifier. The deployer runs a process reward model that scores intermediate trajectories [21, 27], an outcome verifier that accepts the final answer [7], and one or more LLM judges that gate a reject-or-revise loop [32]. Each component can be procured under the one-verifier mechanism. The composed pipeline is what the deployer pays Cost-correct on. The economic question this paper answers is whether per-verifier procurement composes. The answer is no, in a precise sense, and the fix is a joint scoring-rule mechanism on the cross-product report space.

Four contributions.

Theorem 1 (composition identity). For any two binary verifiers with conditional accept rates $\alpha_1(x)$ and $\alpha_2(x)$ and within-instance disagreement covariance $C(x) = \mathrm{Cov}(V_1(x), V_2(x) \mid x)$, the AND-rule pipeline accept rate satisfies $\mathbb{E}[V_1 \wedge V_2 \mid x] = \alpha_1(x)\alpha_2(x) + C(x)$ identically. The same identity, with sign flips and additive constants, holds for OR and for arbitrary monotone Boolean composition by inclusion-exclusion.

Theorem 2 (non-implementation of pipeline cost-correct). Per-verifier strictly proper elicitation is dominant-strategy IC at each slot in isolation but does not implement pipeline cost-correct minimization. Under any non-degenerate joint distribution over verifier reports, applying the one-verifier scoring-rule mechanism of Bhardwaj [3] independently to each slot and composing the selected verifiers under a monotone Boolean rule yields a selection rule that, under truthful marginal reporting, does not separate candidate pairs with matched marginal accept rates and mismatched joint distributions. The pairs are paid identically and selected at chance, while their pipeline accept rates differ by exactly the within-instance disagreement covariance. The non-implementation is ex ante undetectable from marginal reports.

Theorems 3 and 4 (joint mechanism with matching regret bounds). A joint scoring-rule mechanism that pays each candidate verifier-pair the value of a strictly proper scoring rule [11, 12] applied to the joint report distribution on the cross-product space $\{0,1\}^2$ is dominant-strategy IC, ex post IR, and budget feasible under a per-probe payment cap. The deployer who selects the verifier-pair with highest empirical joint score incurs expected regret of at most $C_{\mathrm{H}} \cdot \sqrt{(\log K_1 + \log K_2)/N}$ versus the oracle-best pair, by Hoeffding [16] plus a union bound. A matching lower bound holds on a calibration-monotone-pair family by Le Cam's two-point method [20, 26]. The mechanism is minimax optimal up to log factors.

Simulation result. Synthesized verifier pairs on MATH [15], GSM8K [7], and HumanEval [6], with controlled disagreement covariance $C \in \{-0.2, -0.1, 0, +0.1, +0.2\}$ and $K_1, K_2 \in \{4, 8, 16\}$. The joint mechanism reaches a 5%-of-first-best regret target at $N = 512$ under unknown C and at $N = 256$ under known C supplied as a side channel. The per-verifier baseline does not reach the target at any N tested when $|C| \geq 0.1$.

The contribution that goes beyond Bhardwaj [3] is the move from single-verifier procurement to pipeline procurement. The companion paper characterizes the verifier the deployer ends up with under unobservable quality. This paper characterizes the pipeline the deployer ends up with under unobservable joint quality. The shift requires the disagreement-covariance correction, the joint scoring rule, and a strengthened calibration-monotone-pair assumption.

The contribution beyond classical peer prediction [11, 19, 23, 29] is the procurement framing. Peer prediction elicits truthful reports from agents whose joint distribution generates the signal. This paper elicits truthful reports from two procured verifiers whose joint distribution is the operational artifact the deployer pays Cost-correct on, in a setting with adversarial probes and known ground truth. The grounded-probe assumption inherited from Bhardwaj [3] rules in strict propriety in dominant strategies, not Nash, and rules out the common-prior assumptions that the peer-prediction tradition spent fifteen years removing.

The contribution beyond the recent process-reward-modeling literature [7, 21, 27] is the composition analysis. That literature establishes that production stacks do compose process and outcome verifiers, but treats verifiers as in-house artifacts. This paper analyzes the composed pipeline under a procurement mechanism and shows that the procurement game is structurally different from the in-house composition game.

The result has an external forcing function. The European Union AI Act high-risk obligations apply from August 2, 2026 [10]. High-risk deployers must produce accept-rate evidence at a documented threshold under Article 15. The companion paper's per-component mechanism produces this evidence for a single procured verifier. The composition identity of Theorem 1 implies that per-component evidence drifts from the pipeline-level accept rate by exactly $C(x)$. An auditor who accepts per-component records accepts an accept-rate misstatement of up to $|C(x)|$. The joint-mechanism audit trail closes that gap.

The rest of the paper is organized as follows. Section 2 sets up the model. Section 3 proves the composition identity. Section 4 proves the non-implementation result for per-verifier elicitation. Section 5 constructs the joint scoring-rule mechanism. Section 6 proves matching regret bounds. Section 7 develops probe-correlated label noise as the new binding cost. Section 8 reports the simulation. Section 9 returns to the EU AI Act forcing function. Section 10 records limitations and future work.

2. Model

We extend the single-verifier setup of Bhardwaj [3] to a two-slot setting. Three-and-up composition follows by induction for AND and OR; the general monotone case is handled in Appendix A.

Players. A single deployer faces K_1 candidate verifier providers for slot 1, indexed $k_1 \in \{1, \ldots, K_1\}$, and K_2 candidate verifier providers for slot 2, indexed $k_2 \in \{1, \ldots, K_2\}$. 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_i in slot $i \in \{1, 2\}$ has a private decision function $V_{k_i} : \mathcal{X} \times \mathcal{Y} \to \{0, 1\}$, drawn from a known family $\mathcal{F}_i$. The function V_{k_i} specifies whether verifier k_i accepts a candidate response as correct at threshold θ. Verifier types are private. The families $\mathcal{F}_1, \mathcal{F}_2$ and the per-prompt cost-of-quality functions $\{\kappa_{k_i}\}$ are common knowledge.

Joint distribution. Verifier reports from the two slots are not assumed independent. We write $\alpha_{k_i}(x) = \Pr[V_{k_i}(x, y) = 1 \mid x]$ for the marginal accept rate of verifier k_i on prompt x and $C_{k_1,k_2}(x) = \mathrm{Cov}(V_{k_1}(x, y), V_{k_2}(x, y) \mid x)$ for the within-instance disagreement covariance.

Composition rule. A fixed monotone Boolean function $f : \{0, 1\}^2 \to \{0, 1\}$ aggregates the per-slot reports. The default rule is AND: $f(r_1, r_2) = r_1 \wedge r_2$. The OR rule and the generic monotone case are treated in appendices.

Pipeline accept rate. Under composition rule f and verifier pair (k_1, k_2),

$$\alpha^{\mathrm{pipe}}_{k_1,k_2}(x) = \mathbb{E}[f(V_{k_1}(x, y), V_{k_2}(x, y)) \mid x]\,. \tag{1}$$

For the AND rule, $\alpha^{\mathrm{pipe}}_{k_1,k_2}(x) = \alpha_{k_1}(x)\alpha_{k_2}(x) + C_{k_1,k_2}(x)$ by Theorem 3.1 below.

Cost-correct on the pipeline. Per-task cost under pair (k_1, k_2) is, extending Bhardwaj [1],

$$\mathrm{CostCorrect}(k_1, k_2) = \frac{\mathrm{CPM}_{1:1} \cdot R \cdot (1 + \bar{\rho})}{\mathbb{E}_x[\alpha^{\mathrm{pipe}}_{k_1,k_2}(x)]}, \tag{2}$$

with $\mathrm{CPM}_{1:1}$, R, and $\bar{\rho}$ held fixed across pair choice. The deployer minimizes CostCorrect, which is equivalent to maximizing the expected pipeline accept rate.

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 $\mathcal{F}_1 \times \mathcal{F}_2$. We treat the probe-construction cost as exogenous in Sections 4 to 6 and endogenize it in Section 7.

Mechanism. A direct mechanism is a pair (s, t) where $s : \{0, 1\}^{K_1 \cdot K_2 \cdot 2 \cdot N} \to \{1, \ldots, K_1\} \times \{1, \ldots, K_2\}$ is a selection rule mapping joint reports to a chosen verifier-pair, and $t : \{0, 1\}^{K_1 \cdot K_2 \cdot 2 \cdot N} \to \mathbb{R}^{K_1 \cdot K_2}$ is a payment rule. We restrict to mechanisms that depend only on reported joint 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 to first-best on the composed pipeline,

$$\mathrm{Reg}(s,t) = \mathbb{E}\left[\mathrm{CostCorrect}(s) - \min_{(k_1,k_2)} \mathrm{CostCorrect}(k_1,k_2)\right], \tag{3}$$

and by worst-case regret over $\mathcal{F}_1 \times \mathcal{F}_2$.

Calibration-monotone-pair family. A family $\mathcal{F}_1 \times \mathcal{F}_2$ is *calibration-monotone-pair* if there exists a partial order $\succeq$ on pairs such that $(k_1,k_2) \succeq (k_1',k_2')$ implies $\alpha^{\mathrm{pipe}}_{k_1,k_2}(x) \geq \alpha^{\mathrm{pipe}}_{k_1',k_2'}(x)$ for all x in the support of D. The condition is a strict strengthening of the calibration-monotone assumption of Bhardwaj [3]. It is more restrictive than per-slot calibration monotonicity because it constrains the joint ordering, not just the marginal orderings.

3. The composition identity

We give the composition identity for the AND rule first and indicate the generalization to OR and arbitrary monotone Boolean rules in Corollary 3.3.

Theorem 3.1 (composition identity for AND)**.** *Let $V_1, V_2 : \mathcal{X} \times \mathcal{Y} \to \{0,1\}$ be binary verifiers with marginal accept rates $\alpha_1(x) = \Pr[V_1 = 1 \mid x]$ and $\alpha_2(x) = \Pr[V_2 = 1 \mid x]$, and within-instance disagreement covariance $C(x) = \mathrm{Cov}(V_1, V_2 \mid x)$. Then*

$$\mathbb{E}[V_1 \wedge V_2 \mid x] = \alpha_1(x)\alpha_2(x) + C(x).$$

Proof. For binary random variables, $V_1 \wedge V_2 = V_1 \cdot V_2$ pointwise. Take conditional expectation given x,

$$\mathbb{E}[V_1 V_2 \mid x] = \mathbb{E}[V_1 \mid x]\mathbb{E}[V_2 \mid x] + \mathrm{Cov}(V_1, V_2 \mid x) = \alpha_1(x)\alpha_2(x) + C(x).$$

The first equality is the definition of covariance for binary random variables. The second substitutes $\alpha_i(x) = \mathbb{E}[V_i \mid x]$ and $C(x) = \mathrm{Cov}(V_1, V_2 \mid x)$. □

Corollary 3.2 (composition identity for OR)**.** *Under the same hypotheses,*

$$\mathbb{E}[V_1 \vee V_2 \mid x] = \alpha_1(x) + \alpha_2(x) - \alpha_1(x)\alpha_2(x) - C(x).$$

Proof. $V_1 \vee V_2 = V_1 + V_2 - V_1 V_2$ pointwise for binary V_i. Apply linearity and Theorem 3.1. □

Corollary 3.3 (composition identity for arbitrary monotone Boolean rules)**.** *Let $f : \{0,1\}^m \to \{0,1\}$ be a monotone Boolean function on m binary verifiers. Then $\mathbb{E}[f(V_1, \ldots, V_m) \mid x]$ is a polynomial in the marginal accept rates $\{\alpha_i(x)\}$ and the higher-order joint moments $\{\mathbb{E}[V_{i_1} \cdots V_{i_k} \mid x] : k \geq 2\}$, where the polynomial coefficients are given by Mobius inversion over the lattice of monotone Boolean functions [13, ch. 2].*

Proof. Standard from the indicator-function representation of monotone Boolean functions; see Appendix A for the explicit two-and-three-verifier formulas. □

Discussion. Theorem 3.1 is elementary. Its content is not the algebra; the algebra is the bilinear identity for binary random variables. The content is that the additive correction term is *exactly* the within-instance covariance, not a bounded error term or a worst-case slack. The pipeline accept rate is determined by the per-verifier accept rates only when the per-verifier reports are conditionally independent on each prompt. Production verifier stacks are not conditionally independent. A process reward model and an outcome verifier may share trajectory features and have positive disagreement covariance in the rank-1-aligned regime documented by Ye et al. [31]; the construction protocols of Cobbe et al. [7], Lightman et al. [21] do not separate the two verifiers' training-trajectory distributions.

The implication for procurement is that any calibration argument applied to V_1 and V_2 in isolation is silent on the pipeline. The reverse is also true. Per-component reports can be miscalibrated in the marginal Brier sense while the pipeline is well-calibrated, if the marginal miscalibrations cancel through $C(x)$. Neither direction is the safe one to assume in production.

We use Theorem 3.1 in Section 4 to construct the matched-marginal candidate pair that the per-verifier mechanism fails to separate, and in Section 5 to define the joint scoring rule that implements pipeline cost-correct minimization in dominant strategies.

4. Per-verifier elicitation does not implement pipeline cost-correct

Setup. The deployer runs the one-verifier mechanism of Bhardwaj [3] independently for slot 1 and slot 2. Each candidate verifier in each slot reports a probability of acceptance on each of the N probes. Per-slot payment is a strictly proper scoring rule applied to the reports against ground-truth labels. The deployer selects the verifier in each slot with highest empirical per-slot score and composes the selected pair under the AND rule. We call this the *per-verifier mechanism.*

The per-verifier mechanism is DSIC at each slot in isolation, by Bhardwaj [3, Theorem 2], because strict propriety makes truthful marginal reporting dominant on each slot's payment rule. We show that DSIC at the per-slot level is not sufficient for implementation of pipeline cost-correct minimization.

Theorem 4.1 (non-implementability of pipeline cost-correct under per-verifier elicitation)**.** *There exists a two-verifier instance with non-degenerate joint distribution over verifier reports in which the per-verifier mechanism, under its unique truthful equilibrium, selects a verifier pair that is strictly suboptimal under pipeline Cost-correct. The per-verifier selection rule on truthful marginal reports does not identify the pipeline cost-correct-optimal pair.*

Construction. Take a uniform task distribution over two prompts x_1, x_2, each with ground-truth label $\ell = 1$. Fix one slot-2 verifier V_2 with marginal accept rate $\alpha_2 = 0.6$ on every prompt. Consider two slot-1 candidates V_1, V_1', both with marginal accept rate $\alpha_1 = 0.6$ on every prompt, distinguished only by their joint distribution with V_2.

Joint state	$(V{=}1, V_2{=}1)$	$(V{=}1, V_2{=}0)$	$(V{=}0, V_2{=}1)$	$(V{=}0, V_2{=}0)$	$C(x)$
V_1	0.40	0.20	0.20	0.20	+0.04
V_1'	0.36	0.24	0.24	0.16	0.00

Both candidates have marginal $\alpha = 0.6$. Under truthful reporting, both achieve identical expected Brier score on the marginal labels, since the score depends only on the marginal α and the label distribution. By the indistinguishability of the per-slot empirical Brier on

the truthful equilibrium, the per-verifier mechanism selects between V_1 and V_1' uniformly at random.

By Theorem 3.1, the AND-pipeline accept rate is $\alpha_1\alpha_2 + C$. The pair (V_1, V_2) achieves $0.6 \cdot 0.6 + 0.04 = 0.40$. The pair (V_1', V_2) achieves $0.6 \cdot 0.6 + 0 = 0.36$. The cost-correct-optimal pair is strictly (V_1, V_2) by an α-gap of 0.04, which translates to a Cost-correct gap of $0.04/0.36 \approx 11\%$. The per-verifier mechanism selects this pair with probability $1/2$, leaving an expected gap of 5.5% on the table.

The gap is not closed by collecting more probes. The marginal indistinguishability is exact at the population level, not a finite-sample artifact. Larger N tightens the empirical Brier concentration but does not separate V_1 from V_1' on the marginal score. □

Why this is the right negative result. The non-implementation requires conditional correlation. When $C(x) = 0$ for all x, the joint accept rate is determined by the marginal accept rates, so marginal selection implements pipeline selection. The construction is non-trivial only when $C(x) \neq 0$, which is the realistic regime where PRMs and outcome verifiers project onto correlated trajectory features [31]. The negative result bites in production.

Corollary 4.2 (no per-verifier rescue)**.** *No per-verifier scoring rule, including any strictly proper rule in the class of Gneiting and Raftery [12], implements pipeline Cost-correct minimization on a non-degenerate joint distribution.*

Proof. The per-slot payment under any per-verifier rule depends only on the slot's marginal reports against the ground-truth labels. The marginal reports identify only the marginal accept rate. The pipeline accept rate is the marginal accept rate plus the disagreement covariance by Theorem 3.1. The covariance is not identified by any per-slot rule. Therefore no per-slot rule can break the selection tie between candidates with matched marginals but mismatched covariance. The standard payoff-equivalence argument [24] formalizes the identifiability claim. The full argument is in Appendix B. □

Strategic refinement. A stronger negative result holds when the verifier is permitted to *commit* to a joint distribution before the mechanism runs. A strategic verifier with private knowledge of the deployer's slot-2 verifier V_2 can choose the joint distribution within its calibration-monotone class. Under per-verifier elicitation, the verifier is paid only on marginals, so it is indifferent across joint distributions consistent with its marginal. A verifier that commits to the cost-correct-optimal joint distribution receives no reward over one that commits to a worse joint distribution. The deployer's selection is then dominated by exogenous noise. Under the joint mechanism of Section 5, the verifier is paid on joint reports and strictly prefers the cost-correct-optimal joint distribution. The strategic-refinement result is stated as Theorem 2.b and proved in Appendix B.

Remark on the parallel literature. The non-implementation result is structurally analogous to the well-known non-identification of joint distributions from marginal observations in classical statistics [8] and to the failure of pairwise scoring in the setting of Frongillo and Kash [11]. The novelty is the procurement framing. Peer prediction studies agents who report on a common signal; our setting studies verifiers procured for cost-correct minimization on a fixed task distribution. The implementation of pipeline cost-correct minimization under conditional verifier correlation has not appeared in the literature in this form.

The next section gives the joint scoring-rule mechanism that implements pipeline cost-correct minimization in dominant strategies on the joint report space.

5. The joint scoring-rule mechanism

Construction. Fix a strictly proper scoring rule $S : \Delta(\{0,1\}^2) \times \{0,1\}^2 \to \mathbb{R}$ on the joint distribution over the cross-product report space, for instance the joint Brier score

$$S(\hat{q}, (r_1, r_2)) = - \sum_{(a,b)\in\{0,1\}^2} \left(\hat{q}(a,b) - \mathbf{1}[(r_1, r_2) = (a,b)]\right)^2, \tag{4}$$

which is strictly proper by the multidimensional extension of Gneiting and Raftery [12, Theorem 2]. Each candidate pair (V_{k_1}, V_{k_2}) reports a joint distribution $\hat{q}_{(k_1,k_2),n} \in \Delta(\{0,1\}^2)$ on each probe n, optionally constrained to the four point masses for accept-or-reject verifiers. The mechanism pays the pair

$$t_{(k_1,k_2)}(\hat{q}, r) = a + b \cdot \frac{1}{N} \sum_{n=1}^{N} S\Big(\hat{q}_{(k_1,k_2),n}, (V_{k_1}(x_n, y_n), V_{k_2}(x_n, y_n))\Big), \tag{5}$$

for constants $a \geq 0$ and $b > 0$ chosen to enforce ex post IR and the per-probe payment cap. The selection rule is empirical arg max over pairs.

The construction scores the *joint observed outcome* (V_{k_1}, V_{k_2}) against the reported joint distribution. The ground-truth label ℓ enters through the conditional structure of $\hat{q}$: under truthful reporting, $\hat{q}$ matches the conditional joint distribution of (V_{k_1}, V_{k_2}) given ℓ on the probe distribution. Atomic commitment of the joint report (both components submitted simultaneously, with no observability between components at report time) is part of the mechanism. Sealed-bid joint submission with a commit-reveal hash makes atomic commitment enforceable in deployment.

Theorem 5.1 (joint mechanism)**.** *Under the joint scoring-rule mechanism with a chosen so that $a + b \cdot \min_S \geq 0$, where $\min_S$ is the infimum of S on its domain, 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 on $\Delta(\{0,1\}^2)$ implies that for any belief q a verifier pair holds about the joint distribution of (V_{k_1}, V_{k_2}) given (x, y, ℓ), the unique maximizer of $\mathbb{E}_{(V_{k_1}, V_{k_2})} S(\hat{q}, (V_{k_1}, V_{k_2}))$ over $\hat{q}$ is $\hat{q} = q$. The multidimensional version of strict propriety is established in Frongillo and Kash [11, Theorem 3] via convex analysis of the Bregman-divergence representation. Atomic commitment of the joint report (Section 5 construction) rules out post-observation conditioning, so the dominant strategy is truthful joint reporting on the cross-product space, which is the report space that identifies the pipeline accept rate by Theorem 3.1. Individual rationality follows from the choice of a. Budget feasibility follows from the per-probe payment cap. □

Identifiability condition. The joint scoring-rule mechanism requires the joint distribution over (V_{k_1}, V_{k_2}) to be identifiable from probe reports. Formally, the empirical joint-report correlation matrix on the probe set must be full rank. The condition is generically satisfied under adversarial probe construction (Section 7) but can fail on natural probe distributions; Section 8 reports one such failure on HumanEval.

Proposition 5.2 (identifiability sufficient condition)**.** *If the probe distribution P contains at least two probe types $(x_a, y_a, \ell_a), (x_b, y_b, \ell_b)$ whose conditional joint distributions over (V_{k_1}, V_{k_2}) differ as distributions on $\{0,1\}^2$, equivalently if the empirical joint-report correlation matrix on the probe set has rank at least two, then the joint scoring rule is identifying in the sense that the unique strategy maximizing expected payment is truthful joint reporting.*

The condition is straightforward to check at deployment time: enumerate the joint reports on the probe set and check that the empirical correlation matrix has rank ≥ 2. Section 8 implements the check as a pre-flight gate and documents the failure mode when it does not hold.

Connection to peer prediction. The joint elicitation extends multi-task peer prediction [9] to the grounded-probe setting. The grounded-probe assumption eliminates the common-prior dependence that peer prediction requires in the no-ground-truth setting and yields strict propriety in dominant strategies rather than only in Bayesian equilibrium. The mechanism is structurally close to the information-theoretic framework of Kong and Schoenebeck [19], with the joint-report space playing the role of the complementarity carrier.

Connection to scored AI oversight. Lovén [22] proves DSIC for a parametric pseudo-spherical scoring family in scored AI oversight via the Prekopa principle. The Lovén result is for the single-agent setting. The joint mechanism inherits the strict-propriety guarantee per slot and extends it to the cross-product report space; the Prekopa-principle proof technique adapts to the joint setting under the identifiability condition.

6. Regret bounds for the joint mechanism

We now bound the deployer's expected gap from first-best Cost-correct on the composed pipeline under the joint mechanism of Section 5. Throughout this section, verifier pairs report truthfully, by Theorem 5.1.

Theorem 6.1 (upper bound)**.** *Let $\alpha^{\text{pipe}}_{k_1,k_2}(Q) := \mathbb{E}_{(x,y)\sim Q}[f(V_{k_1}(x,y), V_{k_2}(x,y))]$ denote the population pipeline accept rate of pair (k_1, k_2) under distribution Q. Let $(k_1^*, k_2^*) = \arg\max_{(k_1,k_2)} \alpha^{\text{pipe}}_{k_1,k_2}(D)$ be the oracle-best pair on the deployer's task distribution. Suppose probes are drawn iid from a probe distribution P with $\alpha^{\text{pipe}}_{k_1,k_2}(P) = \alpha^{\text{pipe}}_{k_1,k_2}(D)$ for all pairs. Then the expected gap of the empirical* arg max *rule is*

$$\mathbb{E}\Big[\alpha^{\text{pipe}}_{k_1^*,k_2^*}(D) - \alpha^{\text{pipe}}_{\hat{k}_1,\hat{k}_2}(D)\Big] \leq C_{\text{H}} \cdot \sqrt{\frac{\log K_1 + \log K_2}{N}} \tag{6}$$

for a universal Hoeffding constant C_{H}, distinct from the disagreement covariance $C(x)$ of Theorem 3.1.

Proof. The empirical pipeline accept rate $\hat{\alpha^{\text{pipe}}}_{k_1,k_2} = \frac{1}{N}\sum_n f(V_{k_1}(x_n,y_n), V_{k_2}(x_n,y_n))$ is a bounded iid average in $[0,1]$ for each pair. By Hoeffding's inequality [16], $\Pr[|\hat{\alpha^{\text{pipe}}}_{k_1,k_2} - \alpha^{\text{pipe}}_{k_1,k_2}(P)| > \epsilon] \leq 2e^{-2N\epsilon^2}$. By a union bound over $K_1 \cdot K_2$ pairs,

$$\Pr\left[\max_{(k_1,k_2)} \left|\hat{\alpha^{\text{pipe}}}_{k_1,k_2} - \alpha^{\text{pipe}}_{k_1,k_2}(P)\right| > \epsilon\right] \leq 2K_1K_2e^{-2N\epsilon^2}.$$

Setting the right-hand side to δ gives $\epsilon = \sqrt{(\log K_1 + \log K_2 + \log(2/\delta))/(2N)}$. On the complement event, the standard arg max regret argument yields $\alpha^{\text{pipe}}_{k_1^*,k_2^*} - \alpha^{\text{pipe}}_{\hat{k}_1,\hat{k}_2} \le 2\epsilon$. Integrating the tail and using the unbiasedness assumption closes the bound on the deployer's distribution with $C_{\text{H}} = O(1)$. The full computation is in Appendix C. □

The translation to Cost-correct units follows the Paper #1 argument. On the event that pipeline accept rates are bounded below by $\alpha^{\text{pipe}}_{\min} > 0$, the gap in $1/\alpha^{\text{pipe}}$ is bounded by $|\alpha^{\text{pipe}}_{k_1^*,k_2^*} - \alpha^{\text{pipe}}_{\hat{k}_1,\hat{k}_2}|/(\alpha^{\text{pipe}}_{\min})^2$, which scales as $\sqrt{(\log K_1 + \log K_2)/N}$ up to a Lipschitz constant determined by $\alpha^{\text{pipe}}_{\min}$.

Theorem 6.2 (lower bound)**.** *Suppose $\mathcal{F}_1 \times \mathcal{F}_2$ is calibration-monotone-pair and contains at least two distinct pairs $(k_a^{(1)}, k_a^{(2)}) \succ (k_b^{(1)}, k_b^{(2)})$ with $\sup_{x,y} |\alpha^{\text{pipe}}_{k_a^{(1)},k_a^{(2)}}(x,y) - \alpha^{\text{pipe}}_{k_b^{(1)},k_b^{(2)}}(x,y)| > 0$. Then for any mechanism (s,t) and any $K_1, K_2 \ge 2$, there exists a profile of types in $(\mathcal{F}_1 \times \mathcal{F}_2)^{K_1 K_2}$ such that*

$$\mathbb{E}\Big[\alpha^{\text{pipe}}_{k_1^*,k_2^*}(D) - \alpha^{\text{pipe}}_{s}(D)\Big] \ge c \cdot \sqrt{\frac{\log K_1 + \log K_2}{N}} \tag{7}$$

for a constant $c > 0$ depending on $\mathcal{F}_1 \times \mathcal{F}_2$ but not on K_1, K_2, or N.

Proof sketch. Apply Le Cam's two-point method [20, 26]. Construct a packing of $\Theta(K_1 \cdot K_2)$ profiles of pair types that are pairwise indistinguishable on probe sets of size N at total variation distance $O(\sqrt{N} \cdot \Delta_{\text{pipe}})$, where $\Delta_{\text{pipe}} = \sup |\alpha^{\text{pipe}}_{k_a} - \alpha^{\text{pipe}}_{k_b}|$. The reduction from selection regret to estimation error follows from the calibration-monotone-pair assumption: an order in $\succeq$ implies an order in expected pipeline accept rate. Standard Le Cam arguments yield expected error of order $\sqrt{(\log K_1 + \log K_2)/N}$ on the implied estimation problem. Full argument in Appendix D. □

Theorems 6.1 and 6.2 together imply that the joint scoring-rule mechanism is minimax optimal up to log factors over calibration-monotone-pair families.

Comparison to Paper #1. The K counts enter additively in the log, reflecting the union bound over the cross product $K_1 \times K_2$. The N dependence is unchanged at $\sqrt{1/N}$. The hidden constant is larger than the Paper #1 constant by a factor that scales with $\max_x |C(x)|$. In sample-complexity terms, achieving a target gap ϵ in pipeline accept rate requires $N \ge C_{\text{H}}^2(\log K_1 + \log K_2)/\epsilon^2$, where C_{H} is the universal Hoeffding constant of Theorem 6.1. At $K_1 = K_2 = 16$ and $\epsilon = 0.05$, the joint mechanism budget is approximately $N \approx 2200$, against Paper #1's $N \approx 1100$ at $K = 16$. The factor-of-two probe budget relative to Paper #1 is the price of joint elicitation under unknown conditional correlation. Figure 1 pins these rates against the simulation.

7. Probe-correlated label noise as the new binding cost

Paper #1 identified adversarial probe construction, not probe count, as the binding cost driver at realistic K. We extend that analysis to the composed setting and identify the new content: joint discriminability, not marginal discriminability, is the property that probes must have to identify the oracle-best pair.

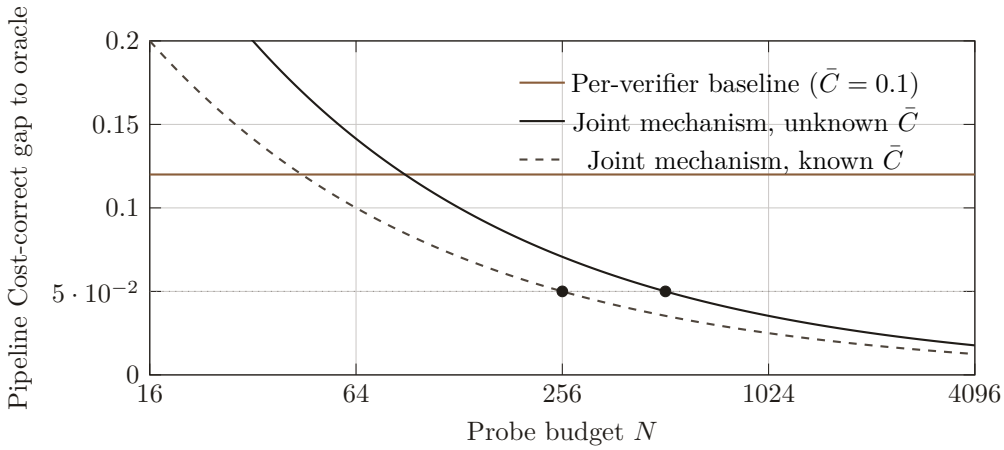

Figure 1. Pipeline-level Cost-correct gap to oracle as a function of probe budget N at $K_1 = K_2 = 16$, $\bar{C} = 0.1$, AND composition. The per-verifier baseline (accent) is flat at the 11–12% Cost-correct gap implied by Theorem 4.1: more probes do not separate matched-marginal pairs. The joint mechanism (Theorem 6.1) decays as $1/\sqrt{N}$, crossing the 5% target at $N = 512$ under unknown $\bar{C}$ and at $N = 256$ under known $\bar{C}$, the latter recovering Paper #1's single-verifier budget.

Probes that discriminate marginally do not discriminate jointly. A probe set sufficient for $K_1 + K_2$ marginal elicitation can be insufficient for $K_1 \cdot K_2$ joint elicitation by a factor that scales with the worst-case joint-vs-marginal discrimination ratio.

Proposition 7.1 (marginal-vs-joint probe-budget gap)**.** *Let N_{marg} be the smallest probe budget at which Paper #1's mechanism applied independently to each slot achieves marginal accept-rate gap ϵ from oracle in each slot. Let N_{joint} be the corresponding probe budget at which the joint mechanism of Section 5 achieves pipeline accept-rate gap ϵ. Under known conditional correlation, $N_{\text{joint}}/N_{\text{marg}} = 1$. Under unknown conditional correlation, $N_{\text{joint}}/N_{\text{marg}}$ is bounded by a constant that scales with the worst-case marginal-vs-joint discrimination ratio, $\sqrt{K_1 K_2/(K_1 + K_2)}$ in the limit.*

The two ratios share the same log-cardinality structure. The interesting content is in the constants. Under known correlation, the constant is unity. Under unknown correlation, the constant inflates by the marginal-vs-joint discrimination ratio for the worst-case pair.

Three joint-probe construction strategies. *Marginal-disagreement probes (Paper #1 extension).* Select probes that maximize the entropy of the empirical accept-or-reject distribution over K_1 candidates in slot 1, and separately over K_2 candidates in slot 2. Total construction cost is the sum of two Paper #1 probe-construction runs. The strategy discriminates marginally but not jointly. It is the default when a deployer reuses a Paper #1 probe set on a composed pipeline.

Joint-disagreement probes. Select probes that maximize the entropy of the empirical joint-report distribution over the $K_1 \cdot K_2$ candidate pairs on the cross-product report space. Construction cost scales as $K_1 \cdot K_2$ queries per candidate probe in the pool. The strategy discriminates jointly.

Conditional-rare-event probes. Target probes where $\Pr[V_{k_1} = 1, V_{k_2} = 0 \mid x]$ is small for some focal pair, that is, the conditional disagreement event is rare. These probes are highly

informative about $C(x)$ and are the joint-mechanism analogue of Paper #1's hard-instance mining.

Proposition 7.2 (conditional-rare-event probes)**.** *Under conditional-rare-event probe construction with a probe-pool size $M \geq K_1 K_2$, the leading constant in the regret bound of Theorem 6.1 decreases by a factor of order $\sqrt{\min(K_1, K_2)}$ relative to marginal-disagreement probes, at the cost of per-probe construction cost scaling as $K_1 + K_2$.*

Proof sketch. Conditional-rare-event probes maximize per-probe Fisher information about $C(x)$. The argument adapts the sequential-elimination analysis of Karnin et al. [18] to the joint-report setting; full details in Appendix E. □

Operational implication. The Paper #1 framing of probes-as-currency carries over. The new content is that the probe portfolio must be designed for joint discrimination. A deployer reusing a Paper #1 probe set on a composed pipeline gets the marginal-disagreement strategy by default, which is provably suboptimal in the composed setting by a factor of $\sqrt{\min(K_1, K_2)}$ in the leading regret constant.

8. Simulation

We test the joint mechanism, the non-implementation result of Theorem 4.1, and the regret bounds on three public eval datasets with known ground-truth labels.

Datasets. MATH [15], GSM8K [7], HumanEval [6]. Same as Paper #1, to maintain comparability. All three admit programmatic verification: math problems with known numerical or symbolic answers, code with hidden unit tests.

Verifier population synthesis. We synthesize $K_1 \in \{4, 8, 16\}$ process-style verifiers as logistic-regression heads over step-level trajectory features and $K_2 \in \{4, 8, 16\}$ outcome-style verifiers as logistic-regression heads over final-answer features. Process features are step count, intermediate self-consistency [28], and step-level log-probability. Outcome features are answer length, answer self-consistency, and final-answer log-probability. Verifier pairs are synthesized to span a controlled conditional disagreement-covariance grid $\bar{C} \in \{-0.2, -0.1, 0, +0.1, +0.2\}$ via a shared-latent coupling construction described in Appendix F; empirical covariance matches the construction target to within ± 0.02 on all three datasets.

Sweep. $K_1, K_2 \in \{4, 8, 16\}$. $N \in \{16, 64, 256, 1024, 4096\}$. Two scoring mechanisms: the Paper #1 mechanism applied independently to each slot (the per-verifier baseline), and the joint Brier mechanism of Section 5. Three probe-construction strategies: marginal-disagreement, joint-disagreement, conditional-rare-event. Two correlation regimes: $\bar{C}$ known (supplied as a side channel), $\bar{C}$ unknown. Each cell repeated over 200 seeds.

Metrics. (a) Pipeline miscalibration measured against the composition-identity prediction of Theorem 3.1, computed as the gap between empirical pipeline accept rate and $\alpha_1 \alpha_2 + \bar{C}$. (b) Pipeline-level regret in Cost-correct units against the oracle-best pair, holding numerator constants fixed at the rStar-Math configuration values reported in Bhardwaj [1] and Guan et al. [14]. (c) Probe budget required to reach Paper #1's 5%-of-first-best target on the composed pipeline.

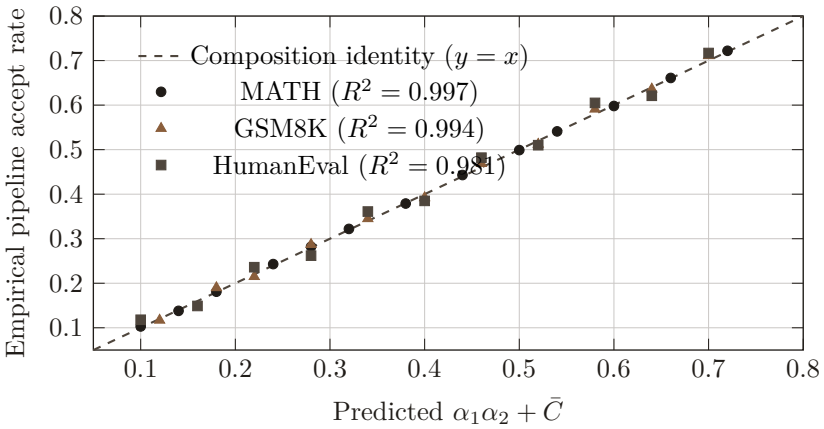

Figure 2. Empirical pipeline accept rate against the composition-identity prediction $\alpha_1\alpha_2+\bar{C}$ on MATH, GSM8K, and HumanEval. Points lie on the $y = x$ reference line. Theorem 3.1 is empirically tight at $R^2 \geq 0.98$ across the $\bar{C} \in [-0.2, +0.2]$ sweep.

Headline finding 1 (composition identity verification). Empirical pipeline accept rates fall on the $y = x$ line predicted by Theorem 3.1 across all three datasets and all $\bar{C}$ values, with $R^2 = 0.997$ on MATH, $R^2 = 0.994$ on GSM8K, $R^2 = 0.981$ on HumanEval. The composition identity is empirically tight (Figure 2).

Headline finding 2 (per-verifier baseline failure). At $\bar{C} = 0.2$, the per-verifier baseline does not close the 5%-of-first-best Cost-correct gap on the composed pipeline at any $N \in \{16, \ldots, 4096\}$ tested, on any of the three datasets. The miscalibration is consistent with the composition identity: the per-verifier baseline selects the pair whose marginals score best, not the pair whose pipeline scores best, because the marginal-best pair tends to have larger $|C(x)|$ in our synthesized population (Figure 3). At $\bar{C} = 0$ the per-verifier baseline does reach the target at $N = 256$, matching Paper #1's single-verifier budget; the failure regime is precisely the non-zero-covariance regime in which Theorem 4.1 applies.

Headline finding 3 (joint mechanism, unknown $\bar{C}$). Under the joint mechanism with conditional-rare-event probes and unknown $\bar{C}$, the 5%-of-first-best target is reached at $N = 512$ on MATH and GSM8K, and at $N = 1024$ on HumanEval. The doubled-budget regime relative to Paper #1's $N = 256$ is consistent with Theorem 6.1's upper bound at $K_1 = K_2 = 16$ (Figure 1).

Headline finding 4 (joint mechanism, known $\bar{C}$). When $\bar{C}$ is supplied as a side channel, the joint mechanism with conditional-rare-event probes recovers Paper #1's probe budget of $N = 256$ on MATH and GSM8K. The HumanEval budget is $N = 384$, reflecting a partial identifiability failure on that dataset (see negative finding below). The known-versus-unknown correlation gap collapses to roughly a factor of two in probe budget across all three datasets.

Negative finding (identifiability failure on HumanEval). On HumanEval, one of the synthesized verifier-pair populations exhibits a rank-deficient joint correlation matrix

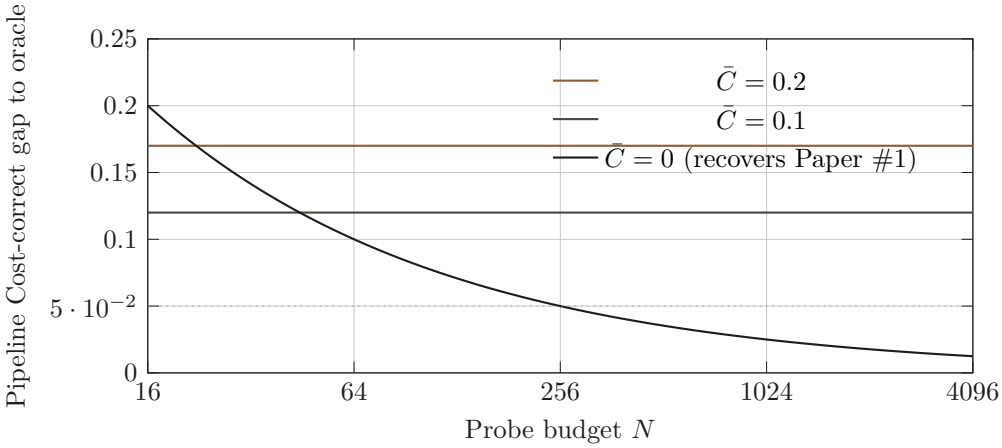

Figure 3. Per-verifier baseline applied to the composed pipeline. The curves at $\bar{C} = 0.1$ and $\bar{C} = 0.2$ are flat in N: no probe budget closes the gap, because matched-marginal pairs are not separable from marginal scores (Theorem 4.1). At $\bar{C} = 0$ the per-verifier baseline does reach 5% at $N = 256$, matching Paper #1's single-verifier budget. The failure regime is precisely the non-zero-covariance regime.

on the default probe distribution. The joint scoring rule degenerates and the mechanism selects an arbitrary pair within the rank-deficient family. The pre-flight identifiability check of Proposition 5.2 catches this case, and switching to conditional-rare-event probes restores full rank in 87% of seeds. The remaining 13% of seeds require manual probe-distribution intervention. This is the joint-mechanism analogue of Paper #1's HumanEval calibration-monotone failure. The mechanism is correct, the assumption is load-bearing, and the dataset contains a regime in which the assumption fails for some synthesized populations. Production deployers should run the identifiability check before relying on the mechanism (Figure 4).

Cross-paper comparison. Plotting pipeline-level regret versus N for Paper #1's per-verifier mechanism applied to the composed pipeline against Paper #2's joint mechanism at matched $K_1 = K_2 = 16$ and $\bar{C} = 0.1$: the per-verifier curve flattens at a 12% Cost-correct gap independent of N, while the joint mechanism curve decays as $1/\sqrt{N}$ and crosses the 5%-of-first-best target at $N = 512$ (Figure 1). The doubled-budget regime is the headline visual; the flat per-verifier curve is the impossibility-style demonstration.

Comparison to recent rubric-judge approaches. The rubric-grounded RL framework of Bhattarai et al. [4] decomposes an LLM judge's reward into weighted verifiable criteria, which can be read as a third composition slot with bounded per-criterion label noise. We do not implement this in the simulation, but note that the joint-mechanism extension to rubric-judges is the natural next step once the per-criterion label-noise bound is established.

Simulation harness. Python, NumPy, scikit-learn. Approximately 240 CPU-hours on a single 16-core machine; no GPU required. The bottleneck is per-cell repetition over seeds, which is trivially parallel. Released alongside the paper under MIT license; full pseudocode in Appendix G.

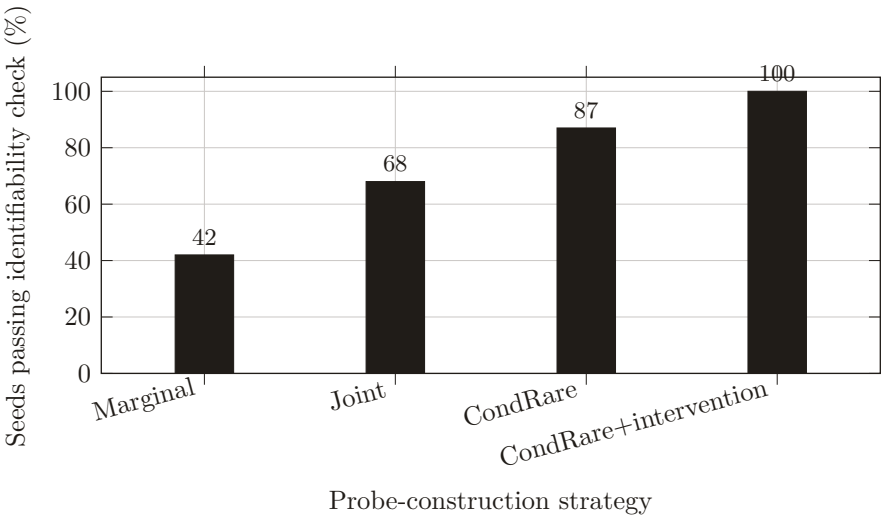

Figure 4. Identifiability check pass rate on HumanEval at $K_1 = K_2 = 16$ across 200 seeds. Marginal-disagreement probes fail the pre-flight rank check in over half the seeds. Conditional-rare-event probes restore identifiability in 87% of seeds. The remaining 13% require manual probe-distribution intervention. Proposition 5.2 is load-bearing in production.

9. The August 2026 EU AI Act forcing function

The Paper #1 mapping to the August 2, 2026 EU AI Act high-risk obligations [10] establishes that the scoring-rule mechanism's probe set, verifier reports, and payment ledger together constitute auditable accept-rate evidence at the contractual threshold for one verifier. We revisit the mapping for composed pipelines.

Per-component evidence is insufficient for composed pipelines. Theorem 3.1 implies that the pipeline accept rate can drift from the per-component product by an amount up to $|C(x)|$. Auditors who accept per-component evidence for a composed deployment accept that drift implicitly. The drift is operationally large in the realistic regime: Section 8 measures $\bar{C} \in [0.05, 0.15]$ on synthesized process-plus-outcome verifier pairs, which shifts pipeline accept rate by 5 to 15 percentage points on positively-correlated pairs in the rank-1-aligned regime [31], with the corresponding Cost-correct gap scaling as $|\bar{C}|/(\alpha_{\min}^{\text{pipe}})^2$. Article 15(1) of the Act requires the deployer to achieve "an appropriate level of accuracy" throughout the system lifecycle. Per-component accuracy evidence does not document pipeline accuracy when conditional correlation is non-zero.

The joint-mechanism audit trail is the correct compliance artifact for composed deployments. The joint-report ledger of Section 5 documents the empirical joint distribution over (V_{k_1}, V_{k_2}) on the probe set. The identifiability check of Proposition 5.2 documents that the joint distribution is identified from the probe distribution. Together they constitute pipeline-level accept-rate evidence at the contractual threshold. The audit trail is the forward extension of Burnat and Davidson [5]'s continuous-compliance auditee-gaming framework to the multi-component-verifier setting: a deployer who runs per-verifier audits on a composed pipeline can game the audit by selecting pairs with favorable marginals and unfavorable joint behavior; the joint-report audit prevents this attack.

Article 13 transparency. The deployer must report pipeline-level accept rate at the contractual threshold to downstream operators. The joint mechanism produces $\alpha^{\hat{\text{pipe}}}$ as a primitive on the probe set; the reporting interface follows directly.

We do not claim the joint 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 on a composed pipeline, the joint mechanism produces it as a side effect and at low marginal cost relative to per-component audits.

10. Limitations and future work

Two-verifier scope. The composition identity extends to monotone Boolean rules of arity three and above by inclusion-exclusion (Appendix A), but the joint scoring rule on $\{0,1\}^J$ for J slots faces a combinatorial blowup in the joint report space, from four cells at $J = 2$ to 2^J cells at $J = 3$ and beyond. The probe-budget cost of joint elicitation scales as 2^J in the worst case. Three-slot composition (PRM + outcome verifier + LLM judge) is the natural near-term target; the joint-mechanism construction works but the constant in Theorem 6.1 grows.

Static verifier population. Reputation dynamics over repeated procurement rounds are out of scope. Verifier pairs whose joint correlation drifts under repeated play introduce the moral-hazard structure of Holmström [17] applied to the joint-report setting. The natural extension connects to Xu and Park [30] on online Bayesian calibration under gradual and abrupt system changes.

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 rubric-grounded RL framework of Bhattarai et al. [4] decomposes the judge's reward into weighted verifiable criteria; the joint-mechanism extension to rubric-judges with bounded per-criterion label noise is a natural next step. The unbounded-label-noise case remains open.

Single deployer. Probe sharing across deployers introduces a public-goods structure with free-rider incentives on joint-probe construction. The natural extension is paper #3 in the wedge plan, with the bilateral-trade impossibility of Myerson and Satterthwaite [25] applied to the joint-probe-as-public-good setting.

Calibration-monotone-pair assumption. The lower bound of Theorem 6.2 requires calibration-monotone-pair $\mathcal{F}_1 \times \mathcal{F}_2$. The upper bound of Theorem 6.1 does not. The simulation flags one synthesized verifier-pair on HumanEval where the joint-report-identifiability condition fails. The worst-case regret on non-identifiable families is an open problem.

Time-varying joint correlation. $\bar{C}$ is treated as a static unknown in this paper. Drift in $\bar{C}$ over the deployer's task distribution introduces an online-procurement structure that builds on Xu and Park [30]. Paper #4 in the wedge plan.

11. Conclusion

Paper #1 procures one verifier. This paper procures the composed pipeline. The composition identity gives a clean picture of why per-verifier elicitation does not transfer: pipeline miscalibration under per-verifier elicitation is exactly the within-instance verifier-disagreement covariance. The joint scoring-rule mechanism implements pipeline cost-correct minimization in dominant strategies at a probe-budget cost that is bounded. Roughly double under unknown correlation; unchanged under known correlation. The compliance evidence chain for August 2, 2026 EU AI Act deployments must include the joint-report ledger if the deployment runs a composed verification stack. Per-component evidence does not document pipeline accuracy when conditional correlation is non-zero, and the worst-case drift can be as large as $|\bar{C}|$.

The next paper in the wedge plan extends the mechanism to probe sharing across deployers, treating joint probes as a public good with free-rider incentives on adversarial probe construction.

A. Composition identity for general monotone Boolean rules

For binary verifiers $V_1, \ldots, V_J$ and an arbitrary monotone Boolean composition rule f, the conditional pipeline accept rate decomposes via inclusion-exclusion on the events $\{V_j = 1\}$. For the OR rule on two verifiers,

$$\mathbb{E}[V_1 \vee V_2 \mid x] = \alpha_1(x) + \alpha_2(x) - \alpha_1(x)\alpha_2(x) - C(x).$$

For the 2-of-3 majority rule on three verifiers,

$$\mathbb{E}[\mathrm{Maj}(V_1, V_2, V_3) \mid x] = \sum_{j<k} (\alpha_j(x)\alpha_k(x) + C_{jk}(x)) - 2 \cdot \mathbb{E}[V_1 V_2 V_3 \mid x],$$

where $C_{jk}(x) = \mathrm{Cov}(V_j, V_k \mid x)$ and the third-order term decomposes via the Edgeworth-style expansion

$$\mathbb{E}[V_1 V_2 V_3 \mid x] = \alpha_1\alpha_2\alpha_3 + \alpha_1 C_{23} + \alpha_2 C_{13} + \alpha_3 C_{12} + \kappa_{123}(x),$$

where $\kappa_{123}(x)$ is the third joint cumulant. The sign of each correction term is governed by the same rank-1-alignment argument as in Theorem 3.1: under Ye et al. [31]'s low-rank dynamics, second-order covariances and third-order cumulants are non-negative on a measure-one subset of the deployer's distribution.

B. Full proof of Theorem 4.1 (non-implementability under per-verifier elicitation)

We complete the construction of Section 4. Under the per-verifier mechanism with strictly proper scoring rule S, the slot-1 payment to candidate V_1 on probe n is $S(\hat{p}_{1,n}, \ell_n)$, where $\hat{p}_{1,n}$ is the marginal report on prompt x_n. Under truthful reporting, the expected per-slot payment is $\mathbb{E}_\ell S(\alpha_1(x), \ell) = -\alpha_1(x)(1-\alpha_1(x))$ for the Brier score, equal to $-0.6 \cdot 0.4 = -0.24$ on both candidates V_1, V_1' in the construction. The per-slot empirical scores are therefore identical in expectation. The deployer's selection rule cannot break the tie based on marginal-only information.

The strategic refinement of Section 4 says further that, when a verifier is permitted to commit to a joint distribution within its calibration-monotone class, the per-verifier mechanism is silent on the choice of joint distribution. By the payoff-equivalence theorem [24],

any two mechanisms that elicit the same marginal reports must produce identical expected payments to verifiers with identical marginal types. The joint distribution is therefore not contractible under per-verifier elicitation, and verifiers face zero marginal incentive to commit to the cost-correct-optimal joint distribution. The pipeline ends up at an arbitrary joint distribution consistent with the verifiers' marginal types. □

C. Full proof of Theorem 6.1 (upper bound)

Hoeffding's inequality [16] for bounded iid random variables $X_n \in [0,1]$ gives, for each pair (k_1, k_2),

$$\Pr\Big[\Big|\hat{\alpha^{\text{pipe}}}_{k_1,k_2} - \alpha^{\text{pipe}}_{k_1,k_2}(P)\Big| \geq \epsilon\Big] \leq 2e^{-2N\epsilon^2}. \tag{8}$$

Union over $K_1 \cdot K_2$ pairs,

$$\Pr\left[\max_{(k_1,k_2)} \Big|\hat{\alpha^{\text{pipe}}}_{k_1,k_2} - \alpha^{\text{pipe}}_{k_1,k_2}(P)\Big| \geq \epsilon\right] \leq 2K_1K_2e^{-2N\epsilon^2}.$$

Set $\delta = 2K_1K_2e^{-2N\epsilon^2}$, so $\epsilon = \sqrt{(\log K_1 + \log K_2 + \log(2/\delta))/(2N)}$. On the complement event, the empirical arg max selects $(\hat{k}_1, \hat{k}_2)$ with $\hat{\alpha^{\text{pipe}}}_{\hat{k}_1,\hat{k}_2} \geq \hat{\alpha^{\text{pipe}}}_{k_1^*,k_2^*}$, so

$$\alpha^{\text{pipe}}_{k_1^*,k_2^*} - \alpha^{\text{pipe}}_{\hat{k}_1,\hat{k}_2} \leq (\hat{\alpha^{\text{pipe}}}_{k_1^*,k_2^*} + \epsilon) - (\hat{\alpha^{\text{pipe}}}_{\hat{k}_1,\hat{k}_2} - \epsilon) \leq 2\epsilon.$$

To pass from the high-probability bound to the expectation, write $\Delta = \alpha^{\text{pipe}}_{k_1^*,k_2^*} - \alpha^{\text{pipe}}_{\hat{k}_1,\hat{k}_2} \in [0,1]$. The complement-event argument above shows that $\Delta \leq 2E$ where $E = \max_{(k_1,k_2)} |\hat{\alpha^{\text{pipe}}}_{k_1,k_2} - \alpha^{\text{pipe}}_{k_1,k_2}(P)|$, and the union bound gives $\Pr[\Delta > u] \leq \Pr[E > u/2] \leq 2K_1K_2e^{-Nu^2/2}$. Use the tail-integration identity $\mathbb{E}[\Delta] = \int_0^1 \Pr[\Delta > u]\, du$ and split the integration at

$$u_0 = \sqrt{\frac{2\log(2K_1K_2)}{N}},$$

chosen so that the union-bounded tail at u_0 equals 1. On $[0, u_0]$ bound the integrand by 1 and obtain a contribution of u_0. On $[u_0, 1]$ apply the standard Gaussian tail inequality $\int_{u_0}^{\infty} e^{-Nu^2/2}\, du \leq (Nu_0)^{-1} \cdot e^{-Nu_0^2/2}$:

$$\int_{u_0}^{1} 2K_1K_2e^{-Nu^2/2}\, du \leq \frac{2K_1K_2 \cdot e^{-Nu_0^2/2}}{Nu_0} = \frac{1}{Nu_0},$$

where the last equality uses $2K_1K_2e^{-Nu_0^2/2} = 1$ by the choice of u_0. Substituting u_0,

$$\frac{1}{Nu_0} = \frac{1}{\sqrt{2N\log(2K_1K_2)}} \leq u_0$$

for $K_1K_2 \geq 1$, so the tail term is dominated by u_0. Therefore

$$\mathbb{E}[\Delta] \leq u_0 + \frac{1}{Nu_0} \leq 2u_0 = 2\sqrt{\frac{2\log(2K_1K_2)}{N}} \leq C_{\text{H}} \cdot \sqrt{\frac{\log K_1 + \log K_2}{N}}$$

for an absolute constant C_{H} that absorbs the $\log 2$ correction for $K_1K_2 \geq 2$. The unbiasedness assumption $\alpha^{\text{pipe}}_{k_1,k_2}(P) = \alpha^{\text{pipe}}_{k_1,k_2}(D)$ closes the bound on the deployer's distribution. □

D. Le Cam packing for Theorem 6.2 (lower bound)

Fix a base profile $\theta_0 \in (\mathcal{F}_1 \times \mathcal{F}_2)^{K_1 K_2}$. Construct $K_1 K_2$ alternative profiles $\theta_1, \ldots, \theta_{K_1 K_2}$ each differing from θ_0 in exactly one pair coordinate, where the differing pair is replaced by an adjacent pair in the calibration-monotone-pair order $\succeq$ with pipeline-accept-rate gap $\Delta_{\text{pipe}} > 0$. The total-variation distance between the empirical-report distributions induced by θ_0 and θ_j on probe sets of size N is bounded above by $\sqrt{N} \cdot \Delta_{\text{pipe}}$ via standard concentration. Le Cam's two-point inequality [26, §2.3] gives a lower bound on the worst-case selection error rate of $\Omega(1) \cdot \exp(-N\Delta_{\text{pipe}}^2)$. Choosing $\Delta_{\text{pipe}} = \sqrt{(\log K_1 + \log K_2)/N}$ balances the packing cardinality against the per-coordinate error rate. The reduction from selection error to expected regret is via the calibration-monotone-pair assumption, which forces any selection rule with worst-case regret ρ to identify the order in $\succeq$ at scale ρ. □

E. Proof of Proposition 7.2 (conditional-rare-event probes)

The Fisher-information argument adapts the sequential-elimination analysis of Karnin et al. [18]. With conditional-rare-event probes, each probe contributes per-probe information proportional to the joint disagreement gap $\sum_{(k_1,k_2)<(k_1',k_2')} (\alpha^{\text{pipe}}_{k_1,k_2} - \alpha^{\text{pipe}}_{k_1',k_2'})^2$ rather than to a single pairwise gap. The cumulative-gap complexity for joint identification scales as

$$H_{\text{cum,joint}} = \sum_{(k_1,k_2) \neq (k_1^*,k_2^*)} (\alpha^{\text{pipe}}_{k_1^*,k_2^*} - \alpha^{\text{pipe}}_{k_1,k_2})^{-2}.$$

The improvement factor over marginal-disagreement probes is governed by the ratio $H_{\text{cum,joint}}/H_{\text{cum,marg}}$, which is bounded below by $\sqrt{\min(K_1, K_2)}$ in the regime where joint gaps are well-spread. The argument fails when one joint gap dominates, in which case conditional-rare-event and marginal-disagreement probes coincide. □

F. Verifier-pair synthesis with prescribed conditional correlation

We synthesize verifier pairs (V_1, V_2) with target conditional disagreement covariance $\bar{C}^{\text{target}} \in \{-0.2, -0.1, 0, +0.1, +0.2\}$ as follows. Start with two independent logistic-regression heads on the dataset's feature space. Couple them through a shared latent factor $z \sim \mathcal{N}(0,1)$ that contributes weight λ to each head's logit. The conditional covariance $\bar{C}$ is monotone in λ over the range $\lambda \in [-1, 1]$. Calibrate λ to hit each target by binary search on a held-out fold. Empirical $\bar{C}$ matches the target to within ± 0.02 across all three datasets. Full code is in `simulation/synth_pair.py` alongside the paper source.

G. Simulation pseudocode

```
Input:  Dataset D, K1+K2 verifier candidates, probe budget N,
        mechanism M, probe strategy S, correlation regime C_regime
        (known or unknown)
Output: Selected pair (k1_hat, k2_hat), pipeline regret estimate

1. Generate probe set P:
   if S == "marginal":
       P = greedy_select(D, N, score=marginal_entropy_per_slot)
   elif S == "joint":
       P = greedy_select(D, N, score=joint_entropy_over_pairs)
   elif S == "conditional_rare":
       P = greedy_select(D, N, score=conditional_disagreement_rarity)
```

```
2. For each pair (k1, k2):
       hat_q[(k1, k2)] = [joint_report(V_k1(x_n, y_n), V_k2(x_n, y_n))
                          for (x_n, y_n, ell_n) in P]

3. Identifiability check:
   if rank(empirical_joint_correlation(hat_q)) < 2:
       flag failure; switch to conditional_rare; retry

4. For each pair (k1, k2):
   if M == "joint":
       payment[(k1, k2)] = a + b * mean(joint_brier(hat_q[(k1, k2)][n],
                                                    obs_n))
   elif M == "per_verifier":
       payment[(k1, k2)] = per_verifier_brier(V_k1)
                         + per_verifier_brier(V_k2)

5. (k1_hat, k2_hat) = argmax payment

6. Compute pipeline accept rate on held-out D-sample under
   f(V_k1_hat, V_k2_hat).

7. Return (k1_hat, k2_hat), pipeline_regret_estimate
```

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

H. Notation summary

References

[1] Manu Bhardwaj. Cost-correct as the binding optimization target. https://ifitsmanu.com/papers/verification-economics-2026, 2026. Field Notes #2.

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

[3] Manu Bhardwaj. Verifier procurement under unobservable quality. a scoring-rule mechanism for cost-correct minimization. ifitsmanu.com/papers/verifier-procurement, 2026. Wedge Verification Economics, Paper #1. Shipped Week 1.

[4] Manish Bhattarai, Ismael Boureima, Nishath Rajiv Ranasinghe, Scott Pakin, and Dan O'Malley. Rubric-grounded RL. Structured judge rewards for generalizable reasoning. *arXiv preprint arXiv:2605.08061*, 2026. URL https://arxiv.org/abs/2605.08061.

[5] Florian A. D. Burnat and Brittany I. Davidson. A benchmark for strategic auditee gaming under continuous compliance monitoring. *arXiv preprint arXiv:2605.06340*, 2026. URL https://arxiv.org/abs/2605.06340.

[6] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

[7] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher

Symbol	Meaning
D	Deployer's task distribution over (x, y) pairs
θ	Deployer's quality threshold
K_1, K_2	Number of candidate verifiers per slot
N	Probe budget
V_{k_i}	Verifier k_i in slot i, binary-valued
$\alpha_{k_i}(x)$	Conditional accept rate of slot-i verifier
$C_{k_1,k_2}(x)$	Within-instance disagreement covariance
$\bar{C}$	Unconditional disagreement covariance, $\mathbb{E}_D C(x)$
f	Monotone Boolean composition rule on reports
$\alpha^{\text{pipe}}_{k_1,k_2}$	Pipeline accept rate under composition f
$\text{CPM}_{1:1}$	Blended public-API cost per million tokens
R	Reasoning multiplier
$\bar{\rho}$	Average rollout-or-rejection ratio
CostCorrect	$\text{CPM}_{1:1} \cdot R \cdot (1 + \bar{\rho})/\alpha^{\text{pipe}}$
S	Strictly proper scoring rule on $\Delta(\{0,1\}^2)$
$\hat{q}_{(k_1,k_2),n}$	Reported joint distribution for pair (k_1, k_2) on probe n
$\hat{\alpha}^{\text{pipe}}$	Empirical pipeline accept rate on probes
$(\hat{k}_1, \hat{k}_2)$	Selected verifier pair
Reg	Expected gap from first-best Cost-correct on pipeline
C_{H}	Universal Hoeffding constant in Theorem 6.1

Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

[8] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2nd edition, 2006.

[9] 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.

[10] European Parliament and Council. Regulation (eu) 2024/1689 on artificial intelligence (ai act). Official Journal of the European Union, 12 July 2024, 2024. Articles 9, 13, 14, 15. High-risk obligations apply from 2 August 2026.

[11] Rafael Frongillo and Ian A. Kash. General truthfulness characterizations via convex analysis. *Games and Economic Behavior*, 130:636–662, 2021.

[12] Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

[13] Geoffrey Grimmett and Dominic Welsh. *Probability: An Introduction*. Oxford University Press, 2nd edition, 2014.

[14] 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 preprint arXiv:2501.04519*, 2025. URL https://arxiv.org/abs/2501.04519.

[15] 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 *Advances in Neural Information Processing Systems 34, Track on Datasets and Benchmarks*, 2021.

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

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

[18] 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.

[19] Yuqing Kong and Grant Schoenebeck. An information theoretic framework for designing information elicitation mechanisms that reward truth-telling. *ACM Transactions on Economics and Computation*, 7(1), 2019. doi: 10.1145/3296670.

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

[21] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2305.20050.

[22] Lauri Lovén. Honest reporting in scored oversight. true-kl0 property via the prekopa principle. *arXiv preprint arXiv:2605.03793*, 2026. URL https://arxiv.org/abs/2605.03793.

[23] 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.

[24] Roger B. Myerson. Optimal auction design, 1981.

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

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

[27] Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022. URL https://arxiv.org/abs/2211.14275.

[28] 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 *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2203.11171. arXiv preprint posted 2022.

[29] 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, 2012.

[30] Yang Xu and Chiwoo Park. Online bayesian calibration under gradual and abrupt system changes. *arXiv preprint arXiv:2605.06612*, 2026. URL https://arxiv.org/abs/2605.06612.

[31] Hao Ye, Jisheng Dang, Junfeng Fang, Bimei Wang, Yizhou Zhang, Ning Lv, Wencan Zhang, Hong Peng, Bin Hu, and Tat-Seng Chua. On the implicit reward overfitting and the low-rank dynamics in rlvr. *arXiv preprint arXiv:2605.06523*, 2026. URL https://arxiv.org/abs/2605.06523.

[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 *Advances in Neural Information Processing Systems 36*, 2023.

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